# Bioinformatic tools for single-cell data analysis in clinical studies

by

Brinda Monian

B S Chemical Engineering and Biochemistry
North Carolina State University, 2013

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

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

FEBRUARY 2020

Signature of Author Signature redacted
Department of Chemical Engineering
September 25, 2019

Certified by Signature redacted
J Christopher Love
Raymond A (1921) and Helen E St Laurent Professor of Chemical Engineering
Thesis Supervisor

Accepted by Signature redacted
Patrick S Doyle
Robert T Haslam (1911) Professor of Chemical Engineering
Chairman, Department Committee of Graduate Theses

# Bioinformatic tools for single-cell data analysis in clinical studies

by

Brinda Monian

Submitted to the Department of Chemical Engineering on September 25, 2019
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Chemical Engineering

## Abstract

Mechanistic understanding of disease has been dramatically enhanced by an explosion of new high-throughput experimental techniques for profiling biological samples, including RNA-Seq, mass spectrometry, and single-cell sequencing. However, the ability to gather exponentially more measurements comes with pitfalls of increased Type I error and reduced interpretability. In theory, single-cell measurements can be helpful in combating this problem, since each sample of cells represents hundreds to thousands of observations. But thinking is still emerging on how best to utilize single-cell data to boost statistics and generate meaningful findings.

This thesis represents several parallel efforts to develop and apply new bioinformatic techniques to generate robust findings from single-cell data. The advances are especially pertinent for small clinical studies in which low sample numbers are limiting. In the first part of the thesis, two classes of methods are introduced: gene module discovery in single-cell RNA sequencing data using sparse PCA, and probability-based metrics for evaluating the degree of association between paired modalities of single-cell data (in this case, single-cell RNA sequencing and paired TCR sequencing data). The methods are shown on two different human datasets, as proof-of-concept and examples of the biological findings capable of being unearthed.

In the second part of the thesis, these methods are applied to larger clinical datasets with questions surrounding acquired tolerance and clinical reactivity in food allergy. In the first study, T-helper cells from peanut-allergic patients undergoing oral immunotherapy were profiled to identify therapy-induced effects and baseline predictors of outcome. Two distinct subsets of expanded TH2 clones were found to be suppressed, but not deleted, by the therapy. In the second study, transcriptional correlates of clinical reactivity were evaluated in peanut-activated memory T-helper cells from peanut-allergic adults. Cells from more reactive patients had higher expression of TH1 and MHC I gene programs, suggesting activation of auxiliary, non-TH2 cell types. In each of these studies, new single-cell analysis techniques were integrated to generate clinical findings with improved robustness and interpretability.

## Acknowledgements

My graduate school experience has been principally shaped by the people who supported me along the way I look back on these years with a feeling of gratitude for all the people and experiences I've had the chance to learn from

Firstly, I would like to thank the patients who participated in the clinical studies that we worked on The scope of biomedical research is evolving such that both scientifically and ethically, it increasingly makes sense to study disease using human samples Doing so, however, requires the great generosity and forward-lookingness of individuals who are willing to participate, often without any benefit to themselves My thesis work would not be possible without such generosity

I am grateful to my adviser, Chris Love, for taking me on as a graduate student six years ago and entrusting me with a place in the lab, intellectual freedom, clinical samples and lab resources, and leadership on projects Chris worked hard to create a wonderful lab environment with fulfilling research projects and a generous group of people, from which I benefitted immensely My thesis committee members, Wayne Shreffler and Doug Lauffenburger, were also an excellent source of guidance I always looked forward to our annual meetings, which were useful and energizing

Early on in graduate school, several senior members of the lab and building took the time to train me and help me find my footing, including Todd Gierahn who taught me so much about single-cell profiling and immunology, Chris Dupont who trained me on the ins and outs of flow cytometry, Greg Szeto who taught me about multiplexed ELISA and good experimental design, and Brittany Goods, who has been both a role model and a friend

Almost all work in our lab was a team effort, and I'm thankful to have had fantastic teammates In the lab, this included Andy Tu, Duncan Morgan, Todd Gierahn, Patrick Petrossian, Rachel Barry, Lionel Lam, Gary Shea, and my wonderful undergraduate researcher, Julia Ginder Our collaborators outside the lab included the Shreffler lab at MGH (especially Wayne Shreffler, Bert Ruiter, and Sarita Patil), Jessica Savage at BWH, and partners at Sanofi MIT staff were also instrumental in keeping research projects moving Noelani Kamelamela and the Bio-Micro Center, everyone at the KI flow cytometry core, Charlie Whittaker, Mariann Murray, and Danielle Camp

Regardless of whether research was going well or poorly, I looked forward to coming to work with friends and labmates who infinitely enriched my graduate school experience I am so grateful to the entire Love lab, as well as our neighbors in the Manalis lab, for creating a fun, welcoming, and motivating work environment (complete with cryptic crosswords)

Finally, I feel the utter impossibility of expressing how much my close friends and family have meant to me during this journey Nikunja and Leia, with whom I've shared hikes, delicious meals, and late-night rants, Heidi, whose friendship I'm not sure I deserve, and who chatted with me about graduate school by Skype countless times, my family – Ashwin, Amma, Appa, and my extended family – who have been incomparably supportive and loving, and finally, those who I acquired while in graduate school, Gokul (who has inspired me and made me better) and Raleigh (who loved me no matter the state of my thesis, as long as it didn't infringe on walks)

# Table of Contents

## List of figures and tables

# 1. Introduction

In the last hundred years, the landscape of human disease has been drastically altered by advancements in therapies and vaccines, changes in environment, diet, and lifestyle, increased use of antibiotics, and several other factors[1,2] While some of these elements reflect great strides in treating infectious disease, others may play a role in the increasing prevalence of noncommunicable disease (such as heart disease, cancer, and type II diabetes), which accounted for 68% of all deaths globally in 2012[3] Several of these diseases, particularly immune-mediated disorders such as autoimmune diseases, cancer, and allergy, have complex mechanisms and present significant challenges for developing treatment strategies in the absence of an obvious pathogen In order to create effective diagnostics and therapies for these diseases, a precise understanding of the underlying cellular and molecular states is required Ideally, human clinical samples will need to be profiled at the deep resolution required to achieve such understanding

In response to this need, there has been a recent explosion of new high-throughput techniques for profiling biological samples, including RNA-Seq, ChIP-Seq, ATAC-Seq, and mass spectrometry for proteomics or metabolomics These powerful methods allow for the unbiased, parallel measurement of all molecules of a particular class, without *a priori* knowledge required of the targets to be analyzed This means that hundreds to thousands of molecules can be profiled simultaneously to determine disease-relevant interactions and signaling pathways, not just individual markers, and that previously unknown biomolecules can be identified as nodes for monitoring or modulation Several of these measurements are now possible at the single-cell level, enabling even more high-resolution mechanistic discovery and tracking of heterogeneous cell populations

This exciting set of innovations presents several opportunities for medical advancement, and significant challenges for data analysis An often-ignored pitfall of the ability to profile samples in a high-throughput, unbiased fashion is the vastly increased probability of Type I error, i e false-positives A closely related challenge is interpretability, sifting through hundreds of features instead of a few can run counter to the goal of identifying a small, reliable set of biomarkers or drug targets Because the field is so nascent, an additional challenge is knowing what research questions are possible and feasible to ask of high-throughput datasets

In the effort to overcome these problems, single-cell measurements could be unexpectedly helpful In a single-cell experiment, each sample theoretically represents hundreds to thousands of independent observations (i e cells), it therefore seems feasible to be able to utilize these data to improve, rather than hurt, statistical power and interpretation But thinking is still emerging on how to go about this in practice

This thesis represents several endeavors to develop and apply bioinformatics techniques for attaining adequate statistics and biological meaning in single-cell data analysis In the remainder of this chapter, I review motivations for single-cell profiling and existing single-cell experimental methods, as well as precedent and prior thinking on how to cleverly use single-cell data to improve, rather than hinder, reproducibility and interpretability Finally, I preview the subsequent chapters of the thesis that describe new advances in data analysis and clinical discoveries that my colleagues and I made in this space

## 1.1. Motivation for single-cell resolution of clinical samples

There is an increasing appreciation that human cells, beyond well-described differences in lineage, are extremely heterogeneous, and that highly specific subsets of cells may be dysregulated in disease For example, it has recently been shown that subsets of exhausted CD8 T cells may be differentially responsive to cancer immunotherapies[4], myelin-reactive T-helper cells have divergent gene expression programs in healthy adults and multiple-sclerosis patients[5], and a small subset of dendritic cells that expresses an antiviral program is associated with elite control of HIV[6] Thus, for most disease contexts there are compelling reasons to study clinical samples at the single-cell level in order to unearth insights regarding specific subtypes of cells

In addition to discovering new disease-relevant cell types, single-cell resolution is advantageous or required for the following goals 1) *Finding rare cells against a large background of unwanted cells.* As an example, antigen-specific T cells are extremely rare in circulation (about 1 in $10^3$ to $10^5$ cells)[7] and cannot be profiled reliably by bulk methods without extremely stringent selection Single-cell resolution relaxes the requirement for stringent selection of rare cell types and enables many new types of studies 2) *Studying multiple cell types from one sample in parallel* Profiling multiple cell types at once may be a practical goal, based on cost or a clinical sample with low cell numbers Alternatively, it may be motivated by the desire to model intercellular communication by profiling signaling cues in multiple cell types, or to survey all cells present in a tissue environment Studying a sample at single-cell resolution can allow for the parallel measurement of most or all cell types present in the dataset 3) *Separating out the effects of changes in cell frequencies from changes in intracellular gene expression levels* Cell frequencies and functional states often change in tandem, and both changes are reflected in bulk readouts such as gene expression In bulk measurements, the two phenomena are impossible to tell apart, and the distinction between them is critical for understanding regulation and mechanism in disease Single-cell measurements allow for the resolution of whether, for example, cells are proliferating and increasing in number, or whether cells are individually becoming more activated

## 1.2. Experimental methods for single-cell profiling

Stemming from an increased appreciation for the roles that rare and heterogeneous cell types play in diseases like food allergy, there has been a proliferation in new experimental techniques for single-cell profiling  Until recently, the only widely available methods for single-cell profiling were flow cytometry and ELISPOT, which could measure up to 17 multiplexed surface or intracellular molecules, and a single secreted cytokine, per cell, respectively[8]  These assays were augmented by the advent of single-cell mass cytometry, or CyToF (which can reliably profile up to 40 molecules per cell with isotope-tagged instead of fluorophore-tagged antibodies)[9], and single-cell protein secretion assays[10] (which can profile 4 to 12 secreted proteins simultaneously from a single cell)  These advances led to new insights on the dynamics of cytokine secretion by T cells[11] and on how to derive signaling networks from single-cell data[12]  Results with these methods, though, were still limited to cellular markers that were known and for which antibodies could be made

At the same time, unbiased “omics” techniques for profiling samples in bulk, such as RNA-Seq, ChIP-Seq, ATAC-Seq, proteomics, and metabolomics, were emerging  Rather than using reagents for a set of pre-specified molecules, these methods relied on sequencing, mass spectrometry, or other measurements that could recognize all analytes of a given class with relatively low bias  For example, in a typical RNA-Seq protocol, all RNA molecules with a poly(A) tail are enriched using poly(dT) binding reagents, amplified using PCR with random priming sequences, and read using next-generation sequencing, in as unbiased a way as possible for different transcripts (barring small inherent biases in amplification and sequencing for transcripts of differing GC content and length)  This class of methods allowed for the measurement of analytes that were not selected *a priori*, and even analytes that were not known, fundamentally changing the way biological samples could be studied

As a natural confluence of discoveries in cell heterogeneity and innovations in high-throughput measurements, massively parallel single-cell techniques have rapidly appeared and been improved upon in the past few years  Single-cell RNA sequencing techniques, for example, include plate-based sorting[13], drop-Seq[14], 10X[15], and Seq-Well[16] (progressively newer and more cost-effective), other measurement platforms include single-cell ATAC-Seq[17] and single-cell ChIP-Seq[18], as well as more specialized measurements like highly multiplexed single-cell FISH[19], dynamic single-cell mass cytometry[20], and multiplexed single-cell protein secretion[10] An extremely recent avenue of innovation has been the development of multimodal single-cell profiling, that is, the ability to measure two types of data on the same cell, such as paired spatial and transcriptomic information[21] Multimodal single-cell measurements are an exciting opportunity to provide critical context and validation to each dataset Overall, these advances allow for the unbiased discovery of precise and rare cell populations, giving additional resolution and knowledge into biological phenomena

## 1.3. Challenges and opportunities in single-cell data analysis

The emergence of massively parallel single-cell techniques is exciting for learning about basic biological phenomena and for uncovering complex mechanisms or biomarkers of disease However, significant challenges exist in converting the often terabytes of data generated by each experiment into actionable knowledge Analysis approaches are still emerging that are suited to the unique challenges of single-cell data, which include high sparsity and noise resulting from tiny amounts of starting material, vulnerability to batch effects, reliably identifying subpopulations of cells, and more Arguably, the biggest challenge to contend with is that the huge number of features generated can vastly increase the chances of false positives and obscure signal to noise in the discovery of potential biomarkers, drug targets, or disease mechanisms

Both high-throughput and single-cell methods represent a double-edged sword for discovering real, reproducible findings On one hand, high-throughput techniques allow for the measurement of thousands of analytes, and single-cell approaches allow for ever-increasing resolution and tracking of narrow subpopulations of cells, which together could enable more precise descriptions of disease states or treatment groups On the other hand, both of these classes of techniques also increase the number of features that can be defined per sample, without selecting for clinically relevant features, and thus also increase the probability of Type I error (false-positives) This downside is evident when one considers the hundreds or thousands of analytes and cell subsets being measured that are irrelevant to the biological context but that may randomly vary with the covariate of interest (such as clinical outcome) and cause a false signal By extension, integrating together multiple modalities of high-throughput data can make for even more potential features and thus false-positives Smaller clinical studies, in which the feature-to-sample ratio is more heavily skewed, are especially vulnerable to this pitfall This problem of irreproducibility has been broadly described in the biomedical literature and has been partially attributed to measuring many relationships without preselection[22]

Hand-in-hand with the difficulty of extracting signal from a sea of false-positives, is the challenge of attaining interpretability of results Biological interpretation and clinically actionable knowledge are a central desired outcome of single-cell studies, but the presence of thousands of features, of which tens or hundreds may be flagged as statistically significant, may be challenging to interpret Gene set enrichment analysis and other pathway analysis tools exist to identify common signaling pathways or network motifs among a myriad of significant genes, but these tools were designed for microarray or bulk RNA-Seq studies and are not easily generalized to single-cell studies (in part because of different library preparation techniques for bulk and single-cell samples, making comparisons to bulk reference datasets inherently biased)

Unexpectedly, the very nature of single-cell data could be helpful in mitigating these challenges In theory, each single cell is a separate observation within a biological sample, so it seems that cells could be utilized as individual samples, rather than just as the basis for additional features, to improve both statistical power and interpretability There is not, however, much precedent for how to envision this idea in practice, but the best example may be a study by Sachs et al in 2005 that derived an entire signaling network architecture *de novo* using single-cell CyToF data[12] The researchers made use of the fact that each cell was in a slightly different state of signaling, so correlations between nodes in the network could be robustly assessed using the thousands of collected observations, in combination with perturbations to the network This study was an example of how single-cell data improved the ability to make biological conclusions by allowing the researchers to organize the measurements into an accurate network As a completely different, clinically-focused example, a recent study showed that mutation calling within single circulating tumor cells could be made more robust by a census-based approach in which a mutation had to be observed in a threshold number of cells in order to be credible[23] In this case, measuring single-cell exomes allowed the researchers to have confidence (and increased statistical power) from the same mutation being detected in multiple cells In both of these studies, the liabilities of individual cells as observations were acknowledged and overcome by analyzing many cells together These studies serve as useful and thought-provoking examples of how one might use single-cell high-throughput data to help, rather than hurt, statistics and interpretability

## 1.4. Thesis objectives

The aims of this thesis work are twofold The first is to make advances in the quantitative analysis of single-cell data to help boost reproducibility and biological interpretation, especially in studies with small numbers of samples The second is to apply these advances to larger clinical studies to help reveal new disease-relevant insights In light of these aims, the thesis is organized into two parts

**Part I: Bioinformatic tools**

- Chapter 2 Gene module discovery in single-cell RNA sequencing data of T cells in pediatric milk allergy Analysis of single-cell RNA sequencing data is hindered by high noise and sparsity in individual genes In this chapter, we show a method to mitigate this challenge by compressing genes co-expressed within many single cells into gene modules without *a priori* knowledge The method is based on sparse PCA which allows for several useful mathematical features such as pseudo-orthogonality of the gene modules and algorithmic transparency We show the usefulness of the method, on a case study with T-helper cells from pediatric milk allergy samples, in discovering and interpreting multiple functional states in milk-reactive T cells that vary longitudinally

- Chapter 3 Analysis of multimodal single-cell data to identify a highly activated, clonotypically distinct state in antigen-reactive T cells An exciting new extension of single-cell profiling is the ability to measure analytes of multiple modalities on the same cell A central question is often the degree of overlap between the modalities, but how to evaluate this quantitatively and statistically is an ongoing question We developed an analysis framework incorporating concepts from probability and information theory to concisely convey the overlap between T-cell receptor (TCR) sequence and paired transcriptome data We apply this framework to a case study of CMV-reactive T-helper cells, isolated at different stimulation times with CMV antigen, and we show a tight and time-dependent association between TCR sequence and T cell state, highlighting the possible role of epitope recognition in shaping the T-cell response Additionally, we identified a highly activated, clonotypically distinct state that peaks at very early time points, is likely associated with specific epitopes, and has not been previously described

**Part II: Clinical applications**

- Chapter 4 Transient suppression, but not deletion, of distinct subsets of TH2 clonotypes in peanut oral immunotherapy Oral immunotherapy (OIT) is a trial-stage treatment for food allergy with a low rate of inducing sustained unresponsiveness to allergen We profiled T-helper cells from twelve peanut-allergic patients experiencing different outcomes from OIT, in order to better understand T cell changes induced by the treatment and cell states that could predict patient outcome Using single-cell RNA sequencing and paired TCR sequencing, we found multiple clonotypically distinct subsets of TH2 cells that were transiently suppressed, but not deleted, by treatment Additionally, we discovered baseline predictors of outcome including TH17 signatures These findings underscore the difficulty of inducing durable reprogramming in OIT and that clinical outcome may be set in stone prior to the start of treatment

- Chapter 5 T-cell correlates of clinical reactivity to peanut allergen Peanut-allergic patients present with a wide range of clinical symptoms and severities, and the underlying immune states contributing to this heterogeneity are not well-understood We profiled circulating peanut-reactive T-helper cells from peanut-allergic patients with high or low clinical reactivity to peanut, in order to discover associated T cell states We found that surprisingly, TH1 and MHC I gene modules were upregulated in cells from patients with higher reactivity to peanut, and that TH2 modules were slightly upregulated in less reactive patients, suggesting that auxiliary effector phenotypes might be correlated with more severe disease

## 1.5. References

1 Guarner, F *et al* Mechanisms of Disease the hygiene hypothesis revisited *Nat Clin Pract Gastroenterol Hepatol* **3**, 275–284 (2006)

2 Allen, L Are we facing a noncommunicable disease pandemic? *J Epidemiol Glob Health* **7**, 5–9 (2017)

3 WHO I Global status report on noncommunicable diseases 2014 *WHO* Available at http //www who int/nmh/publications/ncd-status-report-2014/en/ (Accessed 5th August 2019)

4 Miller, B C *et al.* Subsets of exhausted CD8+ T cells differentially mediate tumor control and respond to checkpoint blockade *Nat Immunol* **20**, 326–336 (2019)

5 Cao, Y *et al* Functional inflammatory profiles distinguish myelin-reactive T cells from patients with multiple sclerosis *Sci Transl Med.* **7**, 287ra74-287ra74 (2015)

6 Martin-Gayo, E *et al.* A Reproducibility-Based Computational Framework Identifies an Inducible, Enhanced Antiviral State in Dendritic Cells from HIV-1 Elite Controllers *Genome Biol* **19**, 10 (2018)

7 DeLong, J H *et al* Ara h 1–reactive T cells in individuals with peanut allergy *J Allergy Clin. Immunol* **127**, 1211-1218 e3 (2011)

8 Perfetto, S P , Chattopadhyay, P K & Roederer, M Seventeen-colour flow cytometry unravelling the immune system *Nat. Rev. Immunol* **4**, 648–655 (2004)

9 Spitzer, M H & Nolan, G P Mass Cytometry Single Cells, Many Features *Cell* **165**, 780–791 (2016)

10 Love, J C , Ronan, J L , Grotenbreg, G M , Veen, A G van der & Ploegh, H L A microengraving method for rapid selection of single cells producing antigen-specific antibodies *Nat Biotechnol* **24**, 703 (2006)

11 Han, Q *et al* Polyfunctional responses by human T cells result from sequential release of cytokines *Proc Natl Acad Sci. U S A* **109**, 1607–1612 (2012)

12 Sachs, K , Perez, O , Pe'er, D , Lauffenburger, D A & Nolan, G P Causal protein-signaling networks derived from multiparameter single-cell data *Science* **308**, 523–529 (2005)

13 Picelli, S *et al* Smart-seq2 for sensitive full-length transcriptome profiling in single cells *Nat. Methods* **10**, 1096–1098 (2013)

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

15 Zheng, G X Y *et al* Massively parallel digital transcriptional profiling of single cells *Nat. Commun* **8**, 14049 (2017)

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

17 Buenrostro, J D *et al* Single-cell chromatin accessibility reveals principles of regulatory variation *Nature* **523**, 486–490 (2015)

18 Rotem, A *et al* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state *Nat Biotechnol* **33**, 1165–1172 (2015)

19 Wang, G , Moffitt, J R & Zhuang, X Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy *Sci Rep* **8**, 4847 (2018)

20 Burg, T P *et al* Weighing of biomolecules, single cells and single nanoparticles in fluid *Nature* **446**, 1066–1069 (2007)

21 Rodriques, S G *et al* Slide-seq A scalable technology for measuring genome-wide expression at high spatial resolution *Science* **363**, 1463–1467 (2019)

22 Ioannidis, J P A Why Most Published Research Findings Are False *PLOS Med* **2**, e124 (2005)

23 Lohr, J G *et al* Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer *Nat Biotechnol* **32**, 479–484 (2014)

# 2. Gene module discovery in single-cell RNA sequencing data of T cells in pediatric milk allergy

This chapter introduces a technique for compressing individual genes in single-cell RNA sequencing data into gene modules in an unsupervised manner  Single-cell RNA sequencing data is inherently noisy and sparse, qualities that are problematic for analyses such as differential expression  We show that using a sparse PCA-based approach, this noise can be mitigated and biologically meaningful gene modules can be generated from measurements of individual genes  Compared to other methods by which researchers have addressed this challenge, our approach is algorithmically transparent and user-friendly  Using a dataset generated from T cells from pediatric milk allergy patients, we show that the method generates gene modules representing both known and unknown T cell functional programs without appreciable loss of information  By tracking the modules longitudinally within patients, we observe that TH1 and NF-kB gene modules are upregulated as a milk-allergic individual ages, regardless of how their allergy progresses  Overall, this gene module approach is simple and powerful for gleaning knowledge about gene programs present in the cells and for dimensionality reduction for downstream tests

## 2.1. Motivation

### 2.1.1. Challenges of dropout in single-cell RNA sequencing data

Single-cell RNA sequencing is a powerful technique for discovering new subpopulations of cells and identifying disease-relevant cell states and signaling pathways A key challenge of single-cell RNA-Seq data, however, is a phenomenon known as dropout[1,2], wherein many, often a majority, of entries in the digital gene expression matrix of cells versus gene counts are zeros (Figure 2-1)

| | TNFRSF1B | CTSH | CXCL5 | MAF | FOXP3 | MS4A1 | BAZ2B | AKAP9 | HIVEP3 |
|---|---|---|---|---|---|---|---|---|---|
| **Cell 1** | 9 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **Cell 2** | 18 | 0 | 0 | 1 | 5 | 0 | 0 | 0 | 0 |
| **Cell 3** | 12 | 0 | 0 | 3 | 0 | 0 | 0 | 4 | 0 |
| **Cell 4** | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 2 | 1 |
| **Cell 5** | 17 | 0 | 0 | 5 | 0 | 0 | 0 | 2 | 0 |
| **Cell 6** | 21 | 1 | 0 | 3 | 0 | 0 | 0 | 0 | 1 |
| **Cell 7** | 9 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| **Cell 8** | 25 | 1 | 0 | 5 | 0 | 0 | 0 | 0 | 1 |
| **Cell 9** | 10 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Cell 10** | 11 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| **Cell 11** | 11 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **Cell 12** | 10 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Cell 13** | 15 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| **Cell 14** | 7 | 1 | 0 | 10 | 1 | 0 | 0 | 1 | 0 |
| **Cell 15** | 14 | 4 | 0 | 2 | 0 | 0 | 0 | 0 | 1 |

Figure 2-1 Snapshot of single-cell RNA-Seq digital gene expression matrix This matrix represents a random snapshot of cells and variable genes in an experiment with milk-reactive memory T-helper cells

Dropout is attributed to two main causes The first is biological mRNA transcription has been described to happen in "bursts", not uniformly, meaning that zeros may simply be a function of transcriptional noise and sampling time[3] Depending on the application, it may be desirable to retain this "real" biological noise The second cause of dropout is technical Capturing the tiny amount of mRNA from a cell is inherently inefficient[1], a rough estimate of the fraction of starting mRNA that is converted to cDNA is proposed to be 2-20% for single-cell RNA-Seq via 10X[4] Thus, many of the zeros are likely missing values due to insufficient mRNA capture (or subsequent under-amplification or under-sequencing)

Regardless of the reason, dropout represents a significant challenge for any quantitative analysis Individual gene measurements are complicated by noise and sparsity, and as a result, only highly-expressed genes tend to score well in differential expression tests (Figure A2-3) Additionally, it is challenging to fully utilize single cells as separate observations, which may be desirable for longitudinal tracking, trajectory-based inferencing, or other analyses, but which is hampered by the noisiness of data within each individual cell In order to adequately address these challenges of dropout, statistics must either account for the zero-inflated nature of the data, or the data must be smoothed in some way

### 2.1.2. Approaches for addressing dropout

Three main approaches have emerged along these lines The first is to employ statistical models that explicitly model dropout[1,5] or the high incidence of zeros[6,7], or that give lower importance to genes likely to be affected by dropout[8] Such tests are designed to decrease the biases produced by dropout on subsequent statistical analyses, by modeling dropout events or zeros probabilistically This approach is the simplest and does not require data manipulation, but is also the narrowest in scope, as it can only be used if the research question of interest can be answered with one of the available models Additionally, imperfectly modeling the molecular events surrounding dropout and biased amplification (which is, to some extent, inevitable) can actually introduce technical artifacts into differential expression or other analyses[9]

The second approach uses imputation to replace some of the zeros or low-count values with higher, imputed values based on predictions of which ones represent dropout events. This strategy, which includes methods such as MAGIC[10], DrImpute[11], and scImpute[12], typically relies on similarity to nearest-neighbor cells as the basis for imputation. In theory, this approach is powerful and broadly useful for any downstream analysis, correcting the data for technical defects in mRNA recovery, but in practice it can be extremely precarious, as imputation can paint over biological noise, cell-to-cell variability, and other real features in the data.

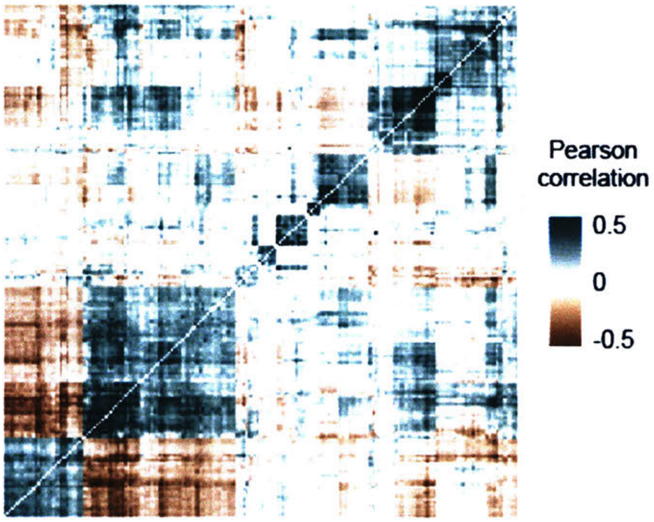


Figure 2-2. Correlation matrix of immune genes in a single-cell RNA-Seq dataset of T cells. Pearson correlation of every gene pair is plotted.

The third approach relies on dimensionality reduction for smoothing, making use of the phenomenon that small subsets of genes tend to be highly co-expressed (Figure 2-2), and that combining genes into gene "modules" may thus be a way to reduce noise from dropout without loss of information. Such a method is ideally done in an unbiased way, without *a priori* information such as published gene sets, since not all disease contexts or cell states have well-described gene pathways, and moreover, not all published gene sets are generalizable to different experimental setups or clinical settings  Existing gene module discovery methods are predominantly based on non-negative matrix factorization (NMF)[13,14,15] or topic modeling[16,17], and have been used to great effect  In addition to reducing noise, this approach has the added advantage of decreased Type I error and improved interpretability over individual genes, because of the reduced number of features and the biological information encoded in the grouping of genes  Drawbacks of this strategy include the inability to directly address the noise present in individual genes, if individual genes are required for any analyses (which they often are)

Both of the smoothing approaches described (imputation and dimensionality reduction) cleverly make use of the single-cell nature of the data as the basis for smoothing, and are feasible for use on single-cell data generated from a variety of sample types and experimental techniques  Almost all existing methods, however, have limitations on ease of implementation (successfully accessing and running the method) and/or algorithmic transparency  We saw a need for simpler techniques that were easier both to implement and to understand intuitively

In light of this, we came up with a new dimensionality reduction method for discovering co-expressed gene modules from genes in single-cell RNA-Seq data  The method is simple and easy to implement, being based on a method for sparse PCA (principal component analysis)[18] Sparse PCA is like PCA in that it discovers co-expressed genes by deriving principal components, i e  the eigenvectors of the correlation matrix of genes, but it additionally imposes a lasso penalty on each component to induce sparseness  This method confers the following beneficial properties  1) the sparsity can be tuned using a single parameter for the desired gene module size depending on the biological application, where smaller modules usually have increased interpretability, 2) gene modules can be made to be pseudo-orthogonal (exact orthogonality is not possible with sparse PCA, unlike PCA), meaning that redundant or highly overlapping gene programs are unlikely, and 3) both negative and positive gene weights can be optionally allowed, which may better represent certain biological phenomena in which genes are negatively regulated with respect to each other  As a case study, we show the method's usefulness in discovering gene modules, and thus enabling better interpretation of longitudinal analysis, on a single-cell dataset of T cells from pediatric milk allergy patients

### 2.1.3. Motivation for case study in pediatric milk allergy

Milk allergy, unlike some other food allergies, is often but not always outgrown in childhood[19] The reasons for acquisition of tolerance to milk in a subset of the population are still being investigated, and are of great interest in the effort to develop better therapies for food allergy For example, correlates of natural tolerance could be useful in directing desired cell states to induce during allergen-specific immunotherapy  To discover such correlates, we collected samples at two time points (about 1-2 years apart) from children whose milk allergy either improved in status, remained the same, or was already resolved, in order to identify immune changes associated with tolerance  We measured gene expression in milk-reactive T-helper cells, a central player in milk allergy  Previous work has shown the transient emergence of Tregs in milk-tolerant patients[20] as well as elevated TH1 function in allergen-stimulated cells from peanut-tolerant patients[21], but a deep and comprehensive survey of the functional phenotypes present in milk-specific T cells was not previously possible  Here we show, using our method for gene module discovery, the diversity of major T-cell states that are present among milk-reactive T cells  We also perform a differential expression analysis to track which cell states are altered with time, and we show that the gene module approach helps in interpreting these changes

## 2.2. Methods

**Pediatric cohort.** Patients were enrolled in accordance with IRB number 2014P000256/PHS at Partners Health  Pediatric milk allergy patients ages 3 to 12 were enrolled for a longitudinal study  Patients were first subjected to a diagnostic panel for milk allergy including a skin-prick test, IgE blood test, and baked and uncooked milk challenges to assess their level of reactivity  Patients then gave two longitudinal blood samples about one year apart, with additional diagnostic tests at or near the second visit to evaluate whether their milk allergy status had changed  At each time point, patients were classified by an allergist as having full milk allergy (able to tolerate neither baked nor uncooked milk), partial milk allergy (able to tolerate baked but not uncooked milk), or no milk allergy (able to tolerate both baked and uncooked milk)  Based on the status at both time points, patients were binned into one of four groups  “Persistent”, meaning the patient had the same status of milk allergy at both time points, “Transient”, meaning the patient’s milk allergy status had improved between the two time points (from full to partial, or from full or partial to no allergy), “Resolved”, meaning that the patient had a previous history of milk allergy but no diagnostic signs of milk allergy at either time point, or “No history,” meaning the patient had never had any history of milk allergy  At each time point, a blood sample was collected and PBMCs were immediately isolated via density-gradient centrifugation using Ficoll-Paque (GE) and frozen in FBS containing 10% DMSO

**PBMC stimulation and sorting.** Cryopreserved PBMCs from pediatric milk-allergic donors were thawed and plated in a 24-well plate at 5M cells/1ml/well in AIM-V medium (Gibco)  PBMCs were stimulated with either milk extract (100ug/ml), anti-CD3/CD28 beads (Dynabeads) or PBS  Cultures were incubated at 37C for 22h before being harvested for staining and flow sorting  3h before harvesting each culture, PE anti-CD154 antibody (BD, clone TRAP1) was added at a dilution of 1 50  For staining, cells were incubated for 25min at 4C with live-dead blue viability dye

(Biolegend), AF700 anti-CD3 (clone), APC-Cy7 anti-CD4 (clone), FITC anti-CD45RA (clone HI100), PE anti-CD154 (clone TRAP1), and APC anti-CD137 (clone 4B4-1) All antibodies were purchased from Biolegend except for PE anti-CD154, which was from BD After staining, cells were sorted on a BD FACS Aria instrument Cells were sequentially gated as lymphocytes, singlets, live cells, CD3+CD4+, and CD45RA-, and were then sorted as either CD154+CD137+/-, CD154-CD137+, or CD154-CD137-

**Single-cell transcriptome sequencing.** Sorted cells were immediately processed for single-cell RNA sequencing via the Seq-Well protocol[22] Briefly, up to 30,000 cells per sample were co-loaded into wells at approximately single-cell occupancy with poly(dT) beads and lysed to allow mRNA to hybridize onto the beads Beads were then pooled and mRNA was reverse-transcribed, PCR-amplified, and prepared for sequencing via the Nextera XT kit Libraries were sequenced on the Illumina Novaseq

**Sequencing data preprocessing.** Raw read processing was performed as in Macosko et al[23] Briefly, sequencing reads were aligned to the hg38 human genome and counted to obtain a digital gene expression matrix of cells versus genes The matrix was filtered to exclude any cells with fewer than 1,000 detected genes or 2,000 detected transcripts (UMIs) Counts were then normalized by cell library size and log2-transformed using the Seurat package in R, and were visualized using a two-dimensional t-SNE projection

**Gene module discovery.** Gene modules were generated from the data using a sparse PCA approach described by Witten et al[18] and in the R package 'PMA' This approach employs an L1-norm penalty which constrains the sum of all gene weights in each component Prior to running sparse PCA, the normalized gene expression matrix (cells as rows, genes as columns) was randomly downsampled to have an equal number of cells from the top 70 (out of 109) samples, in order to prevent the results from being dominated by a few samples and to decrease computational time Genes were filtered down to the union of immune genes (defined as the gene lists on ImmPort at https //www immport org/shared/genelists) and the most variable genes in the dataset, using the 'var genes' command in the R package 'Seurat' Finally, the data was scaled and centered with respect to genes, and sparse PCA was run using the command 'SPC' (with 'orth' parameter set to TRUE and tuning parameter 'sumabsv' tuned as described below) Gene module scores were calculated as the scaled gene expression input matrix multiplied by the output loadings matrix 'v'

**Gating module expression** To classify whether cells were "expressing" a gene module or not, the distribution of module expression across all cells was fit to a mixture model of a Gaussian and a log-normal distribution (to model the case of a bimodal distribution with a clear negative population, where the Gaussian was chosen to model the negative population and the log-normal was chosen to model the positive population), and a single Gaussian (to model the case of a unimodal distribution) The fit with the lower AIC (Akaike Information Criterion) was taken In the case of the mixture model, cells assigned to the log-normal curve were classified as "expressing" the module, and in the case of the single Gaussian, cells at or above one standard deviation above the mean were classified as "expressing" the module

**Differential expression analysis.** Differential expression was performed using a Mann-Whitney U test, to suit the non-normal distribution of most single-cell features, with a Benjamini-Hochberg correction for multiple hypothesis testing

## 2.3. Results: Case study with pediatric milk allergy cohort

### 2.3.1. Milk allergy study design

To demonstrate the performance and features of the gene module discovery method, we ran it on data from T cells sequenced from a pediatric milk allergy cohort using Seq-Well Samples from thirteen children ages 3 to 12 were included in the study (Table 2-1) Patients were selected on the basis of having banked PBMC samples from two visits about 1-2 years apart, and a well-defined milk allergy diagnosis at both visits Based on allergy status at each time point, four patient groups were defined "Resolved", meaning the patient's milk allergy had fully resolved before the first visit, "Transient", meaning the patient acquired partial or complete tolerance to milk during the time between the two visits, "Persistent", which meant that the patient had milk allergy which did not change in severity between the two visits, and "No history", meaning that the patient had never shown signs of milk allergy

PBMCs from each visit were stimulated with milk antigen in order to preferentially activate milk-specific T cells CD4 memory T cells were sorted on the basis of CD154 and CD137, markers for capturing antigen-activated effector and regulatory T cells, respectively (Figure 2-3) CD154+CD137+/-, CD154-CD137+, and CD154-CD137- cells were then profiled using single-cell RNA-Seq This analysis focuses on the CD154+ compartment, for which we recovered the most cells A t-SNE visualization of CD154+ cell transcriptomes shows strong segregation by patient and a possible enrichment of certain cell clusters with time (Figure 2-4)

Table 2-1 Pediatric milk allergy cohort

| Patient ID | Age at enrollment | Gender | Time between visits 1 and 2 (days) | Baseline milk-specific IgE (kU/L) | Group |
|---|---|---|---|---|---|
| PM003 | 8 | M | 417 | 0 76 | Resolved |
| PM073 | 11 | M | 348 | 0 71 | Resolved |
| PM008 | 12 | M | 405 | 10 3 | Transient |
| PM009 | 9 | F | 274 | 2 4 | Transient |
| PM019 | 6 | F | 584 | <0 35 | Transient |
| PM078 | 12 | M | 383 | 2 6 | Transient |
| PM090 | 3 | F | 403 | 3 83 | Transient |
| PM002 | 6 | F | 731 | 14 | Persistent |
| PM026 | 7 | M | 448 | 8 35 | Persistent |
| PM040 | 12 | M | 388 | 21 4 | Persistent |
| PM071 | 5 | F | 549 | 3 63 | Persistent |
| PM049 | 3 | F | 463 | -- | No history |
| PM081 | 3 | F | 378 | -- | No history |

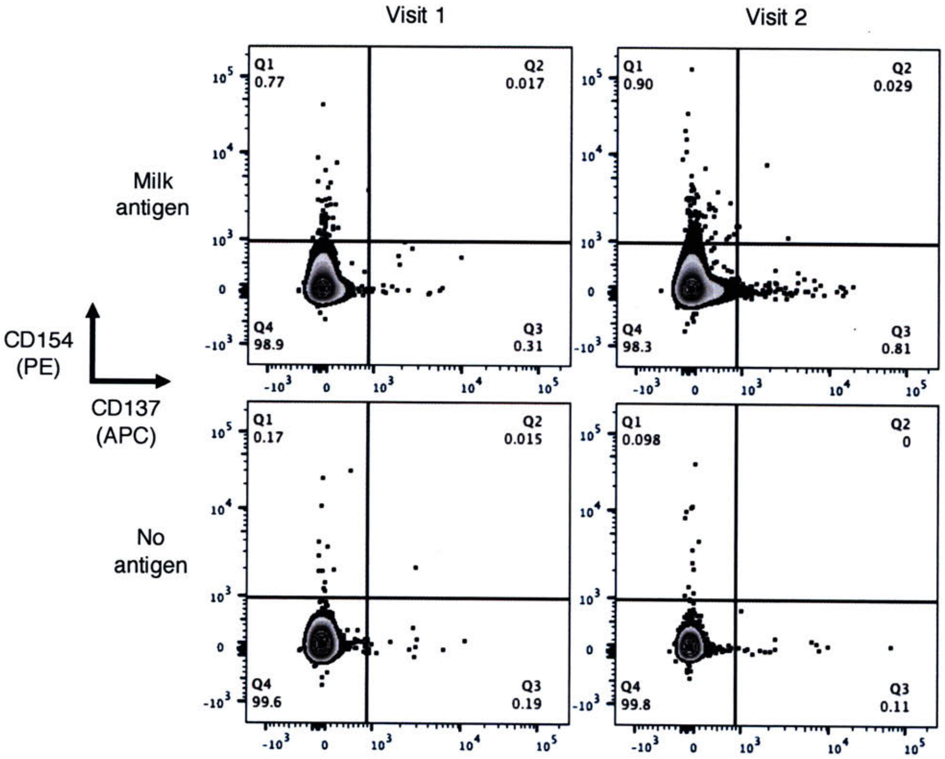


Figure 2-3. Activation of milk-stimulated CD4 memory T cells from a representative milk-allergic individual. Cells are from a "persistent" subject, and are pre-gated as lymphocytes, singlets, and live+CD3+CD4+CD45RA-.

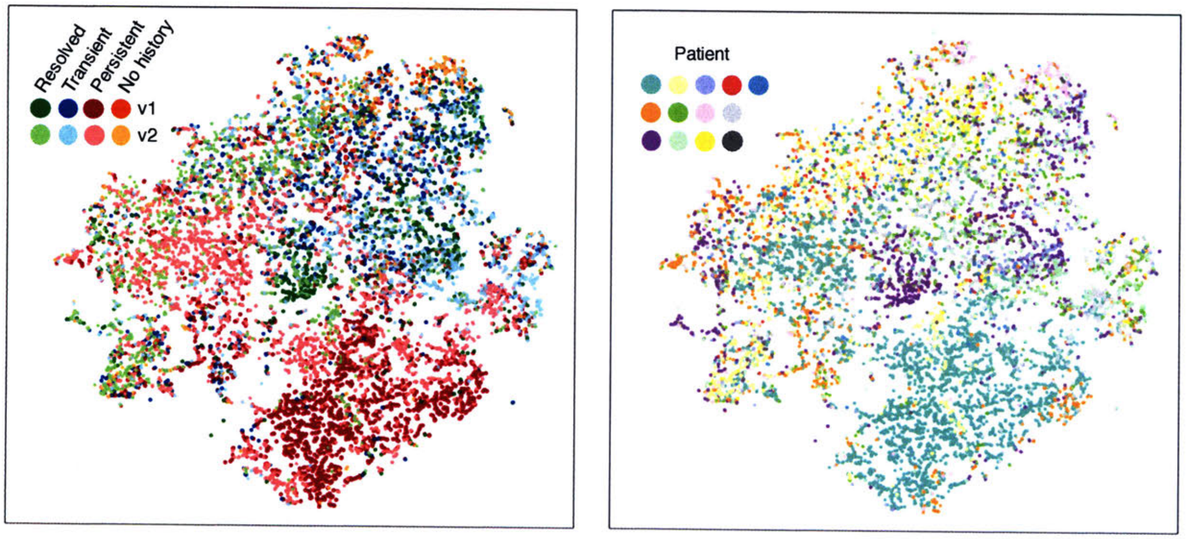


Figure 2-4. t-SNE visualization of CD154+ milk-reactive T-helper cells. Cells are colored by allergy status and visit (visit 1 = v1, visit 2 = v2) (left), and by patient (right).

### 2.3.2. Gene module discovery using sparse PCA

Next, to quantitatively assess what functional states were present among the CD154+ T cells, we employed our gene module discovery method, using the union of immune and variable genes as input features (see 'Methods') The method yielded several modules sorted by percent variance explained, as in PCA, of which we took the top 40 for further analysis based on the perceived "elbow" in the plot of percent variance explained (Figure 2-5) Of note, the percent variance explained by each sparse component differs dramatically from that seen in standard PCA This difference represents features particular to sparse PCA for one, each individual component, being sparse, explains quite little variance Additionally, sparse PCA being a numerical (rather than exact) solution, there is some stochasticity in the percent variance explained by each subsequent component, unlike standard PCA where the percent variance explained by each subsequent component is always a monotonically decreasing curve As a result, and based on the fact that the top gene modules represented dominant T cell programs and other sets of genes that concisely recapitulate known biology, percent variance explained seemed like an overly conservative way to gate modules in this setting, and so a visual "elbow" method was chosen instead Thus, we were able to reduce the feature space from hundreds of input genes to 40 modules, representing a significant benefit for multiple hypothesis testing, provided the gene modules were meaningful

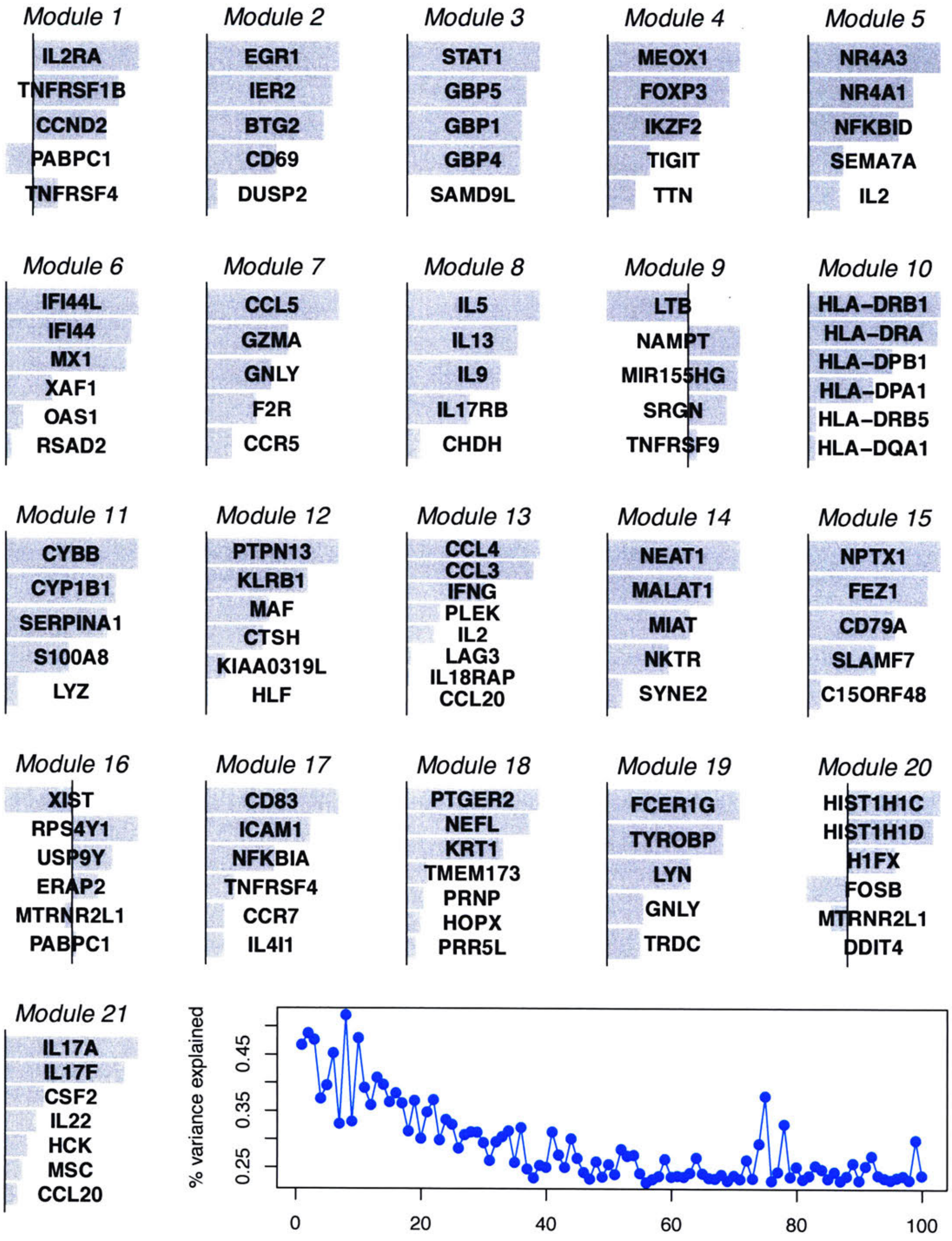


Figure 2-5. Gene modules identified in milk-reactive T cells from pediatric milk allergy patients. Top: Top 21 gene modules, sorted by percent variance explained. The magnitude of each bar represents the weight of the gene in the component. Bars to the left indicate negative weights. Bottom right: percent variance explained by each component.

To examine whether the gene modules were in fact meaningful, we looked closely at the top modules for overlap with known gene programs (Figure 2-5) Reassuringly, several modules corresponded to known T-cell functions, including TH1 function and interferon-induced activation (Module 3), and TH2 function (Module 8) We also observed modules with less obvious functions, such as Module 10, representing MHC II genes (the upregulation of which may indicate potent T-cell activation and signaling[24]), Module 15, a mix of neuronal and immune genes, or Module 16, which contains X-chromosome genes and whose expression cleanly separates cells on the gender of the individual (Figure A2-3)

### 2.3.3. Tracking longitudinal changes in milk-reactive T cells using gene modules

The gene module discovery method was successful at condensing genes into meaningful, known gene programs (and some less-known programs) Next, we wanted to see if the gene modules could improve interpretability in answering clinically focused research questions The central question of interest in this study was what changes occur over time in milk-specific T cells as the individuals age, and whether any of the changes are specific to patients who outgrow their milk allergy For visualizing specific modules, we selected the four modules corresponding to the major T cell functional states TH1, TH2, TH17, and Treg, and tracked their frequencies temporally in all patients with sufficient cell numbers (which, unfortunately, was only about half of the patients in our cohort Patients with no history of milk allergy, for example, did not have sufficient cells for this analysis) Unsurprisingly, each patient had their own distinct distribution of T cell states, and inter-patient variability was much higher than intra-patient variability between time points (Figure 2-6) Interestingly, we observed temporal changes in the form of TH1 induction in not just the transient patient, but in several of the persistent patients We speculate that this could be due either to a universal increase in milk-specific TH1 with age, or early prognostic changes in a subset of the persistent patients that might develop tolerance later on. There was insufficient data in this cohort to conclude which factor might be dominant.

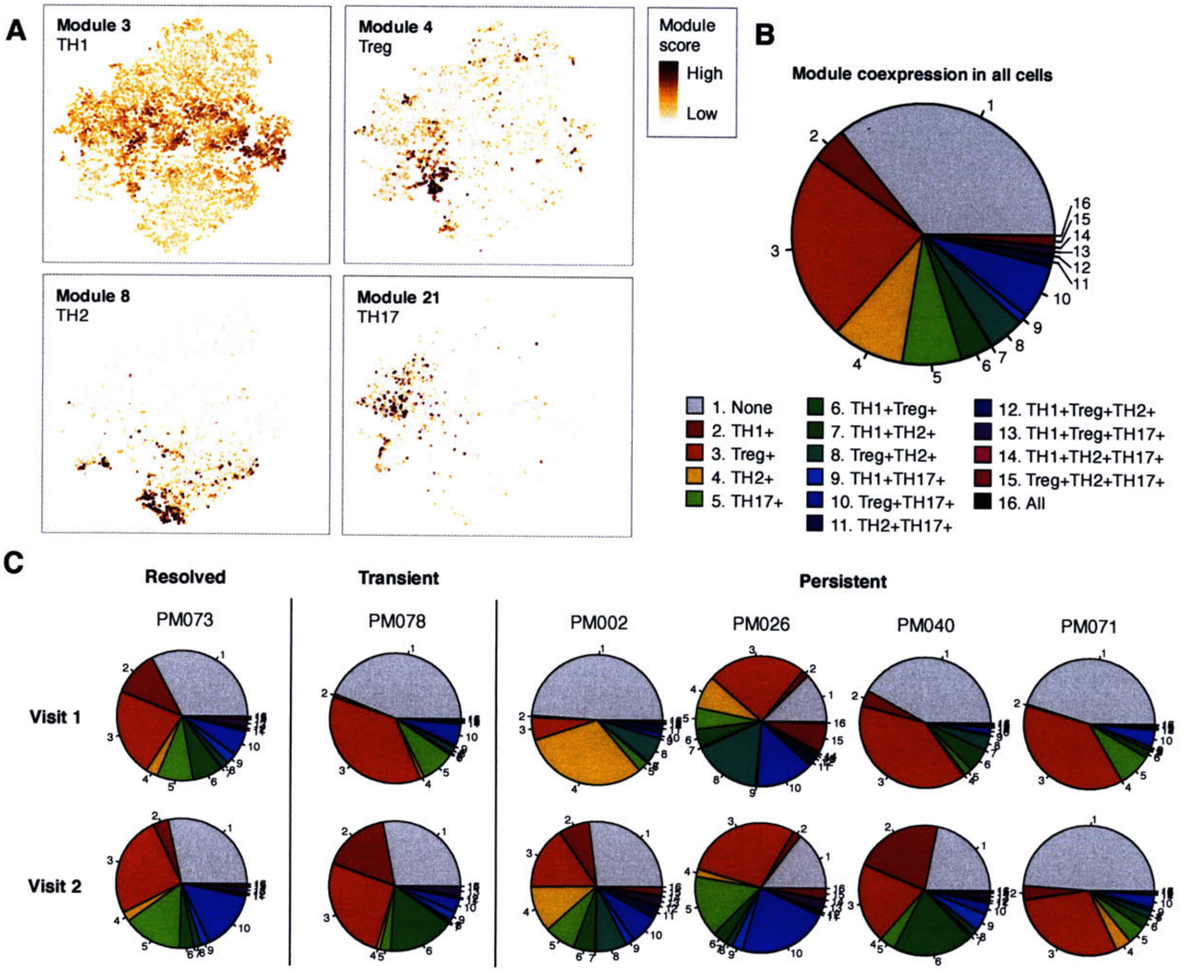


Figure 2-6. Functional T-cell modules identified in milk-reactive T cells and how they vary in frequency over time. A. Overlays of module scores onto the t-SNE visualization. B. Distribution of module co-expression in all cells. Module expression in all cells was gated using a Gaussian mixture model as described in the Methods. Every possible status of module expression in a cell for the four modules is represented with a different color. C. Distribution of module expression over time in individual patients, using the same color scheme for module combinations as in B. Patient samples with at least 100 cells at both time points were included.

To extend the analysis to all gene modules, not just the four major types, we performed differential expression between the two time-points in persistent milk allergy patients. This comparison was done specifically to look for time-specific changes in patients whose milk allergy status did not change. Differential expression was performed identically in both cases on all cells, using a nonparametric test since neither the genes nor the gene modules are normally distributed. Using gene modules, interpretability was aided by the unsupervised grouping of genes into coexpression patterns (Figure 2-7). We observed, for example, that Module 3 included genes that were almost all individually upregulated, but were easier to interpret when grouped together as a single program. Additionally, statistical significance (though irrelevantly significant in this case) was improved by gene modules due to the reduced burden of multiple hypothesis testing. From this analysis, we observed that the TH1 module is indeed upregulated at visit 2 in persistent patients, along with activation programs that are enriched at visit 1.

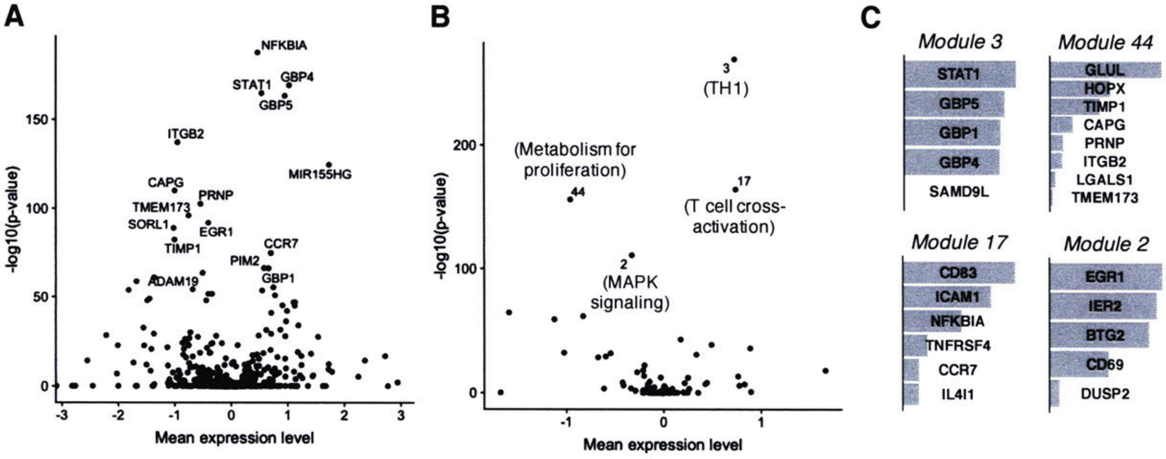


Figure 2-7. Differential expression analysis over time with genes and modules. A. Cells from the four persistent milk allergy donors were tested, with genes up at visit 1 on the left of each plot, and those up at visit 2 on the right. A linear regression was performed with time and patient as covariates. P-values were adjusted for multiple hypothesis testing using a Bonferroni correction. B. Same analysis as A, but with gene modules instead of genes. C. Gene loadings of the top-scoring modules from B.

## 2.4. Discussion

In this work, we show how a simple sparse PCA-based method can be applied to single-cell RNA sequencing data to condense genes into co-expressed gene modules This compression can be useful for smoothing, biological interpretation, and model parsimony when performing quantitative analyses of single-cell RNA-Seq data

This method has several features that may make it more or less attractive compared to other published methods First of all, the fundamental assumption governing any PCA-based approach is that the principal components (i e , gene modules) are linear combinations of features (i e , genes) This is both a drawback and an advantage, the assumption of linearity may not be accurate in all biological settings, but it is often a good approximation and is why PCA-based methods are transparent and easy to understand

Next, as mentioned previously, this particular method of sparse PCA includes the option to constrain all loadings to be positive In contexts where genes negatively regulate each other, allowing negative loadings can more accurately reflect the underlying biology However, having only positive loadings is sometimes easier to interpret and useful for downstream analyses, as scores can only be positive, making them interchangeable with normalized gene expression values in a lot of analyses Having the ability to turn this constraint on or off is a definite benefit

Finally, this method has the option to plug into other PCA-based approaches For example, contrastive PCA is an intriguing new technique that returns principal components enriched in a target dataset compared to a control dataset (by subtracting the covariance matrix of the control from the target and computing PCA on the difference)[25] This approach could, for example, return gene modules only variable in a disease dataset, not a healthy dataset The implementation of sparse PCA used for this work does not, unfortunately, use the covariance matrix as input, so this is not yet an option with the current approach But a user-friendly sparse

PCA algorithm that takes the covariance matrix as input could, in the future, allow for the discovery of contrastive gene modules.

One limitation of this approach is a small bias for highly-expressed genes being included in the gene modules, since these are more likely to have sufficient correlations with other genes (Figure A2-4B). While the genes are z-scored prior to running the sparse PCA, this did not entirely solve the problem of bias towards highly-expressed genes. We suspect that this is a problem common to most dimensionality-reduction methods, and it is certainly most egregious when using individual genes for differential expression analysis, as highly-expressed genes tend to have much lower p-values (Figure A2-4A). Nevertheless, this is a source of bias that we were hoping to address but that was not completely solved with this approach, and that still represents a challenge in single-cell RNA-Seq analysis.

## 2.5. Contributions and acknowledgements

The gene module discovery method presented here was adapted by me, with full credit to the sparse PCA implementation on which it is based. I am grateful to Daniela Witten, the author of the sparse PCA approach, for her insights on aspects of the method.

The pediatric milk allergy study was facilitated by several people, including the patients and their families, Sarita Patil who co-designed the single-cell study and evaluated allergy status for each individual based on patient charts, Jessica Savage and Alex Nee at Brigham and Women's Hospital, who recruited patients and collected blood samples, Jenny Che and Alaina Peterson at Sanofi who isolated PBMCs, Brandon Stone, Neal Smith, and Bert Ruiter who coordinated and performed the PBMC stimulation and flow sorting, Patrick Petrossian and Rachel Barry who helped with the single-cell sequencing experiments, and Wayne Shreffler, Chris Love, and Catherine Jones who helped design the single-cell study. Funding was provided by Sanofi, the Food Allergy Science Initiative at the Broad Institute, and in part by the Koch Institute Support (core) Grant P30-CA14051 from the National Cancer Institute

## 2.6. References

1 Kharchenko, P V , Silberstein, L & Scadden, D T Bayesian approach to single-cell differential expression analysis *Nat Methods* **11**, 740–742 (2014)

2 Grun, D , Kester, L & van Oudenaarden, A Validation of noise models for single-cell transcriptomics *Nat Methods* **11**, 637–640 (2014)

3 Munsky, B , Neuert, G & Oudenaarden, A van Using Gene Expression Noise to Understand Gene Regulation *Science* **336**, 183–187 (2012)

4 Zheng, G X Y *et al* Massively parallel digital transcriptional profiling of single cells *Nat Commun* **8**, 14049 (2017)

5 McDavid, A *et al* Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments *Bioinformatics* **29**, 461–467 (2013)

6 Risso, D , Perraudeau, F , Gribkova, S , Dudoit, S & Vert, J -P A general and flexible method for signal extraction from single-cell RNA-seq data *Nat Commun.* **9**, 284 (2018)

7 Finak, G *et al* MAST a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data *Genome Biol* **16**, 278 (2015)

8 DeTomaso, D & Yosef, N FastProject a tool for low-dimensional analysis of single-cell RNA-Seq data *BMC Bioinformatics* **17**, 315 (2016)

9 Soneson, C & Robinson, M D Bias, robustness and scalability in single-cell differential expression analysis *Nat Methods* **15**, 255–261 (2018)

10 van Dijk, D *et al* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion *Cell* **174**, 716-729 e27 (2018)

11 Gong, W , Kwak, I -Y , Pota, P , Koyano-Nakagawa, N & Garry, D J DrImpute imputing dropout events in single cell RNA sequencing data *BMC Bioinformatics* **19**, 220 (2018)

12 Li, W V & Li, J J An accurate and robust imputation method scImpute for single-cell RNA-seq data *Nat Commun* **9**, 997 (2018)

13 Kotliar, D *et al* Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq *eLife* **8**, e43803 (2019)

14 Shao, C & Hofer, T Robust classification of single-cell transcriptome data by nonnegative matrix factorization *Bioinformatics* **33**, 235–242 (2017)

15 Zhu, X , Ching, T , Pan, X , Weissman, S M & Garmire, L Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization *PeerJ* **5**, e2888 (2017)

16 Bielecki, P *et al* Skin inflammation driven by differentiation of quiescent tissue-resident ILCs into a spectrum of pathogenic effectors *bioRxiv* 461228 (2018) doi 10 1101/461228

17 Dey, K K , Hsiao, C J & Stephens, M Visualizing the structure of RNA-seq expression data using grade of membership models *PLOS Genet* **13**, e1006599 (2017)

18 Wıtten, D M , Tıbshıranı, R & Hastıe, T A penalızed matrıx decomposıtıon, wıth applıcatıons to sparse prıncıpal components and canonıcal correlatıon analysıs *Bıostat Oxf Engl* **10**, 515–534 (2009)

19 Wood, R A *et al* The natural hıstory of mılk allergy ın an observatıonal cohort *J Allergy Clın Immunol* **131**, 805–812 (2013)

20 Shreffler, W G , Wanıch, N , Moloney, M , Nowak-Wegrzyn, A & Sampson, H A Assocıatıon of allergen-specıfıc regulatory T cells wıth the onset of clınıcal tolerance to mılk proteın *J Allergy Clın Immunol* **123**, 43-52 e7 (2009)

21 Turcanu, V , Malekı, S J & Lack, G Characterızatıon of lymphocyte responses to peanuts ın normal chıldren, peanut-allergıc chıldren, and allergıc chıldren who acquıred tolerance to peanuts *J Clın Invest* **111**, 1065–1072 (2003)

22 Gıerahn, T M *et al* Seq-Well portable, low-cost RNA sequencıng of sıngle cells at hıgh throughput *Nat Methods* **14**, 395–398 (2017)

23 Macosko, E Z *et al* Hıghly Parallel Genome-wıde Expressıon Profılıng of Indıvıdual Cells Usıng Nanolıter Droplets *Cell* **161**, 1202–1214 (2015)

24 Hollıng, T M , Schooten, E & van Den Elsen, P J Functıon and regulatıon of MHC class II molecules ın T-lymphocytes of mıce and men *Hum Immunol* **65**, 282–290 (2004)

25 Abıd, A , Zhang, M J , Bagarıa, V K & Zou, J Explorıng patterns enrıched ın a dataset wıth contrastıve prıncıpal component analysıs *Nat Commun* **9**, 2134 (2018)

# 3. Analysis of multimodal single-cell data to identify a highly activated, clonotypically distinct subset in antigen-reactive T cells

Single cells can increasingly be profiled by multiple modalities simultaneously For example, both spatial and transcriptional information can now be collected on single cells in a tumor or in the brain[1,2], and gene expression and protein marker expression can be assayed in parallel using flow sorting and plate-based sequencing[3] Measuring multiple modalities of data invites exciting prospects for uncovering new biology, and significant challenges in integrating and analyzing the two types of data

A key goal with multimodal datasets is often to quantify the extent of association between the modalities, e g how much do specific cell types localize to certain spatial areas of a tumor, or how concordant are transcript and protein expression for a given gene Such analysis has been done qualitatively to great effect[2], but quantitative approaches to assess the degree and statistical significance of overlap are still lacking In this chapter, we show an analysis workflow for quantitatively evaluating overlap between two types of measurements using probability-based metrics

## 3.1. Motivation: Probing the relationship between T-cell receptor and transcriptional state in T-helper cells

Antigen-specific T cells are critically important in a wide variety of immune-mediated diseases, including infectious diseases, autoimmune disorders, allergy, and cancer  T cells function by mounting a response upon recognition of a specific peptide antigen bound to MHC I or II molecules  This recognition occurs through the T-cell receptor, or TCR, which undergoes VDJ recombination in a process similar to antibody development  The nature of the response can vary greatly, in T-helper cells, several major functional states have been identified including TH1 which is important in antiviral and antibacterial immunity, TH2 in anti-helminth immunity, TH17 in antifungal immunity, and Treg in maintaining peripheral tolerance to self-antigens  Additionally, almost all of these types have pathogenic functions, for example, TH2 cells play a role in allergy, TH17 cells in autoimmunity, and Treg cells in subversion of the anti-tumor immune response in cancer  Understanding what causes different T cell phenotypes to arise *in vivo*, and having the ability to modulate them, is therefore of immense scientific and clinical importance

T cell differentiation upon priming by an antigen-presenting cell (APC) is influenced by a combination of cues, including cytokine signaling, engagement between costimulatory receptors and their ligands on the surface of the APC, and signaling through the interaction between the TCR and peptide-MHC complex (pMHC)  The relative importance of these factors is still being understood, especially in humans, since most studies are performed in mice  The role of soluble factors such as cytokines in T-helper cell differentiation is relatively well-established, with elevated IL-12 and IFNα promoting TH1 responses, for example  What is less known is to what extent the TCR-pMHC binding event itself shapes the T-cell response  This effect, if it exists, is important to characterize for understanding basic T cell biology and modulating T cell fates in disease by selection of epitopes (a strategy that is already under way in some clinical settings[4,5])

Prior work on the effect of epitope on T-helper cell differentiation has been largely limited to highly-controlled *in vitro* settings, but the prevailing paradigm is that epitopes causing a higher-avidity interaction of the TCR-pMHC may lead to a TH1 response, those causing lower avidity may lead to a TH2 response, and that lowest-avidity interactions may lead to Treg responses[6,7,8,9] Epitopes qualities leading to other functional T cell states are less delineated It has been noted that these differences in fate could also be due to differences in the density of peptide presented on the surface of the antigen-presenting cell[10] (which could still be a function of epitope, as binding affinity of the peptide itself for MHC could influence surface density) Thus, various confounders exist, and moreover, it remains unclear to what extent these phenomena are relevant in humans in an actual disease response

Studying the link between epitope specificity and T cell fate in humans has seemed insurmountably difficult, as the fate of individual naive T cells cannot be tracked *in vivo*, and even the inference of a clonotype-phenotype relationship using bulk measurements is confounded by the high degree of T-cell heterogeneity It still remains impossible to track individual cells *in vivo*, but as a result of very recent innovations, cells of the same clonotype can now be identified and their functional states measured in parallel, at scale, using paired gene expression and TCR sequencing[11] From this setup, it is possible to identify the phenotypes of cells that likely recognize the same epitope (computational efforts to directly predict epitope from TCR sequence are still not usable for most antigens, but recent algorithms can reliably infer which TCR sequences are directed towards the same antigen[12,13]) This allows for the assessment of whether certain types of TCR sequences tend to be associated with the same T-cell fate, and the subsequent inference of the role that epitopes may play in shaping their differentiated state

To extract meaning from this paired T cell data, standardized analysis techniques do not yet exist An analysis pipeline thus needed to be designed to quantitatively assess overlap We developed a probability-based approach for the likelihood of cells expressing the same T cell functions, given the same or similar TCR sequences This strategy takes advantage of each cell as a separate probabilistic event Some of the concepts used in creating this analysis workflow were borrowed from information theory, an aspect of which focuses on assessing shared information between multiple measurements There are several relevant concepts from information theory, that with careful adaptation can be used for the analysis of paired single-cell data and for biological datasets in general

As a case study to investigate the association between TCR sequence and T cell fate in a well-defined antigen, we isolated CMV-activated memory T-helper cells from humans PBMCs and performed single-cell RNA sequencing via Seq-Well[14], as well as a paired TCR sequencing approach using TCR transcript pulldown and amplification[11] (The reason that TCR transcript sequence could not be gleaned directly from the RNA sequencing data was that most 3' single-cell RNA-Seq protocols do not sequence enough of the transcript to reach the variable regions of the TCR ) CMV was chosen as a common recall antigen likely to have a large number of circulating T cells in some individuals From an *in vitro* time-course in which T-helper cells were sorted after stimulation for 3h to 18h, we show a strong and evolving association between TCR and gene expression state, which suggests that epitope may influence the kinetics of T-cell responses Additionally, we identified a highly-activated, clonotypically distinct state, which preferentially contained a subset of the TCR sequences, some of which were biophysically highly similar, indicating that this state may be the result of encounters with specific epitopes

## 3.2. Methods

**PBMC culture with CMV antigens.** Cryopreserved PBMCs from two healthy donors with ELISPOT-validated responses to CMV were purchased from Cellular Technology Limited (Shaker Heights, Cleveland, OH) PBMCs were thawed and plated in a 24-well plate at 5M cells/1ml/well in AIM-V medium (Gibco) PBMCs were stimulated with CMVpp65 protein at 1 100 (Miltenyi 130-091-823), CMVpp65 peptide pool at 1ug/ml per peptide (Miltenyi 130-093-438), or PBS Cultures were incubated at 37C for 0h, 3h, 6h, 9h, 12h, or 18h before being harvested for staining and flow sorting 3h before harvesting each culture, PE anti-CD154 antibody (BD, clone TRAP1) was added at a dilution of 1 50, this step was omitted for the 0h cultures

**Flow-based enrichment for CMV-reactive T cells.** Cells were washed in PBS and stained for 25min at 4C with Zombie Violet viability dye (Biolegend), then with PacBlue anti-CD8 (clone SK1), APC anti-CD3 (clone UCHT1), APC-Cy7 anti-CD4 (clone RPA-T4), AF488 anti-CD45RA (clone HI100), PE anti-CD154 (clone TRAP1), and BV605 anti-CD137 (clone 4B4-1) All antibodies were purchased from Biolegend except for PE anti-CD154 which was from BD After staining, cells were washed with FACS buffer (PBS with 1% BSA and 1mM EDTA) and sorted on a BD FACS Aria instrument Cells were sequentially gated as lymphocytes, singlets, live cells, CD3+CD4+, and CD45RA-, and then CD154+ cells were sorted

**Single-cell transcriptome sequencing.** Sorted cells were immediately processed for single-cell RNA sequencing via the Seq-Well protocol[14] Briefly, cells were co-loaded into wells at approximately single-cell occupancy with poly(dT) beads and lysed to allow mRNA to hybridize onto the beads Beads were then pooled and mRNA was reverse-transcribed, PCR-amplified, and prepared for sequencing via the Nextera XT kit Libraries were sequenced on the Illumina Novaseq

**Paired single-cell TCR sequencing.** Paired TCR sequencing was performed according to Tu et al[11] Briefly, following cDNA amplification in the Seq-Well protocol, biotinylated capture probes for human TRAC and TRBC regions were annealed to cDNA Magnetic streptavidin beads were then used to enrich the bound TCR sequences, which were then further amplified using human V-region primers and prepared for sequencing using Nextera sequencing handles Libraries were sequenced on an Illumina MiSeq using 150bp-length reads

**TCR sequencing data preprocessing.** TCR sequencing reads were preprocessed according to Tu et al[11] In short, reads were mapped to TCRV and TCRJ IMGT reference sequences via IgBlast, and V and J calls with "strong plurality" (wherein the ratios of the most frequent V and J calls to the second most frequent calls were at least 0 6) were retained CDR3 sequences were called by identifying the 104-cysteine and 118-phenylalanine according to IMGT references and translating the amino acid sequences in between those residues Processed TCR sequences were then paired with the single-cell transcriptome data via the cell barcodes

**Transcriptome sequencing data preprocessing and visualization.** Raw read processing was performed as in Macosko et al[15] Briefly, sequencing reads were aligned to the hg38 human genome and counted to obtain a digital gene expression matrix of cells versus genes The matrix was filtered to exclude any cells with fewer than 500 detected genes or 1000 detected transcripts (UMIs) Counts were then normalized by cell library size and log2-transformed using the Seurat package in R, and transcriptomes were visualized using a two-dimensional t-SNE projection

**Surprisal analysis.** Surprisal was used to determine the tightness of association between a gene module and certain TCRb CDR3 sequences Surprisal is defined as $\ln(P/P_0)$, where P was the probability of two cells, drawn randomly (without replacement) from all cells sharing that TCRb, both expressing a gene module or state In order to account for different gene modules having different fractions of cells expressing, the probability is normalized by $P_0$, the probability of two cells, drawn randomly from all cells, both expressing the gene module $P_0$ represents the prior probability without the constraint of TCRb information Thus the entity $\ln(P/P_0)$ represents the gain in probability due to the constraint of shared TCR sequence

**TCR similarity.** TCRβ CDR3 amino acid sequences were clustered based on pairwise distances generated by the TCRdist method, published by Dash et al[13] Briefly, for two CDR3 sequences of the same length, each amino acid position was compared and a penalty was assessed for every mismatch The penalty, defined as $\min(4 - \text{BLOSUM62}[i, j], 4)$, was between 1 and 4 depending on the similarity of the two amino acid residues being compared, i and j. The overall distance between the 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. Highly similar pairs of TCRs were identified and binned based on their TCRdist distance.

## 3.3. Results

### 3.3.1. Overview of CMV-reactive T-helper cells

The goal of this study was to understand the kinetics of the T-cell response to a defined antigen and to observe if there was a clonotype-phenotype relationship at any time. To do this, we stimulated PBMCs from two CMV-reactive donors with whole CMV pp65 protein or peptide pool, and sorted CD154+ memory CD4 T cells at several time points between 0h and 18h. CD154 was chosen as a sensitive and specific marker for antigen-activated CD4 T cells[16]. Tables A3-1 and A3-2 contain donor information on demographics, CMV reactivity, and HLA background. The temporal pattern in CD154 expression, measured by flow cytometry, is shown in Figures 3-1, A3-1, and A3-2. CD154 expression peaked at very early time points (3h post-stimulation) in both donors and with both peptide and protein antigen, and declined moderately after that. Sorted cells were processed for single-cell RNA-Seq using Seq-Well and paired TCR sequencing using a TCR transcript pulldown technique suitable for 3'-barcoded RNA-Seq libraries[11].

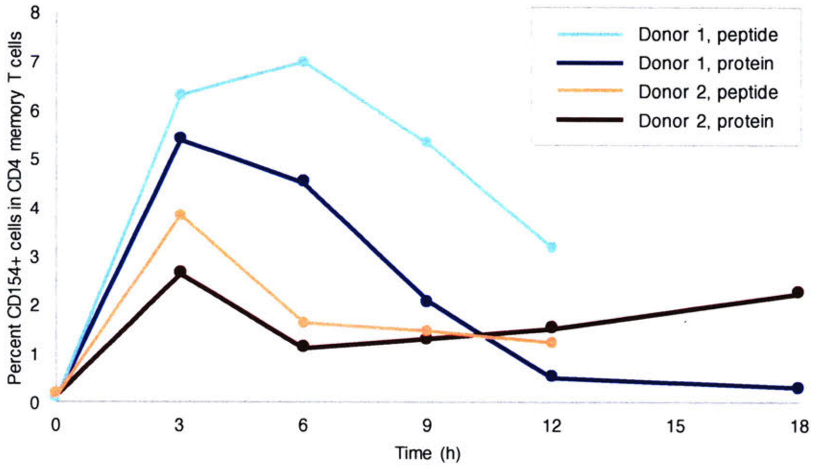


Figure 3-1. Percent CD154+ cells as a function of stimulation time. PBMCs were stimulated with CMVpp65 (whole protein or peptide pool) and sorted on the activation marker CD154. The frequency of CD154+ cells within the CD4 memory T cell compartment is shown.

A t-SNE visualization displays all transcriptomes that were recovered (Figure 3-2A). Qualitatively, there was a definite segregation of responses with time, and a moderate segregation of responses by donor. Paired TCR sequence recovery was efficient, with TCRb sequences recovered for 50% of cells and TCRa sequences recovered for 35%, and recovery was uniform (except for a small cluster of cells in the center of the plot, present in all samples, which could be resting, exhausted, or other cells that were not expressing TCR transcripts). Specific cell clusters were enriched in high clonal size (i.e., the number of cells sharing the same TCRb sequence). To look at the association between clonal expansion and specific T cell programs, we created scores for four major T-helper cell functions (TH1, TH2, TH17 and Treg), overlaid them onto the t-SNE, and observed that TH1 responses peaked early at 3h, and TH2 and TH17 responses peaked much later at 12h (Figure 3-2B, Figure A3-3). Together, these results qualitatively suggest a strong relationship between TH1 function, high clonal size, and early responses at 3h (regardless of peptide or protein antigen).

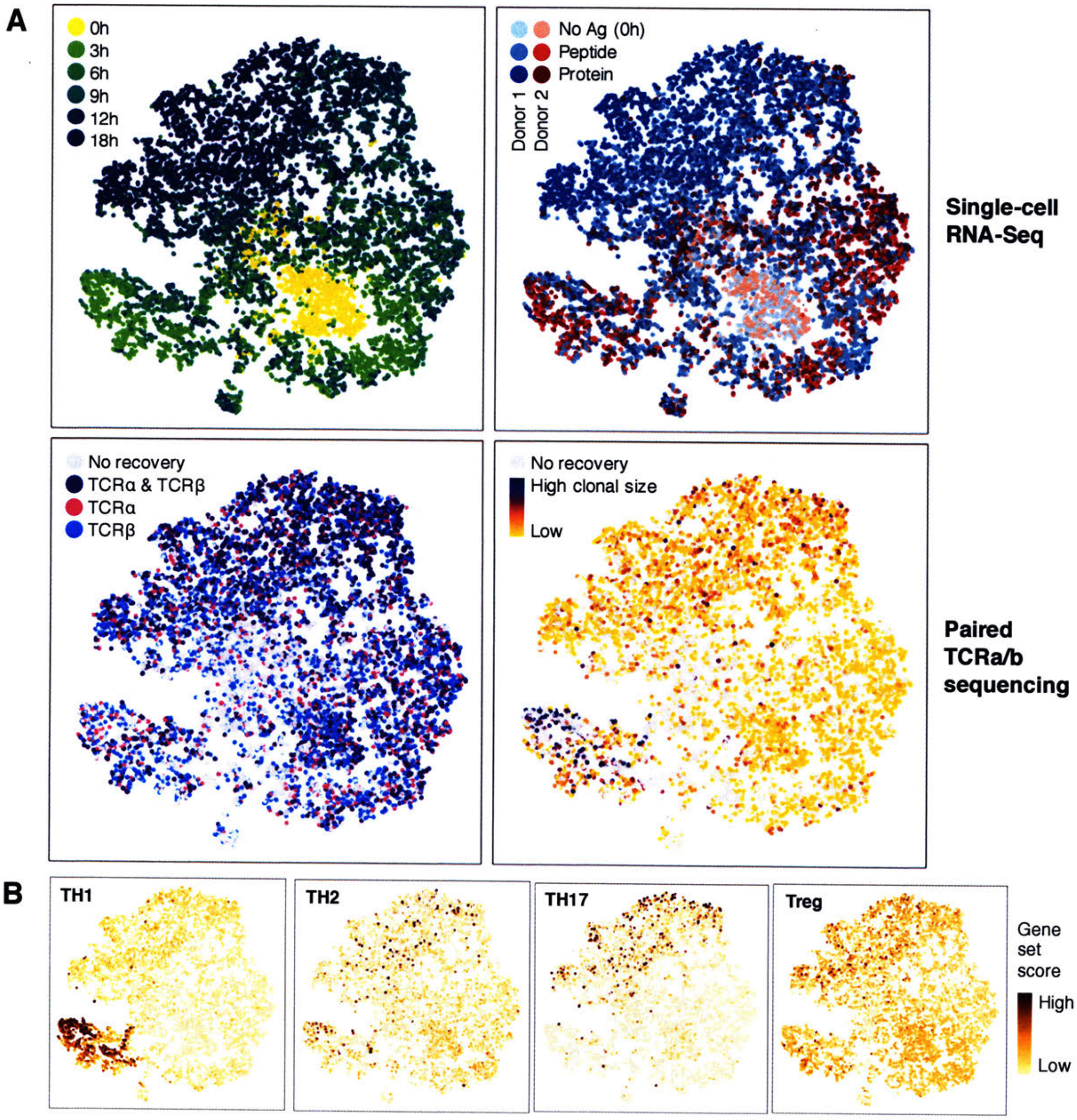


Figure 3-2. Visualizations of paired single-cell transcriptome and TCR sequence dataset. A) t-SNE plots of all single-cell transcriptomes, colored by *in vitro* stimulation time (top left), donor and stimulation antigen (top right), paired TCR sequence recovery (bottom left), or TCRb clonal size, defined as the number of cells with the same TRB CDR3 sequence (bottom right). B) Scores for key functional T-helper states overlaid onto the tSNE plot. Scores were calculated as the sum of expression levels of core transcriptional factors and cytokines for each state: IFNG, TNF, and TBX21 for TH1; IL13, IL5, IL9, IL4, and GATA3 for TH2; IL17A, IL17F, and RORC for TH17; and IL10, TGFB1, TGFB2, TGFB3, and FOXP3 for Treg.

To further explore and quantify the association between TCRb clonal size and time, we plotted the distribution of clonal sizes of all cells over time (Figure 3-3A). Mean clonal size was 1.7 cells at 0h, 9.0 at 3h, 5.7 at 6h, 2.9 at 9h, 2.4 at 12h, and 2.2 at 18h – as expected from the graph, the most expanded cells were found at 3h. This was also confirmed at the individual clonotype level; top expanded clonotypes were detected at most or all time points, but were at their highest frequency at 3h (Figure 3-3B). Interestingly, a few clones reached their highest frequency at 18h, suggesting that the response to CMV could be temporally bimodal, with fast- and slow-responding cells.

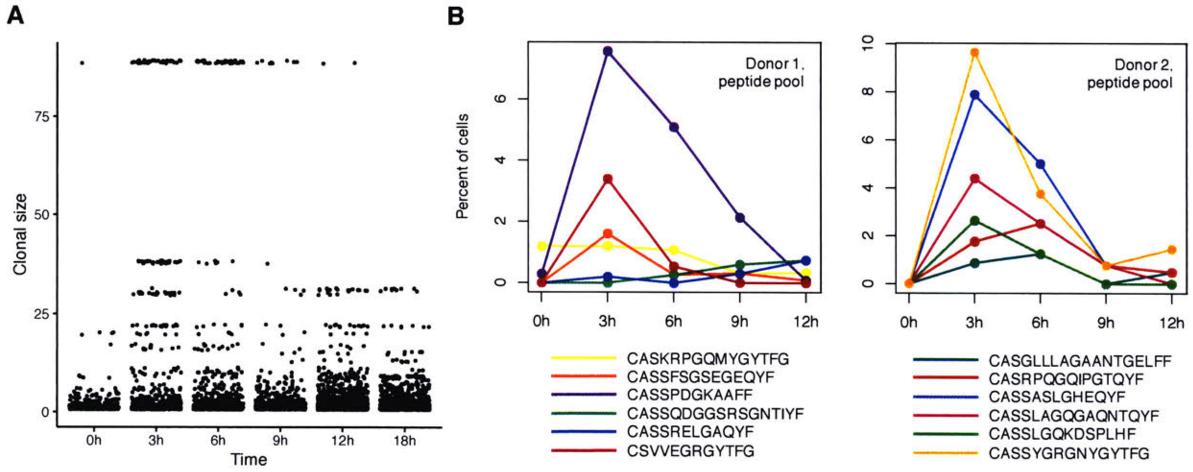
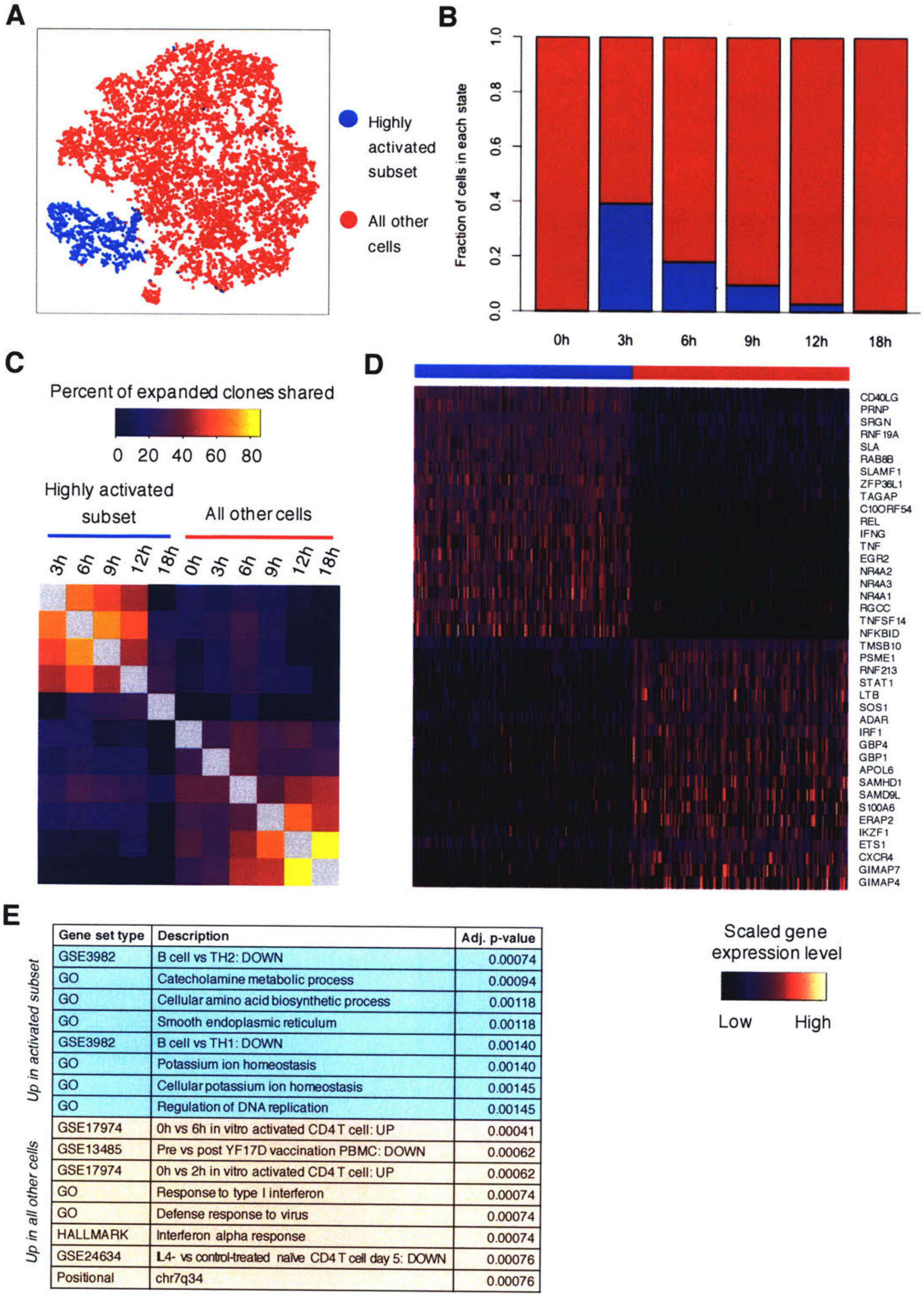


Figure 3-3. TCRb clonotypes peak in frequency early and are detected at most time points. A) Clonal size (defined as the number of cells sharing a TCRb sequence) of all cells detected at each time point. B) Temporal trajectories of six most expanded TCRb clonotypes within the peptide stimulation condition for each donor. Frequency of each TCRb clonotype among all cells with a TCR sequence is plotted on the y-axis. CDR3 sequence is shown in the legend.

### 3.3.2. Identification of a highly activated subset of cells with distinct TCR usage

Apart from TH1, TH2, TH17 and Treg programs, another transcriptional state was evident in the data that might support a bimodal paradigm of T cell response: a distinct subset of cells in the bottom-left of the t-SNE plot that seemed linked to high clonal size (Figure 3-4A; Figure 3-2A).



| | Gene set type | Description | Adj. p-value |
|---|---|---|---|
| Up in activated subset | GSE3982 | B cell vs TH2: DOWN | 0.00074 |
| | GO | Catecholamine metabolic process | 0.00094 |
| | GO | Cellular amino acid biosynthetic process | 0.00118 |
| | GO | Smooth endoplasmic reticulum | 0.00118 |
| | GSE3982 | B cell vs TH1: DOWN | 0.00140 |
| | GO | Potassium ion homeostasis | 0.00140 |
| | GO | Cellular potassium ion homeostasis | 0.00145 |
| | GO | Regulation of DNA replication | 0.00145 |
| Up in all other cells | GSE17974 | 0h vs 6h in vitro activated CD4 T cell: UP | 0.00041 |
| | GSE13485 | Pre vs post YF17D vaccination PBMC: DOWN | 0.00062 |
| | GSE17974 | 0h vs 2h in vitro activated CD4 T cell: UP | 0.00062 |
| | GO | Response to type I interferon | 0.00074 |
| | GO | Defense response to virus | 0.00074 |
| | HALLMARK | Interferon alpha response | 0.00074 |
| | GSE24634 | IL4- vs control-treated naïve CD4 T cell day 5: DOWN | 0.00076 |
| | Positional | chr7q34 | 0.00076 |

Figure 3-4. Identification of a highly-activated, clonotypically distinct subset of peanut-reactive T cells. A) tSNE visualization of all cells, colored by membership in highly activated cluster. B) Frequency of the highly activated subset, among all cells (from both donors and stimulation

conditions), as a function of time. D) Plot of top 20 most discriminating genes (using a ROC test) upregulated inside and outside the activated cluster. Analysis was restricted to the same six donors as before (out of eight). Each dot represents all cells from a donor either inside or outside the highly activated subset, with the color representing the mean gene expression level and the size of the dot representing the percent of cells expressing the gene. F) GSEA results for top eight gene sets enriched in cells inside and outside the highly activated subset. Adjusted p-value (adjusted by the GSEA permutation approach) is shown.

This subset was absent prior to antigen stimulation and peaked early in frequency, comprising about 40% of all cells recovered at 3h (Figure 3-4B). We dubbed it a “highly activated subset” based on its transcriptional signature, which included TCR signaling-induced transcription factor genes such as REL, NR4A1, NR4A2, and NR4A3, and activation markers CD40LG (i.e. CD154, the marker on which the cells were sorted) and IFNG (Figure 3-4D). Most notably, this subset was also clonotypically distinct, with an average of 65% sharing of TCR sequences between time points within the subset, but only 20% sharing of TCR sequences between cells inside and outside the subset, even at the same time point (Figure 3-4C). Gene pathways enriched in the subset were predominantly metabolic programs, suggesting that these cells might be a distinct set of fast-responding cells preparing to undergo proliferation.

### 3.3.3. Probability-based metrics of association between clonotype and transcriptional phenotype

The existence of a previously unknown, highly-activated, clonotypically distinct phenotype in T cells has significant implications, especially if it represents a universal phenomenon in antigen-specific T cells. We looked at whether the signature was present in response to other stimuli, and we found that it was in fact detected, but only with stimuli that signal through the TCR (peanut antigen, CMV, and anti-CD3/CD28 beads, but not PMA/ionomycin) (Figure A3-4). It appears, then, that the transcriptional state is TCR-dependent, we speculate that it could represent a fast-responding subset of cells (perhaps one that is older and closer to senescence)

To understand this state better, and to more deeply probe the relationship between clonotype and T cell function (e g TH1, TH2, etc), we developed a quantitative framework for assessing the overlap between the two modalities of data using probability-based metrics A central question was whether expression of gene modules such as TH1, or overall states such as the highly activated state, was confined to specific clonotypes, or whether it was expressed broadly across many clonotypes This would in turn tell us whether the gene program was likely to represent antigen-specific activation or bystander activation, with the assumption that bystander activation would be more likely to result in cells with random, rather than specific, TCRs To concisely quantify this likelihood, we employed an information theory metric known as ‘surprisal’, which measures the gain in information in one variable due to the knowledge of another variable In this case, we calculated the gain in probability that two cells would both score positively for a state (such as the highly activated subset), given that they both had the same TCR sequence, as a way to assess the tightness of association between specific TCR sequences and gene module expression We also calculated the average relative expansion of cells expressing each module, and we saw that in both donors, expansion and association with specific, not random, TCR sequences were highest in the highly activated state (Figure 3-4) In donor 1, for example, the surprisal for cells of the highly activated subset was about 2 8, representing a 16-fold increase in likelihood of cells both being in the highly activated state if they had the same TCRb sequence TH1 expression also had a moderate association with expansion and specific TCR sequences, but none of the other gene expression programs were as obviously associated with a highly clonal, likely antigen-specific, response

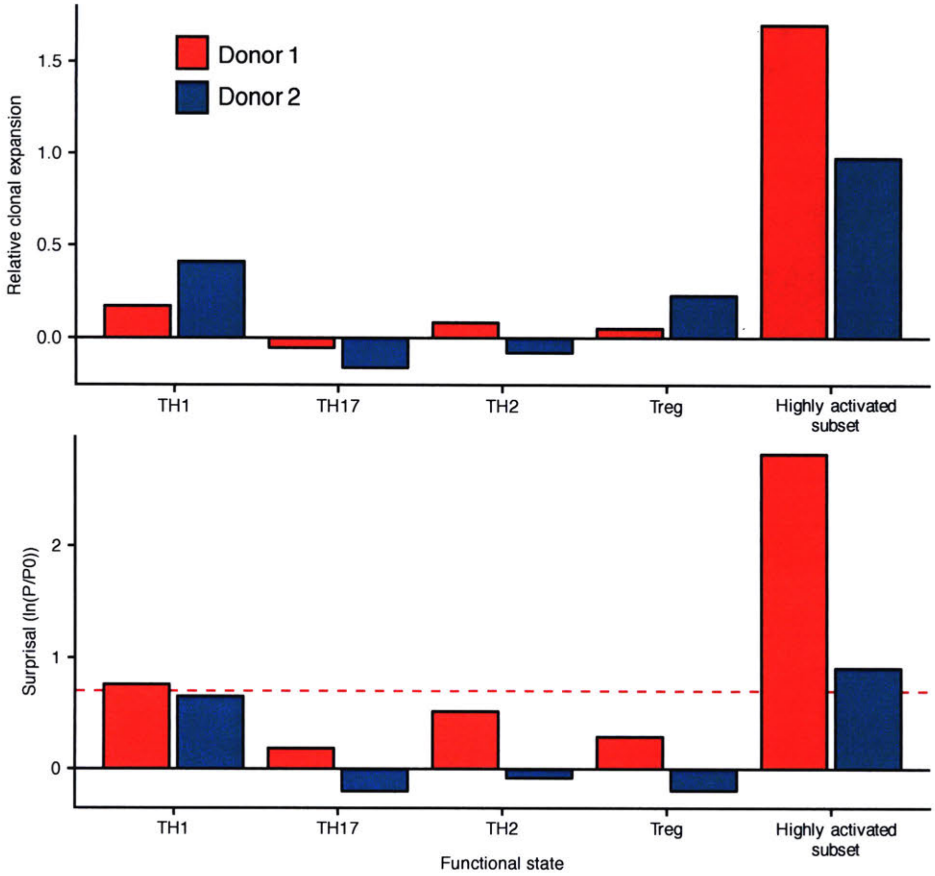


Figure 3-5. Association between TCRb sequence and various transcriptional states. Top: Relative clonal expansion of cells expressing each transcriptional program, defined as the z-score of the TCRb expansion number for all cells that are expressing the gene program, relative to the expansion number for all cells. Bottom: Surprisal for each program, defined as $\ln(P/P_0)$ where P = probability of two cells both expressing a program given that they have the same TCRb, and $P_0$ = probability of any two cells both expressing a program. Whether or not a cell was deemed to "express" a program was defined using a score for the T-helper programs (details in Methods), and for the highly activated subset, whether or not they were in the cluster of highly activated cells. The red dotted line represents a surprisal of ln(2), or a two-fold increase in likelihood due to the constraint of matching TCR sequence.

Having observed and quantified a strong clonotype-phenotype relationship, especially in the highly activated subset, two main explanatory hypotheses exist: 1) T cell fate is influenced by epitope, and cells of the same clone and phenotype represent independent priming events that converged onto the same effector phenotype; or 2) T cell fate may or may not be influenced by epitope, and cells of the same clone represent progeny of the same parent cell that expanded *in vivo* and retained the same functional phenotype in all of its daughter cells Both explanations are intrinsically interesting, but converging upon one would allow us to conclude whether or not epitope has a role in shaping the transcriptional states observed So as a final question, we attempted to discern between the two hypotheses using *similar* TCR sequences instead of exact matches We hypothesized that cells with very similar TCR sequences might be specific for the same epitope Thus, if we observed that cells with similar TCR sequences were co-localized inside or outside the highly activated state, we might conclude that the highly activated state arises from contact with specific epitopes (A positive result would implicate epitope in the differentiation of the highly activated state However, a negative result would not exclude epitope from playing a role, as the avidity of the TCR-pMHC link is known to be important, two highly similar TCR sequences could still have differing avidities for the same epitope, leading to different classes of responses )

To quantify TCR similarity, we used a published technique called TCRdist[13] The technique assigns every pair of TCR sequences a 'distance' based on amino acid similarity in the highly variable regions of the TCRs A distance of 1 to 4, for example, typically indicates a single amino acid mismatch, with the exact value reflecting the similarity of the two amino acids as per the BLOSUM62 matrix For cell pairs at a given TCR distance, we assessed the probability of both cells being co-located inside or outside the highly activated subset and averaged this result across all pairs of cells

The results of this TCR distance analysis by donor is quite striking (Figure 3-6) At a TCR pairwise distance of zero, which indicates an exact match in CDR3 sequence, cells are 90% and 92% likely to be co-localized in or out of the highly activated subset in Donor 1 and 2, respectively This reflects the high degree of association observed earlier for cells of the same clonotype (Figure 3-5) When we extended the analysis to include similar TCRs, we *still* observed an increased likelihood of cells to be co-localized, up to a certain TCR distance (4 in Donor 2, 16 in Donor 1). This suggests that epitope does indeed play a role in shaping the highly activated state, as cells likely to target the same epitope were more likely than expected to share the same state.

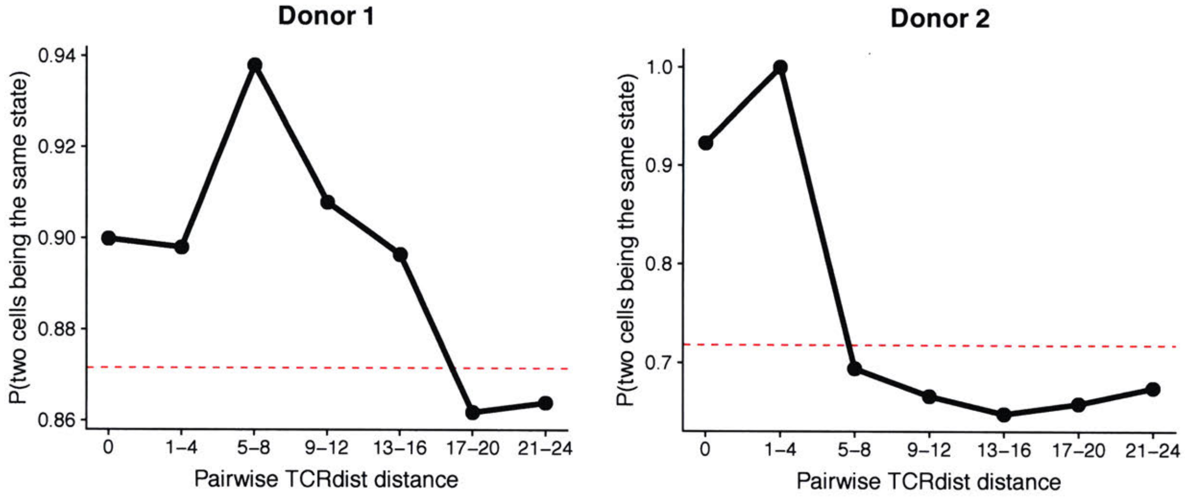


Figure 3-6. Colocalization of highly similar TCR sequences inside or outside the highly activated subset. TCRb CDR3 sequences were scored for their biophysical similarity in a pairwise fashion, using TCRdist (see Methods). Pairs of cells were then binned according to their TCRdist distance. A distance of 1-4, for example, typically represents a single amino acid substitution, with the value of the penalty representing the similarity of the amino acids. A distance of 5-8 typically represents two amino acid substitutions, and so on. The probability of both cells of the pair being co-located either inside or outside the highly activated subset described in Figure 3-4 was averaged across all pairs in the bin and plotted. The red dotted line represents the prior, or the probability of any two cells from the donor being co-located inside or outside the highly activated subset.

## 3.4. Discussion

Paired measurements in single cells are a powerful new way to interrogate biological samples; but quantitative analysis pipelines for multimodal single-cell data are needed, specifically those that harness the single cells as observations. In this study, we present a set of analysis methods for paired single-cell TCR and transcriptome sequencing data, using probability-based metrics.

As a case study, we measured CMV-reactive T-helper cells in PBMCs from two adults and investigated the link between clonotype and phenotype, to understand to what extent epitope binding might shape the T-cell response We found a highly activated subset of cells that was clonotypically distinct, and we showed that this state was associated with specific, not random, TCR sequences Finally, we observed that extremely similar TCR sequences were likely to have the same status, suggesting that epitope might be playing a role in causing the highly activated state

An important question is exactly what this activated transcriptional state represents and whether it aligns with any known phenomena in the literature Two main hypotheses presented themselves 1) The state represents true TCR-induced activation, as opposed to bystander activation This explanation is supported with the gene signatures of TCR-activated transcriptional factors, but it is countered by the fact that highly clonal cells are observed outside the highly activated state 2) The state represents T-follicular helper (Tfh) cells or another fast-responding subset of cells This explanation is also plausible, and upon further investigation of relevant gene signatures, we found that Tfh cells are in fact preferentially enriched by CD154+ selection[17] and that these cells display costimulatory markers such as ICOS, TNFRSF9, TNFRSF4, and PDCD1[18], which we did observe in the highly activated cells We did not observe the Tfh transcription factor Bcl6, however, this is in line with what some others have shown for Tfh cells in the blood (as opposed to in the germinal center)[18,19] Our prevailing hypothesis, then, is that these highly activated cells may represent blood Tfh cells that are primed to respond to antigen stimulation with strong TCR signaling, costimulation, and other pathways Regardless of the hypothesis, the implications of such an activated subset are important for T cell biology and for modulating T cell fate

In addition to the analysis pipeline presented here, it is worth mentioning that many other methods are emerging for the purpose of analyzing multimodal single-cell data Given the vast diversity in goals and types of data generated in single-cell studies, this is unequivocally a good thing for single-cell research  So far, three main classes of techniques have emerged  For sake of simplicity, assuming a paired single-cell workflow in which the first modality represents high-dimensional single-cell data such as RNA-Seq or CyToF, most techniques can be categorized as follows  1) In the case where the second modality represents a categorical feature with few bins (e g  tissue source or methylation status at a single locus), it is often easiest to simply perform statistical tests with this feature as an independent variable[20,21,22]  This allows for direct and intuitive comparison of cell states between bins, however, it is only suitable in the case of one or a few features, and coarse-grain binning of the features must be an appropriate treatment  2) In the case where the second modality is continuous (e g  spatial information or protein expression), statistical learning methods like regression are often used[2,23]  Such methods are simple, intuitive, and powerful, but they tend to ignore even strong associations in subsets of cells, if these cells are not manually selected  3) In the case where it is desirable that visualization, clustering, and identification of cell subsets are done using all modalities of data, new techniques are emerging to successfully integrate the modalities together using dimensionality-reduction and dataset-weighting strategies[24]  This approach is still challenging, though, because it can be unclear how much each modality is weighted, and thus how to interpret the results

The analysis pipeline presented in this chapter is fairly unique among existing methods, as it uses entities from information theory to measure association using probability, rather than using correlation or statistical significance  The uniqueness of this pipeline resulted in response to specific challenges in integrating paired gene expression and TCR sequencing data, the latter of which represents a modality consisting of categorical features with *many* bins, which is not easily analyzed by any of the above methods  Although this framework was developed for a specific use case, it should be generalizable to any multimodal single-cell data which can be discretized into bins  Additionally, probabilities can provide highly tangible and quantitative interpretations compared to correlations or p-values  for example, one can assess *how much more likely* it is for a gene to be expressed under condition A than condition B, a feature that can be particularly desirable in certain settings  Overall, this analysis framework should be a useful contribution to the increasing repertoire of analysis techniques available for multimodal single-cell studies  At the remarkable pace at which new experimental innovations emerge each year, diverse analysis strategies will continue to be required to maximize the utility of single-cell data

## 3.5. Contributions and acknowledgements

This project was inspired by an experiment done with my undergraduate researcher, Julia Ginder, in which we measured the peanut-reactive T cell response in peanut-allergic PBMCs over time and observed intriguing temporal gene expression patterns in expanded T cell clones  To more accurately draw conclusions about what these trends represented, though, we needed to repeat the experiment with a well-defined single antigen and a robust analysis pipeline, as a result, this project with CMV antigen came into being

Paired TCR sequencing was developed and performed by Andy Tu  Seq-Well was developed by Todd Gierahn, who also provided intellectual guidance on this project  Seq-Well experiments were performed with Patrick Petrossian (for whose help at odd hours of the night I am especially grateful)  Sequencing was done by the MIT Bio-Micro center and by the Broad Walk-Up sequencing facility  I am also grateful to Glenn Paradis and the Koch Institute flow core  This work was supported in part by the Koch Institute Support (core) Grant P30-CA14051 from the National Cancer Institute

## 3.6. References

1 Ståhl, P L *et al* Vısualızatıon and analysıs of gene expressıon ın tıssue sectıons by spatıal transcrıptomıcs *Scıence* **353**, 78–82 (2016)

2 Rodrıques, S G *et al* Slıde-seq A scalable technology for measurıng genome-wıde expressıon at hıgh spatıal resolutıon *Scıence* **363**, 1463–1467 (2019)

3 Pıcellı, S *et al* Smart-seq2 for sensıtıve full-length transcrıptome profılıng ın sıngle cells *Nat. Methods* **10**, 1096–1098 (2013)

4 Oldfıeld, W , Larché, M & Kay, A Effect of T-cell peptıdes derıved from Fel d 1 on allergıc reactıons and cytokıne productıon ın patıents sensıtıve to cats a randomısed controlled trıal *The Lancet* **360**, 47–53 (2002)

5 Muller, U *et al* Successful ımmunotherapy wıth T-cell epıtope peptıdes of bee venom phospholıpase A2 ınduces specıfıc T-cell anergy ın patıents allergıc to bee venom *J Allergy Clın Immunol.* **101**, 747–754 (1998)

6 Constant, S , Pfeıffer, C , Woodard, A , Pasqualını, T & Bottomly, K Extent of T cell receptor lıgatıon can determıne the functıonal dıfferentıatıon of naıve CD4+ T cells *J. Exp Med.* **182**, 1591–1596 (1995)

7 Blander, J M , Sant'Angelo, D B , Bottomly, K & Janeway, C A Alteratıon at a sıngle amıno acıd resıdue ın the T cell receptor alpha chaın complementarıty determınıng regıon 2 changes the dıfferentıatıon of naıve CD4 T cells ın response to antıgen from T helper cell type **1** (Th1) to Th2 *J Exp. Med* **191**, 2065–2074 (2000)

8 Tubo, N J *et al* Sıngle naıve CD4+ T cells from a dıverse repertoıre produce dıfferent effector cell types durıng ınfectıon *Cell* **153**, 785–796 (2013)

9 Gebe, J A *et al* Low-avıdıty recognıtıon by CD4+ T cells dırected to self-antıgens *Eur. J Immunol.* **33**, 1409–1417 (2003)

10 Grakouı, A , Donermeyer, D L , Kanagawa, O , Murphy, K M & Allen, P M TCR-Independent Pathways Medıate the Effects of Antıgen Dose and Altered Peptıde Lıgands on Th Cell Polarızatıon *J Immunol* **162**, 1923–1930 (1999)

11 Ang A Tu *et al* Recovery of paıred T cell receptors from massıvely-parallel 3' sıngle-cell RNA-Seq lıbrarıes reveals clonotypıc responses among antıgen-actıvated T cells *In revısıon*

12 Glanvılle, J *et al* Identıfyıng specıfıcıty groups ın the T cell receptor repertoıre *Nature* **547**, 94–98 (2017)

13 Dash, P *et al* Quantıfıable predıctıve features defıne epıtope-specıfıc T cell receptor repertoıres *Nature* **547**, 89–93 (2017)

14 Gıerahn, T M *et al.* Seq-Well portable, low-cost RNA sequencıng of sıngle cells at hıgh throughput *Nat Methods* **14**, 395–398 (2017)

15 Macosko, E Z *et al* Hıghly Parallel Genome-wıde Expressıon Profılıng of Indıvıdual Cells Usıng Nanolıter Droplets *Cell* **161**, 1202–1214 (2015)

16 Chattopadhyay, P K , Yu, J & Roederer, M Lıve-cell assay to detect antıgen-specıfıc CD4 + T-cell responses by CD154 expressıon *Nat Protoc* **1**, 1 (2006)

17 Jıang, W *et al* Identıfıcatıon of murıne antıgen-specıfıc T follıcular helper cells usıng an actıvatıon-ınduced marker assay *J Immunol Methods* **467**, 48–57 (2019)

18 Schmitt, N , Bentebibel, S -E & Ueno, H Phenotype and functions of memory Tfh cells in human blood *Trends Immunol.* **35**, 436–442 (2014)

19 Gowthaman, U *et al* Identification of a T follicular helper cell subset that drives anaphylactic IgE *Science* eaaw6433 (2019) doi 10 1126/science aaw6433

20 Lavin, Y *et al* Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses *Cell* **169**, 750-765 e17 (2017)

21 Kurtulus, S *et al* Checkpoint Blockade Immunotherapy Induces Dynamic Changes in PD-1–CD8+ Tumor-Infiltrating T Cells *Immunity* **50**, 181-194 e6 (2019)

22 Cheow, L F *et al* Single-cell multimodal profiling reveals cellular epigenetic heterogeneity *Nat Methods* **13**, 833–836 (2016)

23 Pliner, H A *et al* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data *Mol Cell* **71**, 858-871 e8 (2018)

24 Welch, J D *et al* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity *Cell* **177**, 1873-1887 e17 (2019)

# 4. Transient suppression, but not deletion, of distinct subsets of TH2 clonotypes during peanut oral immunotherapy

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

Food allergy affects an estimated 8% of children in the US[1], and its reported prevalence and severity are increasing globally[2] Oral immunotherapy (OIT) is an experimental strategy for inducing tolerance to food allergens, but it is ineffective at producing truly sustained, rather than transient, desensitization in most patients[3] Cellular mechanisms that drive therapy-induced changes, and predictive tools for clinical outcome, remain largely unresolved We assessed how populations and clonotypes of T-helper cells were altered during peanut OIT and which of these subsets associated with clinical outcomes in 12 peanut-allergic patients Using single-cell RNA sequencing and paired TCR$\alpha/\beta$ sequencing of peanut-reactive CD4 memory T cells, coupled with an approach for gene module discovery, we observed several distinct functional states among clonally expanded peanut-reactive T cells, including TH2 cells, TH17 cells, and Treg cells TH2 and TH17 programs were transiently suppressed in individual clonotypes during OIT Additionally, TH17 expression was upregulated in T cells from patients with poor clinical outcome These results highlight the impermanence of OIT-induced changes within CD4 T-cell clonotypes, which may reflect non-durable programming

## 4.1. Motivation

Food allergy is an immune hypersensitivity disease characterized by allergen-specific TH2 cells and (in the case of type I hypersensitivity) the production of allergen-specific IgE antibodies These antibodies bind to FcεRI receptors on effector cells such as mast cells, basophils, and eosinophils, and can be cross-linked in the presence of allergen, leading to cellular degranulation and the systemic release of histamine and other mediators[2] The resulting symptoms can range from mild to life-threatening

No FDA-approved treatment for food allergy exists, representing a serious unmet need Allergen-specific immunotherapy is an experimental option with various modalities, the most common being oral immunotherapy (OIT) OIT consists of daily exposure to allergen (e g , peanut flour) by the oral route that is gradually increased over time to induce clinical tolerance Compared to other modalities such as sublingual and epicutaneous immunotherapy, OIT tends to induce higher maximum tolerated doses but also a higher incidence of adverse events[3] Even so, the sustained efficacy of OIT is low, while 80-85% of patients can achieve desensitization (a loss in clinical reactivity with regular consumption of allergen), most food-allergic patients do not attain sustained unresponsiveness (maintenance of tolerance without the need for continued allergen consumption)[3,4]

It is unclear why this state of desensitization is often transient, and more broadly, how OIT causes immune tolerance in the first place, but several observations have been made about immune changes during immunotherapy OIT induces changes in allergen-specific serum IgE titers – typically a rapid increase upon initial allergen exposure, followed by a slow decrease to pre-baseline levels – and high baseline levels of IgE are predictive of poor outcome[5] Additionally, allergen-specific IgG4 has been shown to increase over time and may exert a protective function by competing for allergen binding and disrupting IgE cross-linking and signaling[6] These B-cell changes may be directed by T-helper cells, which play an important role in the maintenance of B cells and antibody levels in the periphery. Understanding allergen-specific T-helper cell changes, therefore, is critical to unraveling cellular mechanisms in OIT. OIT-induced changes in T-helper cells, however, are much less widely reported and agreed upon, in part due to the difficulty in isolating and tracking rare allergen-specific T cells longitudinally. Frequencies of circulating allergen-specific TH2 cells, and their expression of TH2 cytokines, may decrease[7] or be suppressed by anergic gene programs[8], and patients who achieve sustained unresponsiveness may have a higher frequency of Tregs post-treatment[9]. Whether T cells are predominantly altered by clonal anergy, deletion, Treg-based suppression, or other factors, is still unresolved. Clarifying T-cell changes in OIT at high resolution could be important in explaining observations in B cells and the associated variability in patient outcomes.

## 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 studies8, 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, seven 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 frozen in FBS with 10% DMSO  After the study was completed, PBMCs from a patient at all time points were simultaneously thawed, washed with PBS, and cultured in AIM-V medium (Gibco) with 100 $\mu$g/ml peanut extract for 20h  (Peanut 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 1 100 dilution for the last 3h After harvesting, the cells were labeled with anti-CD3-AF700 (BD Biosciences, UCHT1), anti-CD4-APC-Cy7 (BD Biosciences, RPA-T4), anti-CD45RA-FITC (BD Biosciences, HI100), anti-CD154-PE (BD Biosciences, TRAP1), anti-CD69-AF647 (BioLegend, FN50), anti-CD137 APC (clone 4B4-1), and Live/Dead Fixable Violet stain (Thermo Fisher, cat no L34955) Cells were then sorted on a FACS Aria II 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")

**Single-cell RNA-Seq.** Sorted subsets of CD4 memory T cells were processed for single-cell RNA sequencing using the Seq-Well platform as previously described[12] A portion of each cDNA library was reserved for paired TCRa/b enrichment. The rest was barcoded and amplified using the Nextera XT kit and sequenced on the Illumina NovaSeq

Raw read processing was performed as in Macosko et al[18] Briefly, sequencing reads were aligned to the 'hg38' reference human genome, collapsed by unique molecular identifier (UMI), and counted to obtain a digital gene expression matrix of cells versus genes These counts were then filtered to exclude any cells with fewer than 500 genes or 1000 UMIs and normalized by library size per cell and a log2 transformation

**Paired single-cell TCRa/b sequencing.** Paired TCR sequencing was performed according to Tu et al[13] Briefly, following cDNA amplification, biotinylated capture probes for human TRAC and TRBC regions were annealed to cDNA Magnetic streptavidin beads were used to enrich the bound TCR sequences, which were then further amplified using human V-region primers and prepared for sequencing using Nextera sequencing handles Libraries were sequenced on an Illumina MiSeq using 150bp-length reads

TCR sequencing reads were preprocessed according to Tu et al[13] In short, reads were mapped to TCRV and TCRJ IMGT reference sequences via IgBlast, and V and J calls with "strong plurality" (wherein the ratios of the most frequent V and J calls to the second most frequent calls were at least 0 6) were retained CDR3 sequences were called by identifying the 104-cysteine and 118-phenylalanine according to IMGT references and translating the amino acid sequences in between those residues Processed TCR sequences were then paired with the single-cell transcriptome data via the cell barcodes

**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 (uniform manifold approximation and projection)[19] Clustering was performed on the top 10 principal components using the Louvain graph-clustering method

**Gene module discovery.** Coexpressed gene modules were generated based on a sparse PCA approach described by Witten et al[14] and implemented in the R package "PMA" This method employed an L1 norm penalty to reduce and eliminate gene loadings that contributed less 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 prevent the results from being dominated by a few samples and to decrease computational time Genes were filtered down to the union of immune genes (as defined by the sets of gene lists available on ImmPort at https //www immport org/shared/genelists) and the variable genes in the dataset using the 'var genes' command in the R package "Seurat" 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 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

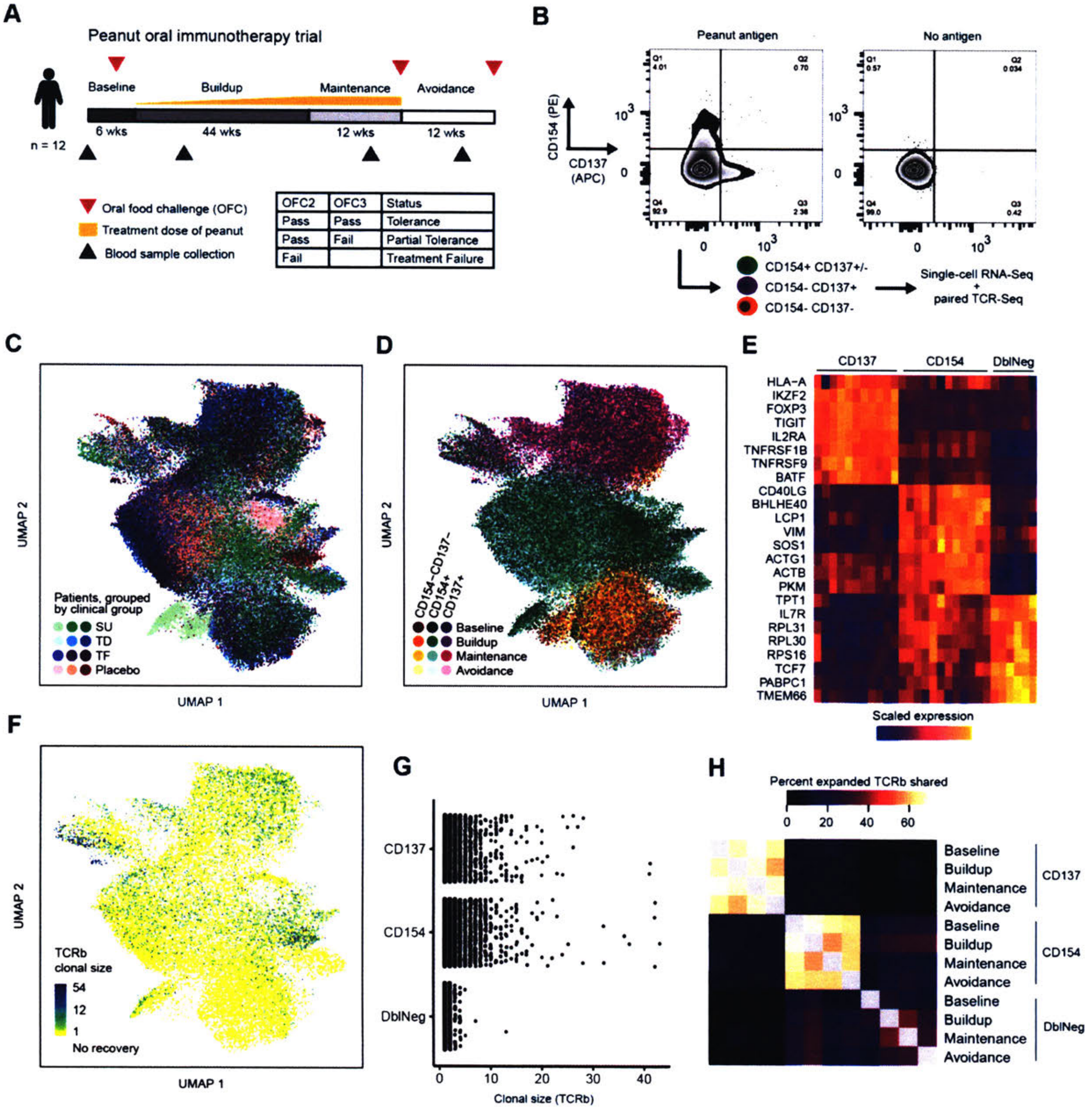
**Distance analysis of TCR sequences.** Pairwise similarity of TCRb CDR3 sequences was evaluated using an adapted version of the TCRdist method published by Dash et al15 Briefly, for two TCRb CDR3 amino acid sequences of the same length, each residue position was compared and a penalty was assessed for 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 0, 1-4, 5-8, 9-12, 13-16, 17-20, and 21-24

**Probability-based association between TCR and gene expression.** Probability-based analysis was used to determine the tightness of association between a categorical transcriptional feature (such as cluster or status of gene module expression) and TCRb 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 a TCRb CDR3 sequence, both expressing a gene module 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 cluster $P_0$ represents the prior probability without the constraint of TCRb 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

## 4.3. Results

### 4.3.1. CD154+ and CD137+ peanut-reactive T cells have distinct transcriptional states and TCR repertoires

To examine the dynamics of the T cell response induced by OIT, we profiled peanut-reactive T-helper cells longitudinally from patients undergoing peanut OIT (Figure 4-1A)  The 40-patient trial (NCT01750879) consisted of daily ingestion of peanut flour or placebo  The oral dose increased every two weeks for 44 weeks (buildup), and was then held at a maximum dose for 12 weeks (maintenance)  This phase was followed by a 12-week period of strict avoidance to assess the durability of desensitization  Clinical outcomes were evaluated by oral food challenges (OFCs) at the end of the maintenance and avoidance phases, and were defined as “tolerance” (passing both food challenges), “partial tolerance” (passing the maintenance challenge but failing the avoidance challenge), and “treatment failure” (failing the maintenance challenge)  Peripheral blood samples were collected longitudinally for single-cell analysis from 12 of the enrolled patients (3 each of tolerance, partial tolerance, and treatment failure outcomes, as well as 3 placebo patients, all of whom had treatment failure outcomes)



| OFC2 | OFC3 | Status |
|---|---|---|
| Pass | Pass | Tolerance |
| Pass | Fail | Partial Tolerance |
| Fail | | Treatment Failure |

Figure 4-1. CD154+ and CD137+ peanut-reactive T cells recovered from oral immunotherapy patients have distinct transcriptional signatures and TCR repertoires. **a**, Peanut OIT design and definition of outcomes. Clinical outcomes were defined as shown based on two oral food challenges towards the end of the study. Samples from 12 of the patients, three treatment patients with each outcome plus three placebo patients, were selected for single-cell profiling. **b**, Peanut-reactive memory T-helper cells enriched by FACS, following stimulation with peanut antigen (left). Stimulation with no antigen (right) is shown for comparison. Cells are pre-gated on live singlet CD3+CD4+CD45RA- cells, and sorted as CD154+CD137+/-, CD154-CD137+, or CD154-CD137- ("DblNeg"). **c**, Two-dimensional UMAP visualization of all single-cell transcriptomes, colored by sorted subset and time point. TO (tolerance), PT (partial tolerance), TF (treatment failure), and

PL (placebo) refer to the four clinical groups **d**, UMAP visualization colored by patient (each triad of colors represents a clinical group) **e**, Differentially expressed genes upregulated in each sorted subset, identified using a ROC test Rows represent z-scored gene expression values, and columns represent average expression of all cells in a patient **f**, Clonal size for every cell, overlaid onto the UMAP Cells without paired TCRβ recovery are colored in gray **g**, Distribution of clonal sizes (defined as the number of cells sharing a TCR sequence), within each sorted subset, by unique TCRβ CDR3 **h**, Heatmap of the percentage of TCRβ shared between conditions, defined as the number of unique TCRβ CDR3 sequences detected in both conditions divided by the geometric mean of the number of unique sequences in each of the two conditions "Condition" was defined as a sorted subset at a time point, with sequences from all patients pooled together

PBMCs from four time points were cultured with peanut extract for 22h Peanut-reactive T cells were then enriched via FACS using CD4 memory markers and the activation markers CD154 (a marker for antigen-stimulated T-helper cells[10]) and CD137 (a recently-reported marker for isolating activated Tregs[11]) (Figure 4-1B, A4-1) This workflow was chosen in order to elicit activation of a broad range of peanut-specific T cells with minimal bias for epitopes or HLA types Because activation-based sorting may contain non-specific bystander cells, additional filtering was later done using TCR sequences to identify clonally expanded cells, as a proxy for antigen-specific cells Sorted memory CD4 T cells were processed for single-cell RNA sequencing (scRNA-Seq) via Seq-Well[12] and paired single-cell TCRα/β sequencing using a new 3'-sequencing approach[13]

In total, we recovered high-quality transcriptomes for 134,129 cells (74,646 CD154+, 41,186 CD137+, and 18,297 CD154-CD137-) There was a strong association between transcriptional state and sorted subset, as well as smaller associations with patient (Figure 4-1C, 4-1D, 4-1E) Despite normalizing for technical factors such as library size and frequency of mitochondrial genes, these patient associations remained, suggesting inherent biological differences rather than batch effects (Figure A4-2) Top genes differentiating between CD154+ and CD137+ included (unsurprisingly) CD40LG and actin genes in CD154+ cells, and TNFRSF9 and the regulatory markers FOXP3 and TIGIT in CD137+ cells To see if these striking transcriptional differences between CD154+ and CD137+ cells were reflected in TCR usage, we turned to the paired TCR sequences recovered for each cell Using a new methodology for paired TCRa/b sequencing[13], we recovered TCRa sequences for 60% of cells and TCRb sequences for 70% of cells (with both sequences detected in 45% of cells) (Figure A4-2) We identified clonal cells on the basis of matching TCRb CDR3 sequences – although we note that the vast majority of expanded TCRb sequences were paired with a single TCRa (Figure A4-3) Clonal size, defined as the number of cells sharing a TCRb CDR3 sequence, was then plotted for each sorted subset Compared to two negative controls, clonal sizes of 10+ were exclusively present in CD154+ and CD137+ cells, confirming a strong selection of expanded clones among peanut-activated cells, but clonal sizes of the two activated subsets were comparable, indicating that both might be enriching for comparable frequencies of peanut-specific T cells The expanded clones largely localized within certain areas of the UMAP, indicating an association between expansion and transcriptional state (Figure 4-1F, 4-1G) We then examined if clonotypes were shared across time points or between the CD154+ and CD137+ compartments 55% of expanded clones were present at multiple time points, but notably, clones were almost exclusively present in either CD154+ or CD137+ cells, suggesting fundamental lineage or epitope specificity differences between the two sorted cell subsets (Figure 4-1H)

### 4.3.2. Clonally expanded T cells are associated with specific gene expression modules

To observe the finer-grain cell subsets present among peanut-reactive T cells, we developed an approach to score and visualize narrow gene programs expressed in each cell The approach combines co-expressed genes into gene modules in an unsupervised manner using sparse principal component analysis (PCA)[14] As a PCA-based approach, it is algorithmically transparent and easy to implement The sparsity of each principal component (i e , gene module) is tuned by a single parameter to limit the number of genes included in each component  To further aid interpretation, we restricted the input genes to the union of immune genes (defined using ImmPort gene lists) and variable genes in the dataset  Using this strategy, we discovered several modules that recapitulated known functional states of T cells, such as TH2 function, TH17 function, TH1 function, MHC II upregulation, and regulatory T-cell function (Figure 4-2A, A4-4, A4-5)  These diverse functional states of memory T-helper cells were present across most or all patients, suggesting a role for each in the peanut-specific response

We next wanted to assess the degree of clonality and TCR specificity for each gene module corresponding to a major functional T-helper state, and whether or not these gene modules were co-expressed within cells  To test this quantitatively, we asked what gene modules were associated with highly expanded cells and with specific, rather than random, TCR sequences (Figure 4-2C)  We analyzed the relative clonal expansion of module-expressing cells, and we also calculated the fold-change in likelihood of two cells of the same TCRb both expressing a gene module, if they had the same TCR  The latter metric was used to evaluate the association with specific, not random, TCRs, which one would hypothesize to be associated with an antigen-specific, rather than bystander, response  As expected, the TH2 module had among the highest expansion and association with specific TCRs in both CD154+ and CD137+ cells, with TH1 cells also scoring as highly clonal and TCR-specific (Figure 4-2C)  A heatmap of top expanded clones showed qualitatively that expression of a module was fairly consistent within each clonotype, with often 100% of cells expressing the module  Additionally, a strong overlap between TH1 and TH17 expression, but not with TH2 or Treg expression, was evident, highlighting the coexpression of different gene programs within peanut-reactive T cells

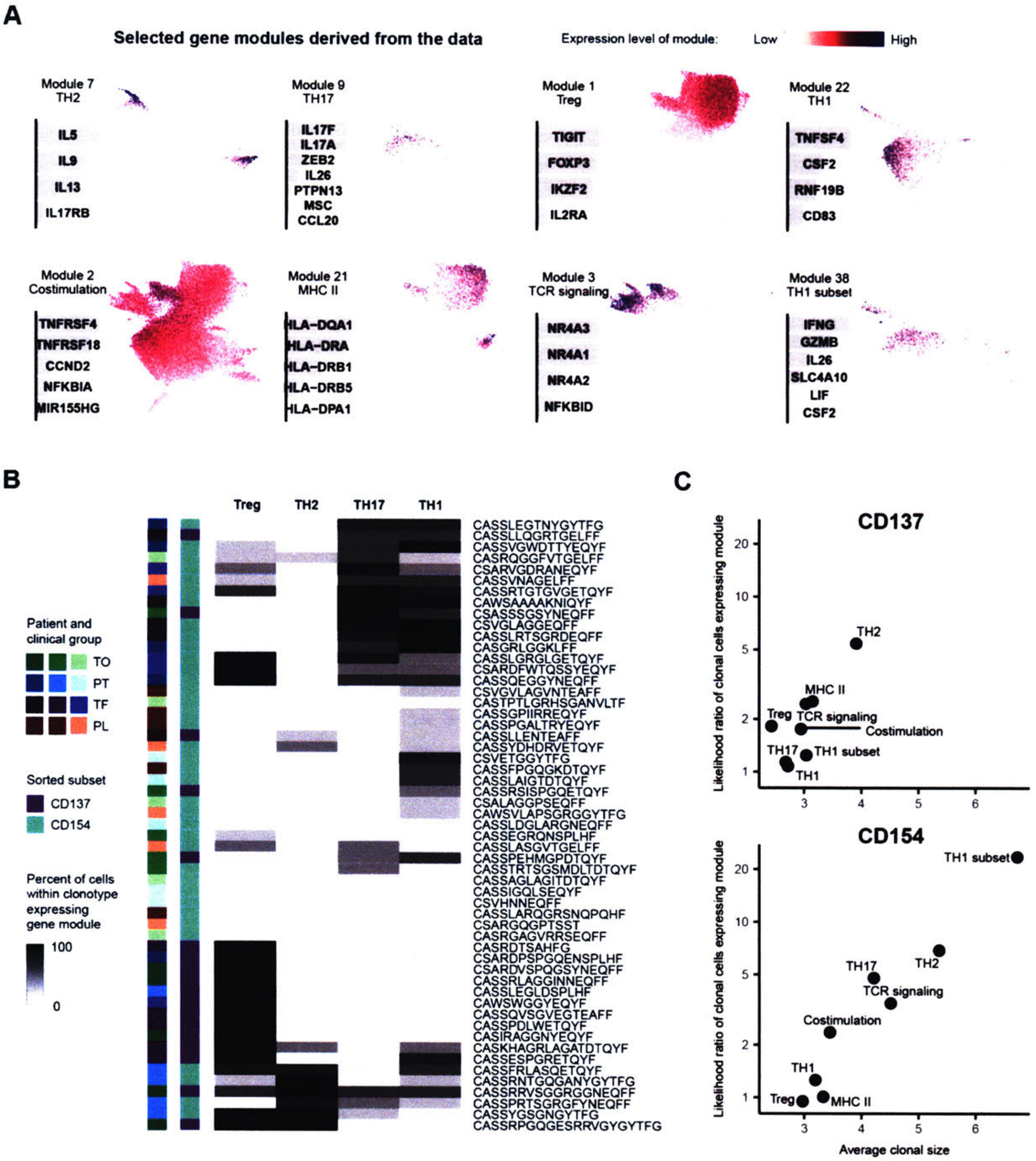


Figure 4-2. Unsupervised gene module discovery highlights several transcriptional programs associated with clonal expansion in peanut-reactive T cells. **a**, Selected gene modules discovered from the data using a sparse PCA approach. For each module, the weights of each contributing gene, and an overlay of module score on the UMAP coordinates, are shown. **b**, Heatmap of top expanded clonotypes and the percent of cells within each clonotype expressing each module (cells were gated as "expressing" a module using the CD154-CD137- cells as a negative population). Clonotypes (rows) are annotated with the majority patient and sorted subset in which

the clonotype was detected **b**, Average clonal size and association with specific TCR sequences for each gene module Clonal size, defined as the number of cells with the same TCRb sequence, was calculated across all cells and then averaged for those expressing the module Association with specific TCR sequences was defined as the fold-change in likelihood of observing two cells both expressing a gene module, if they had the same TCRβ

### 4.3.3. Multiple clonotypically distinct subsets of peanut-reactive TH2 cells exist

An interesting observation that emerged from the gene module analysis was that TH2 cells appeared to consist of multiple distinct cell subsets To probe this further, we selected TH2-expressing cells and re-visualized them alone We found three distinct clusters (Figure 4-3A), which were present in most or all patients Based on the differentially expressed genes enriched in each cluster, these appeared to represent a Tfh-like population (high in costimulatory markers, CXCR5, and PD-1), a deviated Treg state (identical to FOXP3+ CD137+ cells except for the additional expression of TH2 cytokines), and a GATA3-high population (with high levels of GATA3, IL17RB, and CHDH) (Figure 4-3B) All subsets shared similar levels of clonal expansion, but clonotypes were, surprisingly, highly restricted to a single subset (Figure 4-3C, 4-3D)

To understand whether this restriction was the result of convergence onto certain CDR3 motifs, we performed a TCR distance analysis to see whether highly similar TCRs (not just exact TCRs) also tended to be of the same subset TCR distance was assessed using a previously published method[15], in which an exact CDR3 match scored as 0 and each individual amino acid substitution added a penalty of 1-4 We found that for extremely similar TCR sequences (distance < 9), there was an increased likelihood of both cells being Tfh-like (Figure 4-3E) This phenomenon was not present for highly similar-TCR cells that were GATA3hi or deviated Tregs Two groups of closely related CDR3s in the Tfh-like subset that resulted from this analysis are shown (Figure 4-3F) Intriguingly, this result suggests a convergence onto common epitope recognition motifs only for the Tfh-like cells, which might indicate a relatively narrower set of target epitopes for that subset

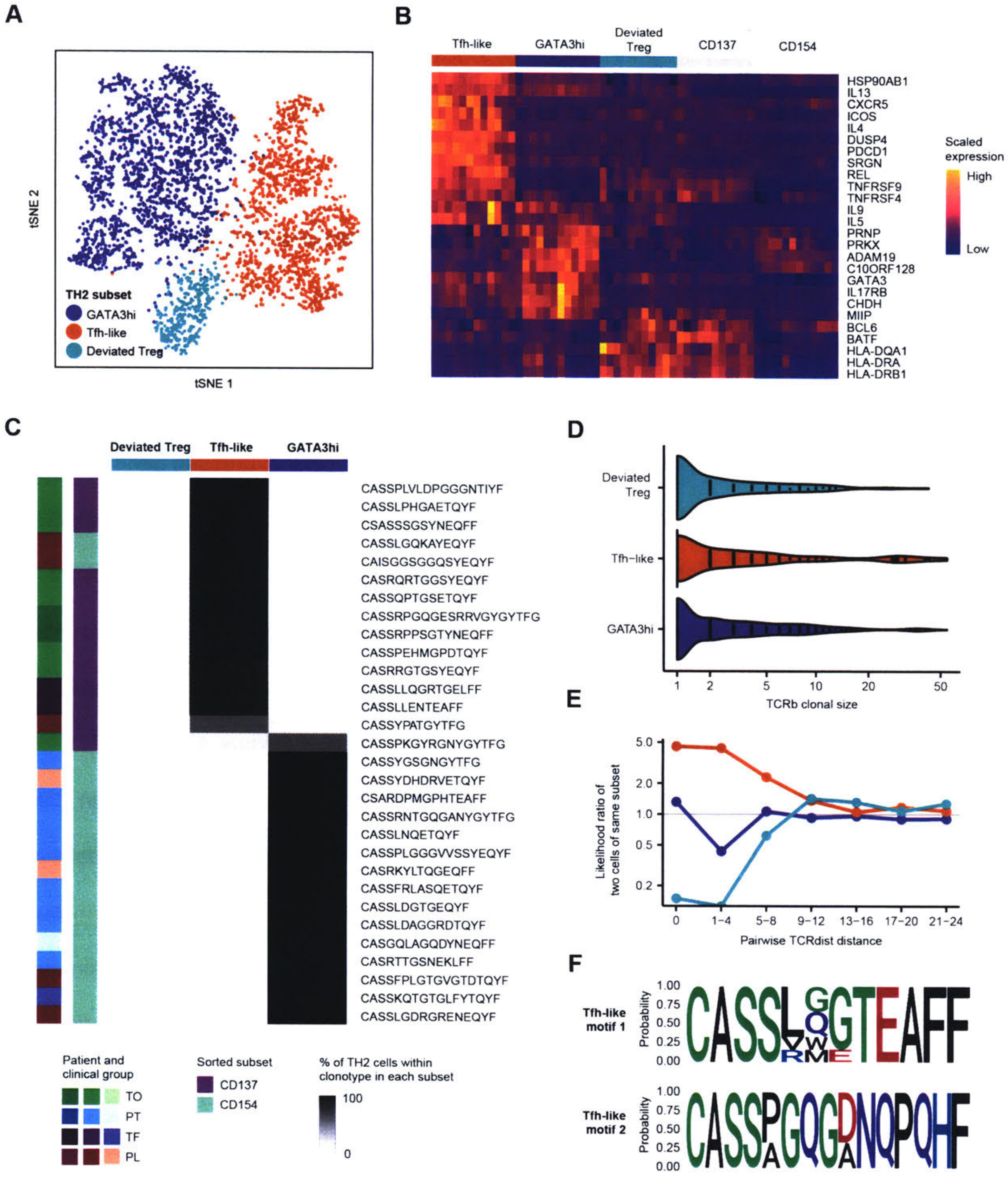


Figure 4-3. Clonotypically distinct TH2 subsets differ in TFH, Treg, and TH2 qualities. **a,** UMAP visualization of TH2-scoring cells, re-clustered and colored by cluster identity. **b**, Differentially expressed genes upregulated in each TH2 subset; genes were identified using a ROC test. Rows represent z-scored gene expression values, and columns represent average expression of all

cells ın a gıven patıent **c**, Heatmap of the percentage of TH2 TCRβ CDR3s shared between condıtıons, defıned as the number of unıque CDR3 sequences detected ın both condıtıons dıvıded by the geometrıc mean of the number of unıque CDR3 sequences ın each of the two condıtıons "Condıtıon" was defıned as a TH2 subset at a partıcular tıme poınt, wıth sequences from all patıents pooled together **d**, Dıstrıbutıon of TCRβ clonal sızes for cells ın each subset, wıth all cells (left) or all TH2-scorıng cells (rıght) **e**, Probabılıty of cells wıth hıghly sımılar TCRs to be ın each subset Paırwıse TCR dıstance usıng the TCRdıst method ıs plotted on the x-axıs On the y-axıs ıs the probabılıty of two TH2-scorıng cells belongıng to a subset, gıven that theır TCRβ CDR3 sequences are a certaın dıstance apart The dotted lıne represents the prıor probabılıty of any two TH2 cells belongıng to the subset **f**, Selected TCRb CDR3 motıfs present among Tfh-lıke TH2 cells The probabılıty of each amıno acıd appearıng at each posıtıon, among the selected clones wıth each motıf, ıs plotted

### 4.3.4. TH2 and TH17 clonotypes are functionally suppressed, but not deleted, during OIT

Wıth a knowledge of the cell subsets present among peanut-reactıve cells and theır clonotypıc assocıatıon, our next goal was to track T cell clones from these subsets longıtudınally and see how frequency and functıon were altered durıng OIT The objectıve was to shed lıght on the unresolved questıon of whether OIT predomınantly ınduces deletıon, anergy, or another fate ın peanut-specıfıc T cells In TH2-expressıng cells, we fırst observed that the vast majorıty of expanded TH2 clonotypes were present at all four tıme poınts, wıth a smaller number that were mıssıng at later tıme poınts (Fıgure 4-4A) There was a small dıfference ın thıs pattern wıth respect to TH2 subset, wıth Tfh-lıke clonotypes beıng the most lıkely to be detected at all four tıme poınts Overall, however, thıs result compellıngly suggested that most peanut-reactıve TH2 clones were not deleted from the perıphery over tıme To see ıf suppressıon of TH2 clones was present, we then looked at the longıtudınal expressıon of the TH2 gene module wıthın clonotypes of each subset, and we observed a strong ablatıon ın TH2 expressıon durıng maıntenance and a slıght rebound ın expressıon at avoıdance Thıs was especıally the case ın all subsets except for the Tfh-lıke subset, whıch nevertheless was slıghtly suppressed compared to baselıne (Fıgure 4-4B) TH2 clonotypes thus appear to be functıonally suppressed, rather than deleted, by the treatment, a fındıng that agrees wıth the often-transıent nature of the desensıtızatıon ınduced by OIT The differential suppression of the different TH2 subsets suggests that OIT might possibly act in a cell-subset- or epitope-biased manner

To extend this analysis to non-TH2 cells, we looked at the longitudinal expression of all four major functional gene modules (TH1, TH2, TH17, Treg) within expanded CD154+ clonotypes We observed a similar suppression in TH17 expression compared to baseline, but not in TH1 or Treg function (Figure 4-4C) This result again suggested cell-subset-biased effects of OIT, since only TH2- and TH17-expressing clonotypes were functionally suppressed by treatment Finally, we looked at the association of the major functional gene modules in expanded CD154+ cells with clinical outcome, and interestingly, we found that TH17 expression was strongly correlated with poor clinical outcome even at baseline (Figure 4-4D) TH2 expression, surprisingly, did not correlate with clinical outcome either based on baseline levels or by longitudinal trends, as TH2 expression of clonotypes from all patients was similarly suppressed The enrichment of TH17 expression among patients with poor clinical outcome suggests a pathogenic role for non-TH2 cell types in food allergy Moreover, the fact that TH17 clonotypes were also only functionally suppressed, and not deleted, highlights the difficulty in achieving truly persistent, rather than transient, outcomes following OIT

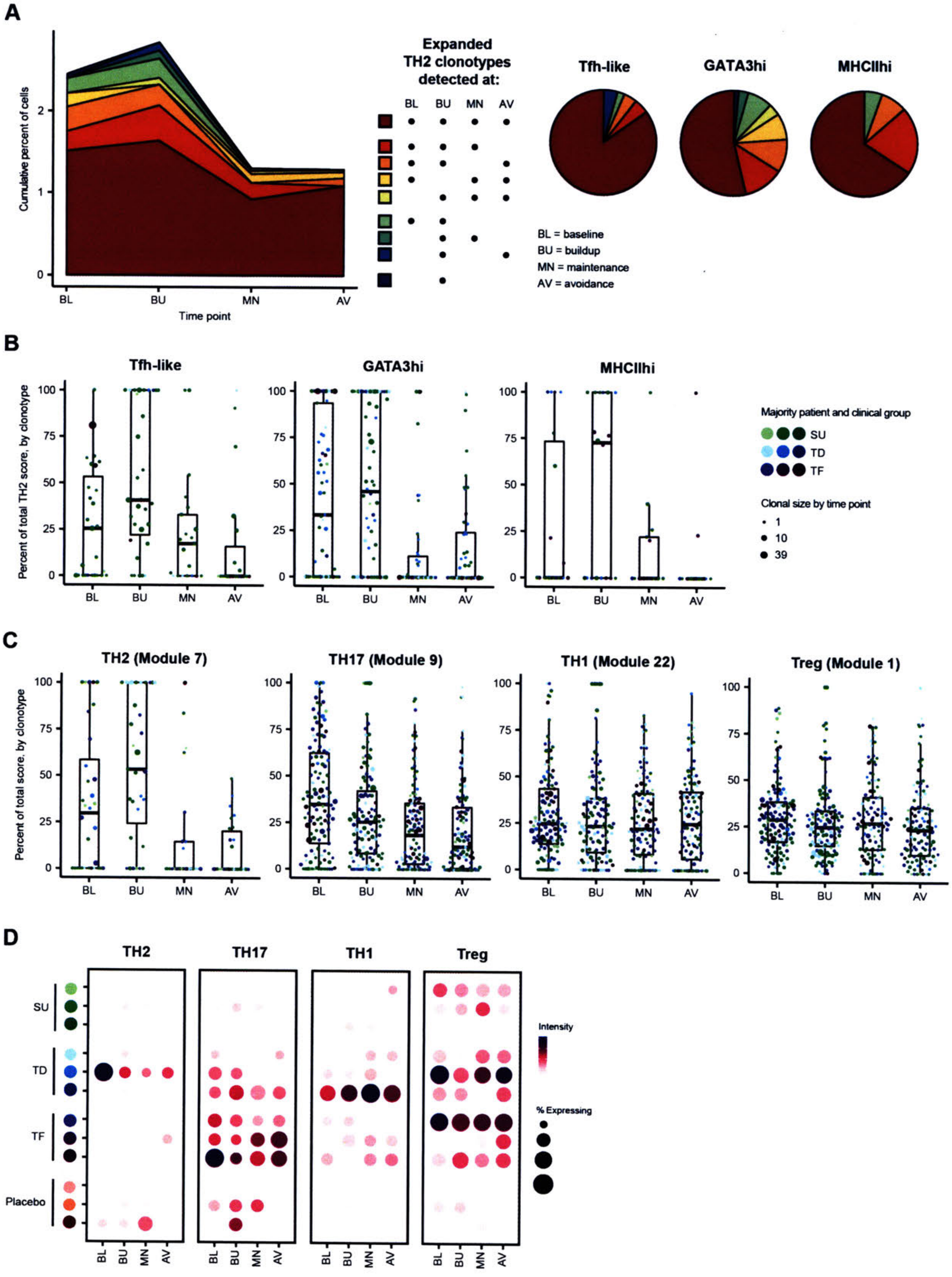

A
Cumulative percent of cells
Time point
BL
BU
MN
AV
Expanded TH2 clonotypes detected at:
Tfh-like
GATA3hi
MHCIIhi
BL = baseline
BU = buildup
MN = maintenance
AV = avoidance
B
Tfh-like
GATA3hi
MHCIIhi
Percent of total TH2 score, by clonotype
Majority patient and clinical group
SU
TD
TF
Clonal size by time point
1
10
39
C
TH2 (Module 7)
TH17 (Module 9)
TH1 (Module 22)
Treg (Module 1)
Percent of total score, by clonotype
D
TH2
TH17
TH1
Treg
SU
TD
TF
Placebo
Intensity
% Expressing

Figure 4-4 TH2 and TH17 clonotypes are functionally suppressed, but not deleted, by OIT **a,** Temporal patterns of expanded TH2 clonotypes Clonotypes in treatment patients with at least eight cells detected were used, of which at least one cell had to score as TH2 Left Stacked area plot with each colored ribbon representing a temporal pattern The width of each ribbon at each time point represents the frequency of all cells of that temporal pattern at that time point, normalized to the number of cells with TCRβ recovery at that time point Right Pie charts of temporal patterns of all cells in each TH2 subset **b**, TH2 module score, averaged by clonotype, over time Clonotypes with TH2 expression in at least one cell and detection at two or more time points were included Each dot represents the average expression for one clonotype at a time point, with dot size representing number of cells averaged Dot color represents the majority patient in which the clonotype was detected **c**, Scores of modules corresponding to major T cell functions, averaged by clonotype, over time In each plot, CD154+ clonotypes with module expression in at least one cell and detection at two or more time points were included Dot color and size legend from **b** apply **d**, Single-cell dot plot visualization of gene module expression, aggregated by patient (row) and time point (column) from all expanded clonotypes used in **c** Dot size indicates percent of cells expressing the module, and dot color indicates the mean module score for all cells in the sample

## 4.4. Discussion

As an experimental treatment for food allergy, OIT has been shown to induce changes in circulating allergen-specific IgE and IgG4 titers and variable outcomes in patients, from transient desensitization to more sustained tolerance There is less concurrence on treatment-induced changes in allergen-specific T-helper cells, which could be important in directing antibody changes and patient outcomes In this work, we profiled circulating peanut-reactive T-helper cells from patients undergoing peanut OIT, with the goals of learning the effects of treatment on the function and frequency of individual clonotypes, and T-cell correlates of clinical outcome Using single-cell RNA sequencing and paired TCR sequencing, we found a diversity of transcriptional modules among peanut-reactive T cells, including TH2, TH17, Treg, TH1, and MHC II expression, and specific patterns of TCRb clonal expansion associated with each Within the TH2 module, there were three distinct subsets of TH2 cells (a GATA3hi subset, a Tfh-like subset, and a deviated-Treg phenotype) that were clonotypically distinct We then tracked expression within clonotypes longitudinally, and we found that in all three TH2 subsets, expression of the TH2 module was transiently suppressed, but the clonotypes were not deleted We saw a similar transient suppression in TH17-expressing clonotypes Finally, we looked for baseline predictors of clinical outcome and found that higher TH17 expression was associated with treatment-failure, which could point to pre-defined pathogenicity in the treatment-failure patients contributing to poor outcome Overall, these results suggest that OIT may induce non-durable reprogramming in a majority of peanut-reactive T cells

Intriguingly, we observed multiple subsets of TH2 cells that were clonotypically distinct This segregation of TH2 cells has not been described before, to our knowledge Upon investigation of whether these two subsets aligned with any TH2 subsets previously described in the literature (such as TH2A cells[7] or Tfh13 cells[16]), we saw good concordance between Tfh13 cells and the Tfh-like TH2 cells in our dataset, based on gene expression Reassuringly, although we did not observe appreciable expression of the Tfh transcription factor Bcl6, the study describing Tfh13 cells shows a similarly low expression of Bcl6 We could not determine whether one or both of the subsets aligned with the pathogenic TH2A phenotype described previously The distinct TCR usage of the subsets described in our study, and the evidence of TCR convergence at least in the Tfh-like TH2 cells, suggests that epitope recognition could play a role in directing these phenotypes to emerge

We did not observe an emergence in peanut-reactive Treg cells over time, as some others have reported[9] Instead, Treg expression levels in both CD137+ and CD154+ clonotypes, and the number of clonotypes expressing the Treg module, were relatively constant over time This discrepancy could be due to the fact that we measured gene expression upon peanut stimulation, which might yield different findings from profiling resting peanut-specific T cells, or that when peanut-reactive T cells are studied holistically with a diverse antigen pool, the emergence of Tregs may not be significant Regardless of the reason, the implications of Treg emergence (or not) during OIT is important The appearance of Tregs has been extensively described in natural tolerance and SIT[17], but findings thus far are mixed in OIT[5], if peanut-specific Tregs do not in fact emerge during OIT, it means that different modes of tolerance induction could favor different T-cell mechanisms

Taken together, these results indicate that the majority of OIT reprogramming appears to be nondurable, as most TH2-expressing clonotypes are transcriptionally suppressed by the treatment, but are not deleted, and some regain function after the avoidance phase Importantly, this finding gives resolution to a general decrease in TH2 function that has been observed in OIT previously, but whether due to suppression or deletion was impossible to know without longitudinal tracking of T-cell clonotypes Additionally, clinical outcome is correlated with the elevated expression of TH17 function at baseline that is mildly supressed by the treatment, suggesting that outcome may be relatively set at baseline and resistant to being altered Indeed, innovations to OIT such as anti-IgE have been shown to reduce the incidence of adverse events but not appreciably increase the rate of sustained unresponsiveness[3] Clinical outcome may be driven by diverse pathogenic pathways outside of TH2 signaling, that are more difficult to displace and reprogram

## 4.5. Contributions and acknowledgements

This project was conducted in collaboration with Wayne Shreffler's lab at MGH The clinical trial (NCT01750879) and PBMC isolation and banking were conducted by Wayne Shreffler's clinic PBMC stimulation and sorting was done by Bert Ruiter at MGH Seq-Well experiments were performed with Andy Tu, Patrick Petrossian, and Rachel Barry Paired TCR sequencing was done by Andy Tu Sequencing was done by the MIT Bio-Micro center and by the Broad Walk-Up

sequencing facility Analysis was done with the help of Duncan Morgan and Neal Smith This work was supported in part by the Koch Institute Support (core) Grant P30-CA14051 from the National Cancer Institute

## 4.6. References

1 Gupta, R S *et al* The prevalence, severity, and distribution of childhood food allergy in the United States *Pediatrics* **128**, e9-17 (2011)

2 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)

3 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)

4 Burks, A W *et al* Oral Immunotherapy for Treatment of Egg Allergy in Children *N. Engl J. Med* **367**, 233–243 (2012)

5 Vickery, B P *et al* Sustained unresponsiveness to peanut in subjects who have completed peanut oral immunotherapy *J Allergy Clin Immunol* **133**, 468-475 e6 (2014)

6 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)

7 Wambre, E *et al.* A phenotypically and functionally distinct human TH2 cell subpopulation is associated with allergic disorders *Sci Transl Med* **9**, eaam9171 (2017)

8 Ryan, J F *et al* Successful immunotherapy induces previously unidentified allergen-specific CD4+ T-cell subsets *Proc Natl Acad Sci U S A* **113**, E1286-1295 (2016)

9 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)

10 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 (2006)

11 Bacher, P *et al* Regulatory T Cell Specificity Directs Tolerance versus Allergy against Aeroantigens in Humans *Cell* **167**, 1067-1078 e16 (2016)

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

13 Ang A Tu *et al* Recovery of paired T cell receptors from massively-parallel 3' single-cell RNA-Seq libraries reveals clonotypic responses among antigen-activated T cells *In revision*

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

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

16 Gowthaman, U  *et al*  Identification of a T follicular helper cell subset that drives anaphylactic IgE  *Science* eaaw6433 (2019)  doi 10 1126/science aaw6433

17 Shreffler, W  G , Wanich, N , Moloney, M , Nowak-Wegrzyn, A  & Sampson, H  A Association of allergen-specific regulatory T cells with the onset of clinical tolerance to milk protein  *J Allergy Clin  Immunol*  **123**, 43-52 e7 (2009)

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

19 McInnes, L , Healy, J  & Melville, J  UMAP  Uniform Manifold Approximation and Projection for Dimension Reduction  *ArXiv180203426 Cs Stat* (2018)

# 5. T-cell correlates of clinical reactivity to peanut allergen

Parts of this work appear in A A Tu, T M Gierahn, B Monian, et al, *in revision* and B Ruiter, N P Smith, B Monian, et al, *in revision*

Peanut allergy manifests with a wide range of clinical symptoms and severities, and the underlying immune states contributing to this heterogeneity are not well-understood Particularly in T-helper cells, which are a central component of the food allergic immune response and which have a diversity of functional states, the combination of cell states contributing to different clinical phenotypes is not fully known and has been difficult to study at scale Using single-cell RNA sequencing and paired TCR sequencing, we profiled circulating peanut-reactive T-helper cells from peanut-allergic patients with high or low clinical reactivity to peanut The goal of the study was to discover and interpret T cell states correlated with reactivity We found that TH1 and MHC I gene modules were upregulated in cells from patients with higher sensitivity to peanut, and that, surprisingly, a TH2 gene module actually had slightly upregulated expression in hyporeactive patients, suggesting a possible temporal dependency and a non-TH2 component to the severity of the T-cell immune response in food allergy

## 5.1. Motivation

As of 2010, peanut allergy affects more than 1% of Americans and tends to be higher in severity and more permanent than other food allergies[1] In addition to the risk for dangerous allergic reactions, having a food allergy can also induce daily anxiety Knowledge of the severity of an allergy and the likely risk of a reaction could be useful in decreasing uncertainty and improving an allergy patient's quality of life Most diagnostic methods for food allergy, while reasonably accurate at assessing whether an individual has a particular food allergy, do not accurately predict (and are not designed to test) differences in clinical reactivity, e g , whether 10 or 1000 mg of peanut is necessary to cause an allergic reaction[2] The gold-standard for gauging clinical reactivity is a double-blind, placebo-controlled oral food challenge and the appearance of allergic symptoms at a particular dose of allergen, but this test is cumbersome and stress-inducing A rapid and simple test for allergen reactivity represents an unmet need in the management of food allergy

Perhaps even more importantly, molecular or cellular biomarkers of clinical reactivity could be useful in understanding mechanisms of severity in food allergy Previous mechanistic work is scant, but one study identified an increased frequency of circulating peanut-responsive CD154+ memory T-helper cells expressing TH2 cytokines in children with high reactivity[3], an observation supported but not statistically significant in a related study in infants[4] Another study reported that in symptomatic and asymptomatic patients with low levels of peanut-specific IgE, symptomatic patients had IgE specific to certain peptide epitopes from the peanut proteins Ara h1 and Ara h2, suggesting that antibody diversity and repertoire, rather than pure titer, may be relevant for clinical severity[5] These results suggest that complex mechanisms, involving both the intensity of the T-cell response and the breadth of the B-cell response, might be at play in determining severity in peanut allergy

We sought to address whether we could identify T-helper cell correlates of clinical reactivity in peanut allergy, as T cells represent an important component of the allergic response that might be influencing the degree and quality of the B-cell response. For samples from patients with differing clinical reactivities, we relied on a clinical trial for peanut oral immunotherapy, for which several patients were recruited who failed the inclusion criteria by not reacting during a baseline oral food challenge with peanut. These patients were deemed "hyporeactive" to peanut. Additionally, several "reactive" patients did show allergic symptoms during the baseline challenge with peanut. We collected PBMCs from both groups of patients, stimulated the cells with peanut antigen, sorted activated memory T-helper cells, and performed single-cell RNA sequencing via Seq-Well[6] in order to study peanut-reactive T cells at both high resolution and high throughput. We also employed paired TCR sequencing using a new TCR transcript pulldown and amplification approach[7], to interpret any relevant cell states in the context of their likely antigen specificity.

## 5.2. Methods

**Patient enrollment.** Peanut-allergic subjects were screened for participation in a clinical trial for peanut oral immunotherapy at the Food Allergy Center at Massachusetts General Hospital (NCT01750879). All subjects were recruited with informed consent, and the study was approved by the Institutional Review Board of Partners Healthcare (protocol number 2012P002153). Individuals were screened on the basis of having a previous diagnosis of peanut allergy and peanut- and Ara h 2-specific serum IgE titers of > 0.35kU/L (ImmunoCAP, Thermo Fisher). Patients then underwent a double-blind, placebo-controlled oral food challenge of up to a maximum dose of 300mg of peanut protein. Individuals who tolerated up to this dose without objective symptoms were deemed "hyporeactive", whereas those who did react at this dose or lower were deemed “reactive”  Two weeks prior to the food challenge, a blood sample was collected for PBMC isolation

**PBMC isolation and culture with peanut antigens.** PBMCs were isolated from blood via density-gradient centrifugation (Ficoll-Paque Plus, GE Healthcare)  Cells were frozen in fetal bovine serum with 10% DMSO  At a later date, cryopreserved PBMCs were thawed and plated in a 24-well plate at 5M cells/1ml/well in AIM-V medium (Gibco)  Cells were stimulated with 100ug/ml delipidated peanut extract (following x protocol), xxx anti-CD3/CD28 beads (Dynabeads, Thermo Fisher), or PBS  Cultures were incubated at 37C for 20h before being harvested for staining and flow sorting  3h before the end of the culture, PE anti-CD154 antibody (BD, clone TRAP1) was added at a dilution of 1 50, this step was omitted for the 0h cultures

**Flow-based enrichment of peanut-reactive T cells.** Cells were washed in PBS and stained for 25min at 4C with Zombie Violet viability dye (Biolegend), then with APC anti-CD3 (clone UCHT1), APC-Cy7 anti-CD4 (clone RPA-T4), AF488 anti-CD45RA (clone HI100), PE anti-CD154 (clone TRAP1), and BV605 anti-CD137 (clone 4B4-1)  All antibodies were purchased from Biolegend except for PE anti-CD154, which was from BD  After staining, cells were washed with FACS buffer (PBS with 1% BSA and 1mM EDTA) and sorted on a BD FACS Aria instrument  Cells were sequentially pre-gated as lymphocytes, singlets, live cells, CD3+CD4+, and CD45RA-, and were then sorted as either CD154+ or CD154-

**Single-cell transcriptome sequencing.** Sorted cells were immediately processed for single-cell RNA sequencing via the Seq-Well protocol[6]  Briefly, cells were co-loaded into wells at approximately single-cell occupancy with poly(dT) beads and lysed to allow mRNA to hybridize onto the beads  Beads were then pooled and mRNA was reverse-transcribed, PCR-amplified, and prepared for sequencing via the Nextera XT kit  Libraries were sequenced on the Illumina Novaseq

**Paired single-cell TCR sequencing.** Paired TCR sequencing was performed according to Tu et al[7] Briefly, following cDNA amplification in the Seq-Well protocol, biotinylated capture probes for human TRAC and TRBC regions were annealed to cDNA Magnetic streptavidin beads were then used to enrich the bound TCR sequences, which were then further amplified using human V-region primers and prepared for sequencing using Nextera sequencing handles Libraries were sequenced on an Illumina MiSeq using 150bp-length reads

**TCR sequencing data preprocessing.** TCR sequencing reads were preprocessed according to Tu et al[7] In short, reads were mapped to TCRV and TCRJ IMGT reference sequences via IgBlast, and V and J calls with "strong plurality" (wherein the ratios of the most frequent V and J calls to the second most frequent calls were at least 0 6) were retained CDR3 sequences were called by identifying the 104-cysteine and 118-phenylalanine according to IMGT references and translating the amino acid sequences in between those residues Processed TCR sequences were then paired with the single-cell transcriptome data via the cell barcodes

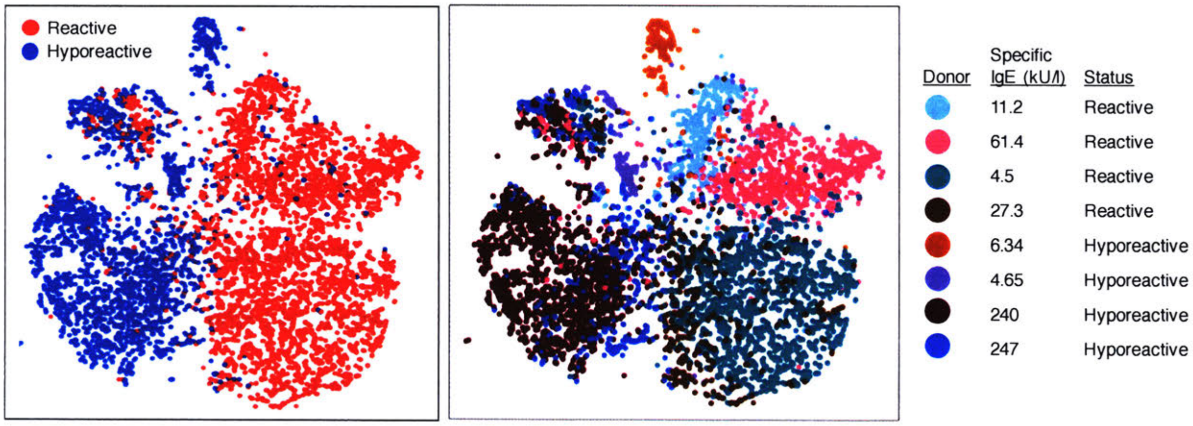
**Transcriptome sequencing data preprocessing and visualization.** Raw read processing was performed as in Macosko et al[8] Briefly, sequencing reads were aligned to the hg38 human genome and counted to obtain a digital gene expression matrix of cells versus genes The matrix was filtered to exclude any cells with fewer than 500 detected genes or 1000 detected transcripts (UMIs) Counts were then normalized by cell library size and log2-transformed using the Seurat package in R, and transcriptomes were visualized using a two-dimensional t-SNE projection

**Logistic regression.** A classifier was built to discover discriminating features between single-cell transcriptomes in reactive and hyporeactive patients using logistic regression in the R package 'glmnet' Scores for the top 50 gene modules (sorted on percent variance explained) were selected as input for the classifier A random subset of 25% of the cells was held out as a test set, and the remaining 75% was used to train the model Optimal model size was selected using a lasso penalty and 10-fold cross-validation. Misclassification rate of the selected model was then calculated and reported for cells in the test set.

## 5.3. Results

### 5.3.1. Survey of peanut-reactive T cells recovered from eight peanut-allergic patients

Using t-SNE, we first visualized the peanut-reactive T cell transcriptomes collected from all eight patients (Figure 3-1). A strong separation between transcriptomes from clinically reactive and hyporeactive patients was immediately evident. We also observed that each patient's cells had a unique transcriptional signature, which we confirmed was not likely to be due to batch effects by looking at library size and number of genes detected by patient (Figure A5-1). Intriguingly, a distinct cluster in the top left of the t-SNE was composed of cells from all patients; this cell subset is the subject of Chapter 3 of this thesis and represents a highly activated state in antigen-stimulated CD4 T cells (Figure A5-2).



| Donor | Specific IgE (kU/l) | Status |
|---|---|---|
| ● | 11.2 | Reactive |
| ● | 61.4 | Reactive |
| ● | 4.5 | Reactive |
| ● | 27.3 | Reactive |
| ● | 6.34 | Hyporeactive |
| ● | 4.65 | Hyporeactive |
| ● | 240 | Hyporeactive |
| ● | 247 | Hyporeactive |

Figure 5-1. tSNE visualization of transcriptomes of all CD154+ peanut-reactive T cells. Cells recovered from eight donors are colored by clinical reactivity to peanut (left) or by donor (right). At right is a key with peanut-specific IgE and clinical reactivity information for each donor.

To uncover transcriptional states associated with the visual differences between reactive and hyporeactive patients in the t-SNE plot, we ran the gene module discovery method described in Chapter 2 of this thesis. This allowed us to group thousands of genes into tens of coexpressed gene programs, aiding in both biological interpretation and model parsimony. The gene modules encompassed a wide variety of T cell programs, both known and new (Figure 5-2). For example, Module 3, 12, and 13 represented TH2 function, Treg function, and TH17 function, respectively. However, we also saw modules that did not obviously correspond to known T cell programs (Module 1, 5, 11, and others) which could be pursued in the discovery of new gene networks.

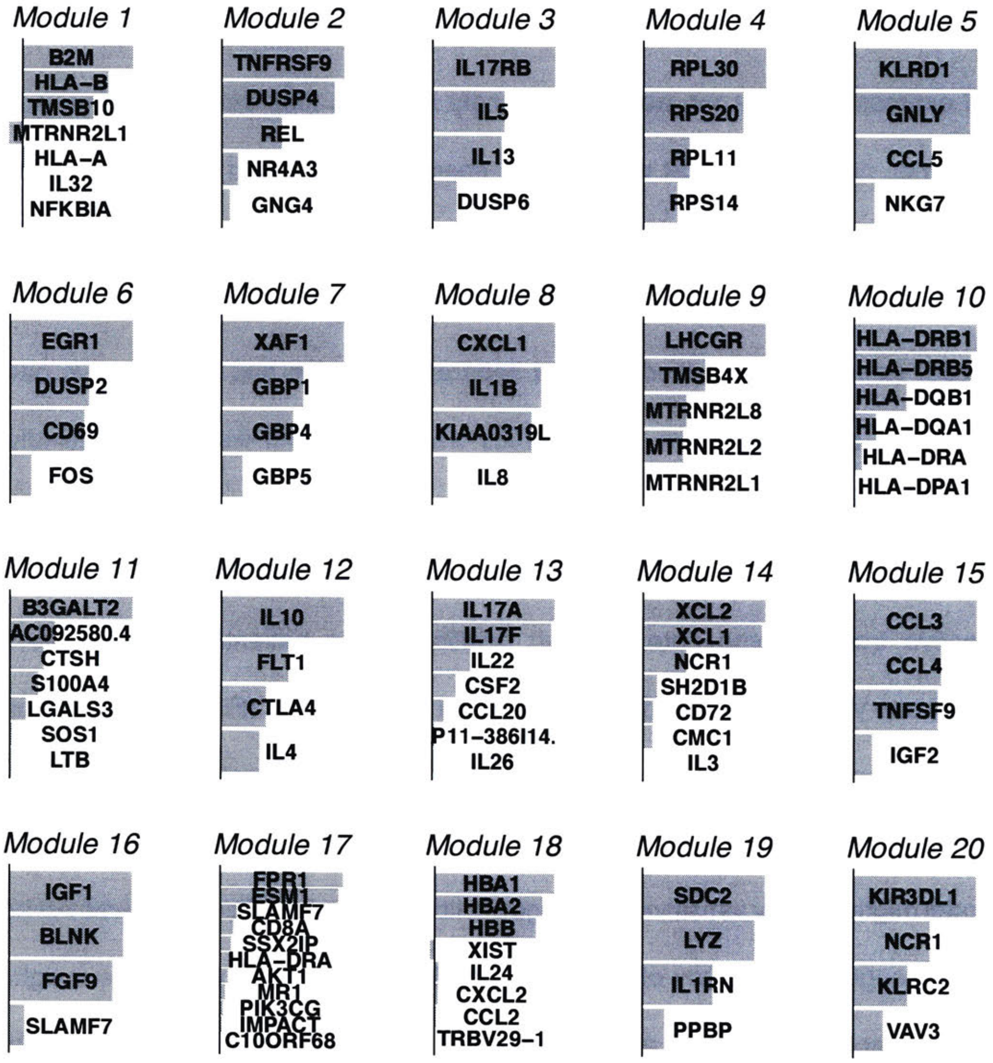


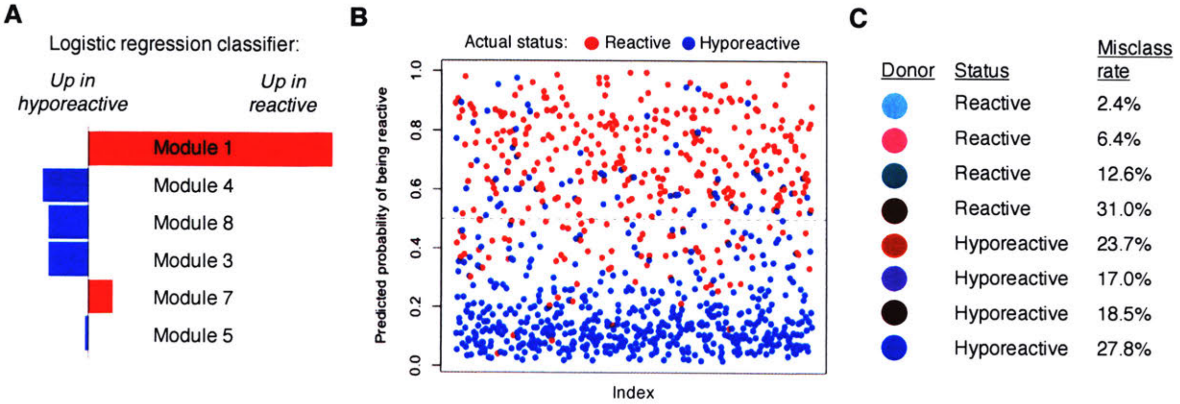
Figure 5-2. Gene module discovery in peanut-reactive T cells. Using the unsupervised approach described in Chapter 2 of this thesis, co-expressed gene modules were derived from the entire

dataset of eight donors The magnitude and direction of each bar represent the weight and its sign of each gene in each component The top 20 gene modules based on percent variance explained are shown

### 5.3.2. Classification of cell states using logistic regression

We ran a logistic regression using the top 50 gene modules to classify cells from reactive and hyporeactive patients We used individual cells as observations, instead of aggregating together all cells from each patient, which shifted the analysis question slightly and allowed us to make use of significantly more data as a result In this vein, we also selected classification instead of differential expression, since we wanted to identify changes that were conserved across cells from all patients, not changes that were patient-specific (which could have arisen in differential expression, given that individual cells from a small number of patients could strongly skew the analysis) Using ten-fold cross-validation and a lasso penalty (Figure A5-3), we identified an optimal model size of six gene modules that separated reactive and hyporeactive patient cells (Figure 5-3) The misclassification rate for cells was 15 1% and was, unsurprisingly, slightly different by patient, recalling to mind the patient-specific transcriptional signatures in Figure 5-1 To evaluate the performance of the model on an independent cohort, we collected CD154+ peanut-reactive transcriptome data from an additional nine patients (five reactive and four hyporeactive) one year later and applied the classifier to these cells (Figure A5-4) Unfortunately, these cells were processed using an updated library preparation strategy, and it appeared that there were strong global shifts in the classification success of cells from each patient, suggesting that the model might be very sensitive to changes in library recovery Thus, this was not an entirely useful way to gauge the accuracy of the classifier, although it did provide some insight into the classifier's likely lack of robustness across different library preparation techniques and batches

Looking into the components most important to the classification model, the top gene module, surprisingly, consisted of several MHC I genes (B2M, HLA-B, HLA-A) that were upregulated in reactive patient cells (Figure 5-3A). This result suggests that the T cells from the reactive patients could be preparing for cell division via proliferation or otherwise upregulating MHC I. A module containing TH1-related genes GBP1, 4, and 5 (Module 7) was also upregulated in reactive patients. Interestingly, a TH2 gene module (Module 3) was mildly associated with hyporeactive status, which has several possible intriguing explanations that are explored in Discussion. To further interpret and understand these results, we next looked into the paired TCR sequences of these cells to examine their clonality.



| Donor | Status | Misclass rate |
|---|---|---|
| ● | Reactive | 2.4% |
| ● | Reactive | 6.4% |
| ● | Reactive | 12.6% |
| ● | Reactive | 31.0% |
| ● | Hyporeactive | 23.7% |
| ● | Hyporeactive | 17.0% |
| ● | Hyporeactive | 18.5% |
| ● | Hyporeactive | 27.8% |

Figure 5-3. Classification of T cells from reactive and hyporeactive patients. A) Coefficients of a lasso-penalized logistic regression model for classifying the cells. The top 50 gene modules were selected as features for input into the regression. The size of each bar indices the magnitude of the relevant coefficient, and the color indicates whether it is upregulated in reactive (red) or hyporeactive (blue) patients. B) Classification success for all cells in the test set (ordered randomly on the x-axis). Cells are colored by their observed status (reactive or hyporeactive patient) and the y-axis shows the predicted probability of being reactive. Cells with a probability of above 0.5 were classified as "reactive", and below as "hyporeactive". The overall misclassification rate was 15.1%. C) Misclassification rate for cells in the test set, by donor.

### 5.3.3. Degree of association between TCR sequence and transcriptome

Once we had identified modules that were associated with high or low clinical reactivity, we next wanted to determine whether these gene modules represented bystander or antigen-specific activation, and what their degree of clonality was. To answer this question, we turned to the paired

TCR sequencing, for which recovery was efficient and uniform across cells (with TCRb sequences recovered for 55% of cells and TCRa sequences recovered for 32%)  To visually assess the overlap between clonal expansion and gene expression, we overlaid TCRb clonal size (defined as the number of cells sharing the same TCRb sequence) onto the tSNE plot and compared it to overlays of the top gene modules of interest (MHC I and TH2) (Figure 5-4)  There appeared to be an association between high clonal size and TH2 expression, but interestingly, not between high clonal size and MHC I expression  However, quantitative confirmation of this trend was needed, and more specifically, it was important to know whether expression of modules such as MHC I was confined to specific TCR clonotypes, or whether it was expressed broadly across many clonotypes  This would in turn tell us whether the module was likely to represent antigen-specific, or bystander, activation, with the assumption that bystander activation would be more likely to result in cells with random, rather than specific, TCRs

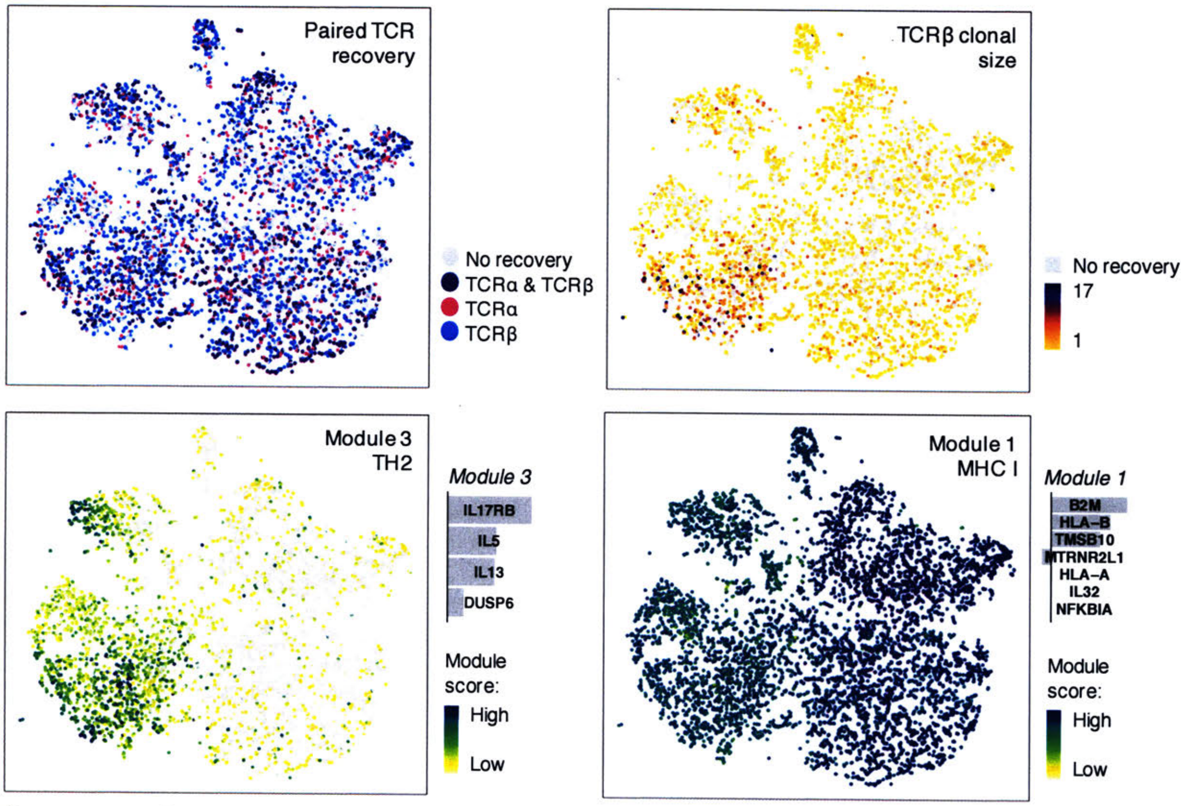


Figure 5-4. Qualitative association between clonal expansion and gene expression. tSNE visualization of all cells, colored by: paired TCR sequence recovery for each cell (top left); clonal size, defined as the number of cells sharing the same TCRß sequence (top right); expression of module 3, summarizing TH2 function (bottom left); and expression of module 1, summarizing MHC I and other genes (bottom right).

To concisely quantify the extent of overlap between gene module expression and specific TCR sequences, we calculated the 'surprisal' metric described in Chapter 3 of this thesis. (This metric represents the gain in information about a cell's module expression due to its TCRb sequence; essentially it evaluates the tightness of association between specific TCR sequences and module expression.) We also assessed the average relative clonal expansion of cells expressing each module, which is important in inferring the *in vivo* expansion, and thus degree of activation to peanut, of each cell subset. From this analysis, we saw intriguing trends in the top discriminating modules (Figure 5-5). For example, in Module 1 (the module most specific for reactive-patient cells), we saw no enrichment above average for clonal expansion or for specific

TCRb clonotypes via surprisal. This could indicate an elevated bystander response in reactive cells due to higher levels of earlier peanut-specific activation. However, in Module 7, the other module specific for reactive-patient cells representing a TH1-related pathway, we saw average levels of clonal expansion, but a high degree of surprisal, i.e. a strong association with specific rather than random TCR sequences. We speculate that cells expressing this module could be peanut-specific TH1 cells that may be too young to have undergone sufficient exposures to be highly clonally expanded *in vivo*. In modules upregulated in hyporeactive patients, we again saw varied clonotypic features. Unsurprisingly, TH2 cells (Module 3) were both highly expanded and highly associated with specific TCR sequences. However, the other two modules (Module 4 and Module 8, representing ribosomal proteins and innate inflammatory pathways, respectively) were associated with below-average expansion and random association with TCR sequences. These are interesting discriminators for which there is no obvious hypothesis. Taken together, these results suggest a myriad of different T-helper functions, including possible pathogenic roles for TH1 and other non-TH2 cells, that are relevant in clinical reactivity to allergen.

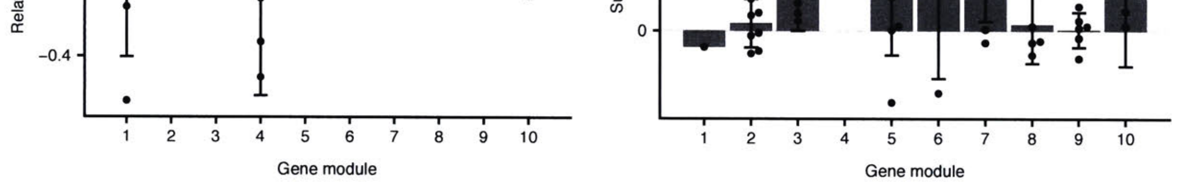


Figure 5-5. Patterns of TCR association with the top 10 gene modules. Each data point represents cells from one patient. Left: Relative clonal expansion of cells expressing each module, defined as the z-score of the TCRb expansion number for all cells within a patient that are expressing the gene module, relative to the expansion number for all cells within the patient. Right: Surprisal for each module, defined as $\ln(P/P_0)$ where P = probability of two cells within a patient both

expressing a module given that they have the same TCRb, and $P_0$ = probability of any two cells within a patient both expressing a module Surprisal represents the tightness of association between specific TCRb sequences and module expression Whether or not a cell was deemed to "express" a module was defined using a distribution-based threshold (details in Methods) Modules are annotated at the top of each plot with the sign of their coefficient in the classification model (red triangle = up in reactive patients, blue inverted triangle = up in hyporeactive patients) Points are missing for the surprisal graph whenever there are no module-scoring expanded cells in a patient (e g the entire surprisal bar is missing for Module 4 because there were no expanded cells in any patient expressing Module 4)

## 5.4. Discussion

In this study, we profiled peanut-reactive T cells from patients with differing clinical reactivities in order to discover and interpret immune correlates that could classify the patients by reactivity We observed a variety of T-helper cell states, which we organized into modules using the gene module discovery approach described previously Upon running a classifier using the module scores as features, we discovered that MHC I and TH1-related gene programs were elevated in cells from reactive patients, and that some of these cells were associated with specific, rather than random, TCRb clonotypes In hyporeactive patients we observed, surprisingly, that TH2 responses were slightly elevated and associated with highly clonal cells, and that other pathways related to innate inflammation might be upregulated

Elevated TH1 function in cells from reactive patients is an interesting result, because it suggests a possible pathogenic role for TH1 cells in severity of food allergy TH1 cells have been proposed as protective and associated with tolerance to milk allergen in children[9], but have also been described as exacerbating allergic asthma in mice[10] These results highlight the complexity and highly context-dependent results regarding non-TH2 subsets of T-helper cells in allergy

Even more surprising is the higher TH2 function in adults with lower reactivity to peanut This is counter to previous studies in infants and young children[3,4], the discrepancy suggests a possible time-dependent effect, either *in vivo* or *in vitro* While TH2 responses might correlate with or even be causal for clinical reactivity early in life, these early responses may evolve differentially in patients who remain reactive (perhaps including the emergence of TH1 cells) versus those who become less reactive Another possible reason for the discrepancy is that TH2 cells from reactive patients might indeed be present at similar frequencies, but have been activated at a different stimulation time *in vitro* In other words, the chosen time point favored a certain collection of cells and discriminating features that may not be holistically representative of all peanut-specific cells *in vivo* This selection bias for T cells that become activated at a certain rate is inherent to the experimental setup and the time-dependent nature of T-cell responses However, regardless of the selection in the milieu of peanut-reactive cells due to patient age or *in vitro* stimulation time, the classifier presented here is a biomarker of clinical reactivity and a window into the complex roles of different T-helper types in shaping the severity of a food-allergic immune response

## 5.5. Contributions and acknowledgements

This project was conducted in collaboration with Wayne Shreffler's lab at MGH The clinical trial (NCT01750879) and PBMC isolation and banking were conducted by Wayne Shreffler's clinic PBMC stimulation and sorting was done by Bert Ruiter at MGH Seq-Well experiments were performed with Andy Tu, Gary Shea, and Patrick Petrossian Paired TCR sequencing was done by Andy Tu Sequencing was done by the MIT Bio-Micro center and by the Broad Walk-Up sequencing facility This work was supported in part by the Koch Institute Support (core) Grant P30-CA14051 from the National Cancer Institute

## 5.6. References

1 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)

2 Peeters, K a B M *et al* Does skin prick test reactivity to purified allergens correlate with clinical severity of peanut allergy? *Clin Exp Allergy* **37**, 108–115 (2007)

3 Chiang, D *et al* Single-cell profiling of peanut-responsive T cells in patients with peanut allergy reveals heterogeneous effector TH2 subsets *J Allergy Clin Immunol* **141**, 2107–2120 (2018)

4 Weissler, K A *et al* Identification and analysis of peanut-specific effector T and regulatory T cells in children allergic and tolerant to peanut *J Allergy Clin Immunol* **141**, 1699-1710 e7 (2018)

5 Beyer, K *et al.* Measurement of peptide-specific IgE as an additional tool in identifying patients with clinical reactivity to peanuts *J Allergy Clin Immunol.* **112**, 202–207 (2003)

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

7 Ang A Tu *et al* Recovery of paired T cell receptors from massively-parallel 3' single-cell RNA-Seq libraries reveals clonotypic responses among antigen-activated T cells *In revision*

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

9 Turcanu, V , Maleki, S J & Lack, G Characterization of lymphocyte responses to peanuts in normal children, peanut-allergic children, and allergic children who acquired tolerance to peanuts *J Clin. Invest* **111**, 1065–1072 (2003)

10 Hansen, G , Berry, G , DeKruyff, R H & Umetsu, D T Allergen-specific Th1 cells fail to counterbalance Th2 cell–induced airway hyperreactivity but cause severe airway inflammation *J Clin Invest* **103**, 175–183 (1999)

# 6. Conclusions and outlook

The emergence of new high-throughput single-cell techniques, and a concomitant decrease in their cost, represents an exciting new set of directions for biomedical studies In particular, clinically-applied research now benefits from the ability to deeply profile and track rare cell subsets, identify intercellular communication events that may be dysregulated in disease, and discover entirely unknown biomarkers or drug targets But as the use of these techniques grows, an increasing challenge is how to leverage these data to help rather than hurt statistical power and biological interpretability While there is some precedent on cleverly utilizing single-cell data to maximize reproducibility and interpretability, thinking on how best to perform quantitative analyses of single-cell data is still emerging

In Part I of this thesis, I demonstrated two new bioinformatic approaches for improving the reproducibility and meaningfulness of single-cell high-throughput data analyses Each approach was applied to a case study with human samples in order to highlight its advantages, features, use cases, and limitations The first approach was a method for gene module discovery, which aimed to reduce the inherent noise, murkiness, and propensity for false-positives of analyses involving individual genes in single-cell RNA sequencing data, by meaningfully compressing them into co-expressed genes We demonstrated the usefulness of this approach on longitudinal samples from pediatric milk allergy patients We observed that several gene modules neatly recapitulated known T-cell programs such as TH2, TH17, and MHC-II upregulation, and that in milk-reactive T cells, TH1 and NF-kB gene modules increased in expression as the individual aged (regardless of whether they outgrew their allergy or not) This case study highlighted the usefulness of the modules method, both in compressing the feature space without loss of information, and in making longitudinal changes easier to interpret The second bioinformatic approach was a framework for integrating multiple modalities of single-cell data and quantitatively assessing their degree of overlap The approach utilized probability-based metrics in a new way to intuitively capture the association between two modalities of single-cell data We applied this framework to a dataset of paired single-cell RNA sequences and TCR sequences of CMV-reactive T cells from two adults, to investigate whether TCR sequence was associated with T cell state We found that there was in fact a strong, time-dependent association, and that a highly-activated, TCR-selective state existed at early stimulation times These results suggested that there was in fact a significant association between epitope specificity and T cell fate, which is an intriguing finding for modulating T cell fates

In Part II of this thesis, the bioinformatic tools developed in Part I were integrated as part of a holistic approach for analyzing single-cell data from two larger clinical studies in food allergy, with the goal of better understanding acquired tolerance and clinical reactivity in peanut allergy In the first study, we profiled T cells from patients undergoing peanut oral immunotherapy and discovered two subsets of peanut-reactive TH2 cells with completely distinct TCR repertoires Both subsets were transiently suppressed, but not deleted, by the treatment, corroborating the historical impermanence of OIT-induced desensitization In the second study, we profiled T cells from peanut-allergic patients with high and low clinical reactivity to peanut allergen to discover correlates of clinical status Using single-cell RNA-Seq and paired TCR sequencing, we surprisingly noticed a strong upregulation in TH1 and TH17 pathways in more reactive patients Cells expressing these modules had average levels of clonal expansion and upregulation of antiviral pathways, suggesting a role for non-TH2 cells in promoting or enhancing pathogenic responses to allergen In both studies, the combined application of novel experimental and bioinformatic techniques allowed for new disease-relevant insights

The work described in this thesis hopefully provides precedent and guidelines on how to make effective use of single-cell data, especially in small clinical studies where low sample numbers are a concern for achieving reproducible results The aim is that these methods, and the results generated from applying them to various cohorts, are broadly useful to the enhanced understanding of disease through single-cell profiling The methods presented here, however, are by no means an exhaustive set of solutions to the challenges faced in analyzing high-throughput single-cell data of clinical samples Numerous hurdles are still to be overcome, including the following 1) *Biases in library recovery and cell state due to batch effects* An often-unavoidable feature of clinical studies, batch effects are inherent whenever experimental factors like user, sample holding or processing time, or protocol differ even slightly These are especially evident in transcriptional profiling studies, as mRNA levels are fast-changing and thus sensitive to batch effects Batch-effect correction algorithms exist for single-cell RNA sequencing data, but are still imperfect and have the possibility of introducing artifacts into the data 2) *A lack of systematic ways to identify cell types in high-throughput data* Highly specific subsets of cells have been well-described at the protein level using flow cytometry and histology, but these described states do not always map cleanly to states observed by transcriptional or other profiling techniques Matching observed states in a dataset with described cell types in the literature is usually an arduous, manual process that can often still be inconclusive 3) *The ubiquitous challenge of interpreting gene or protein hits that are not well-annotated.* Despite the exponentially growing presence of published gene sets and databases, many genes, proteins, and metabolites are still not well-annotated and therefore difficult to interpret if they emerge as top hits in an analysis Additionally, it appears that the function of genes is highly context- and tissue-specific, so findings are often not generalizable from one study to the next This is an ongoing challenge that will slowly be addressed with more studies in basic biology as well as more robust analysis techniques for comparing findings across published datasets

# 7. Appendix

## Appendix A2. Gene module discovery in single-cell RNA sequencing data

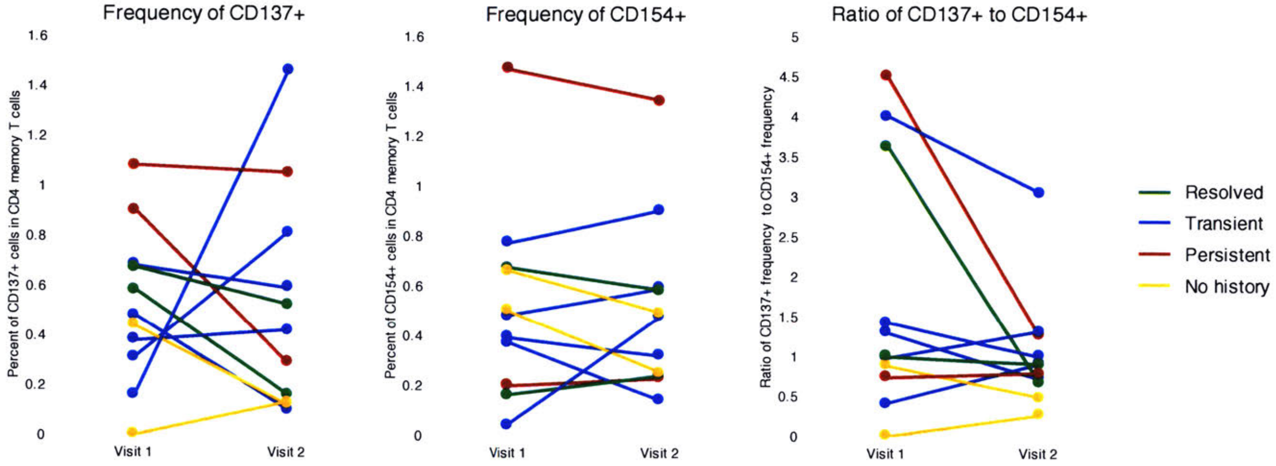


Figure A2-1. Changes in frequencies of activated T cells from pediatric milk-allergic patients. Frequencies of CD137+ and CD154+ cells (and the ratio of the two) among CD4 memory T cells is plotted against time. Each line represents a patient's cells, colored by their status: Resolved (history of milk allergy but no diagnostics signs of it at either time point); Transient (diagnosis status changed between visits 1 and 2); Persistent (patient had the same level of milk allergy at both visits); and No history (patient had no history of milk allergy).

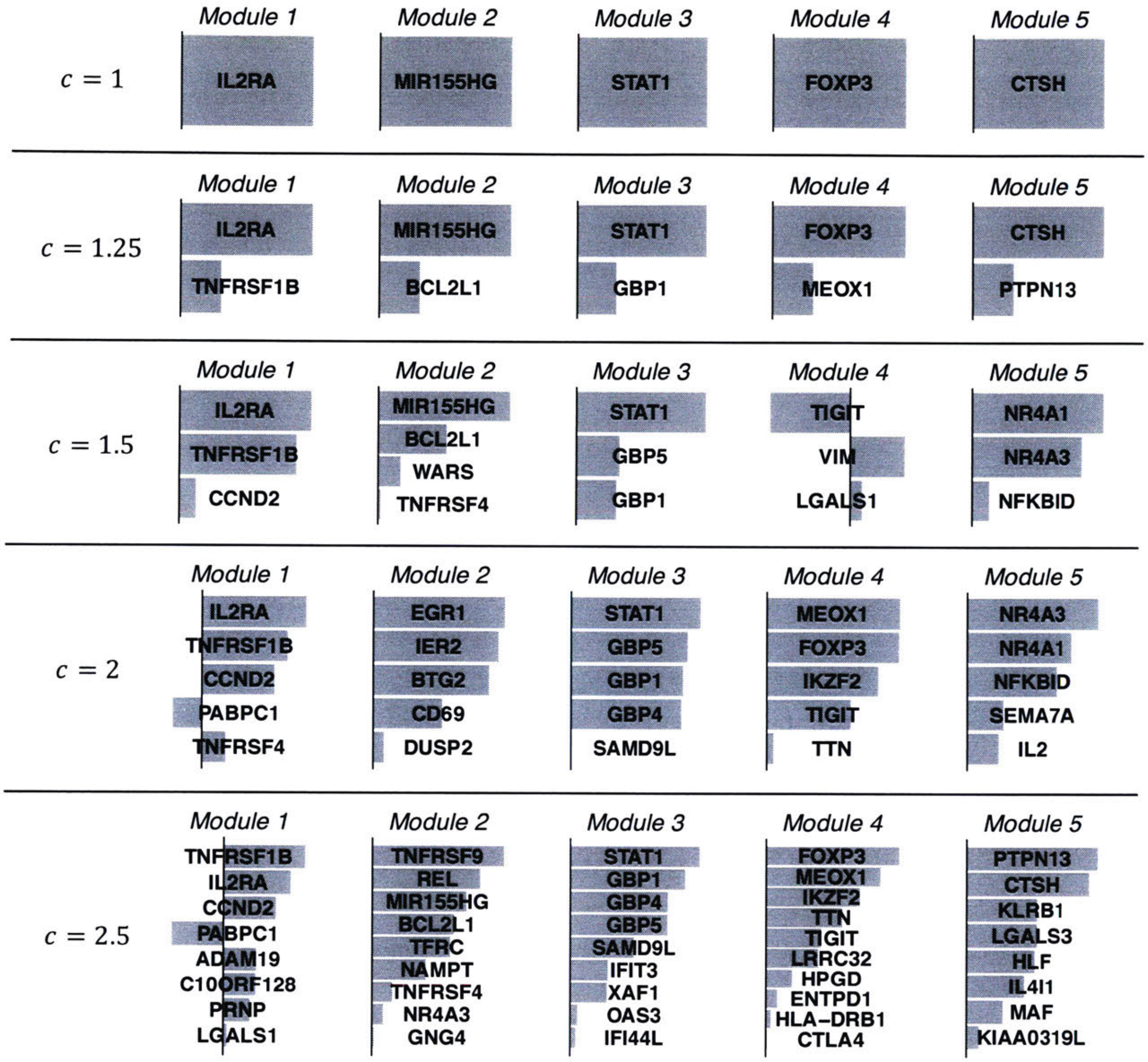


Figure A2-2. Effect of tuning parameter on gene modules. The tuning parameter, *c*, is the sum of the absolute value of all gene weights allowed in each component. The top five gene modules generated from sparse PCA runs with five different tuning parameter settings are shown. In each bar graph, the magnitude and direction of each bar represent the weight and sign of the relevant gene in the module. Scales may not be the same between bar graphs.

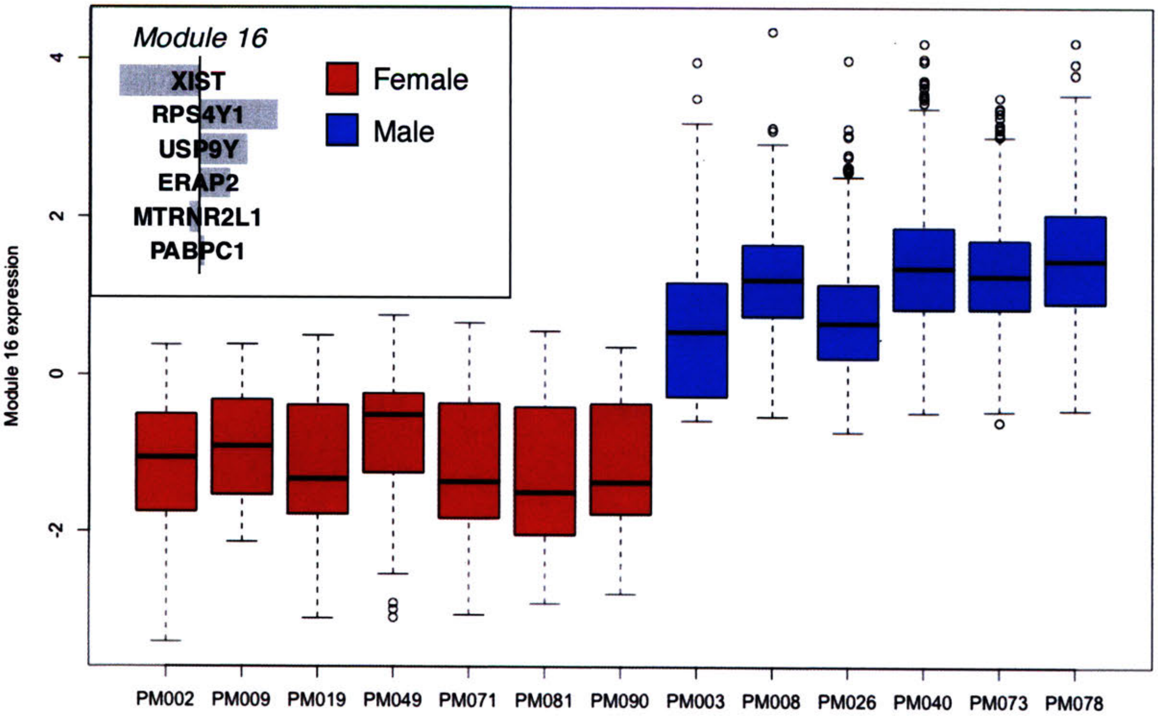


Figure A2-3. Module 16 expression in cells by patient, colored and grouped by gender. All cells from both time points are included for each patient. Inlay: Module 16 gene loadings; the magnitude and direction of each bar represent the weight and sign of the relevant gene in the module.

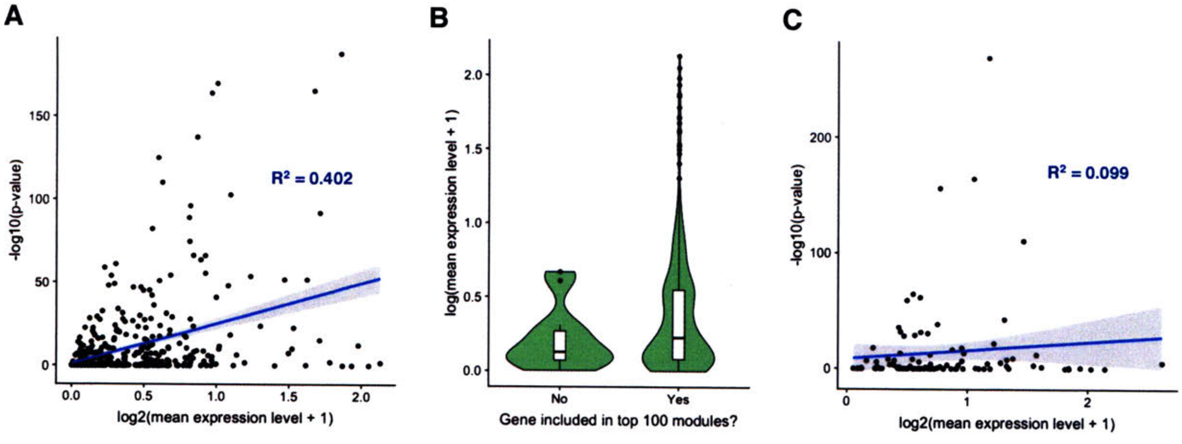


Figure A2-4. Relative bias for highly-expressed transcripts in genes and modules. Data is from the four persistent donors, and the statistical comparison being made is between all cells at visit 1 and all cells at visit 2. A) Adjusted p-value versus log-transformed mean expression of every gene that was used as input to the sparse PCA. B) Mean expression of genes that were, or were not, present among the loadings of the top 100 gene modules. C) Adjusted p-value versus log-transformed mean expression of the top 100 gene modules.

## Appendix A3. Analysis of multimodal single-cell data of antigen-reactive T cells

Table A3-1. Demographic information and CMV reactivity of donors.

| Donor | CTL ID | Age | Gender | Race | ELISPOT with CMV pp65 peptide pool (Class I/II) |
|---|---|---|---|---|---|
| 1 | LP_58 | 54 | F | Hispanic | 861 |
| 2 | LP_335 | 39 | F | Hispanic | 707 |

Table A3-2. HLA Class II allele information of CMV-reactive donors.

| Donor | DRB 1 | DRB 3/4/5 | DQA | DQB | DPA | DPB |
|---|---|---|---|---|---|---|
| 1 | DRB1*04:01<br>DRB1*08:02 | DRB4*01:03 | DQA1*03:03<br>DQA1*04:01 | DQB1*03:01<br>DQB1*04:02 | not tested | not tested |
| 2 | DRB1*01:02<br>DRB1*11:02 | DRB3*02:02 | not tested | DQB1*03:01<br>DQB1*05:01 | DPA1*01:03<br>DPA1*02:01 | DPB1*02:01G<br>DPB1*14:01G |

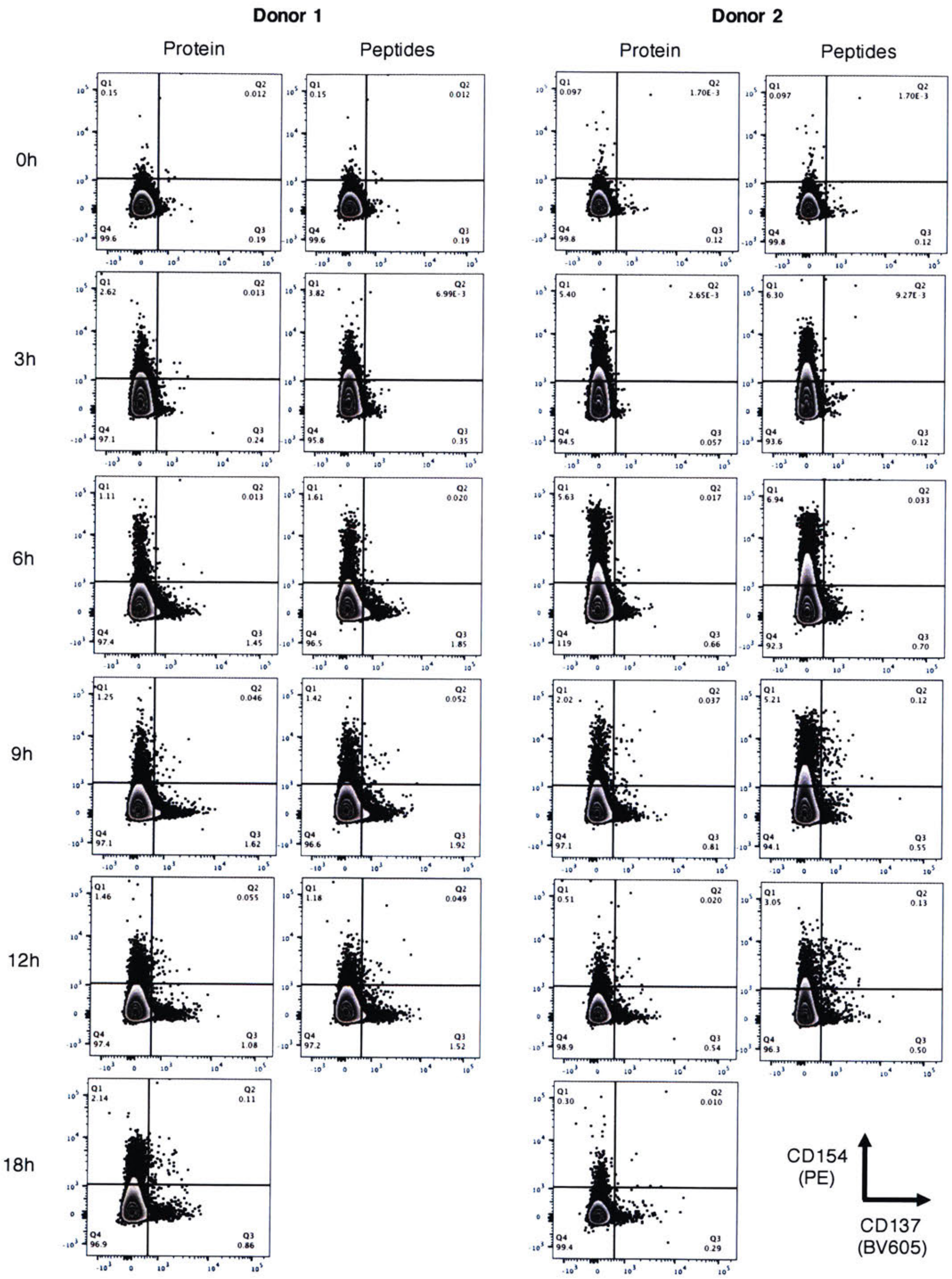


Figure A3-1. Activation markers upregulated on T-helper cells in response to CMV stimulation. Cells are pre-gated on lymphocytes, singlets, live CD3+ cells, and CD4+CD45RA- cells. On the x-axis is the MFI of CD137 signal, and on the y-axis is the MFI of CD154 signal. Rows represent stimulation times, and columns represent stimuli. “Protein” refers to CMVpp65 whole protein, and “Peptides” refers to an overlapping peptide pool of the CMVpp65 amino acid sequence. The 18h condition was omitted for the peptides as it was expected that activation in response to

peptides would occur earlier. Each set of two columns represents cells from one healthy, CMV-reactive donor.

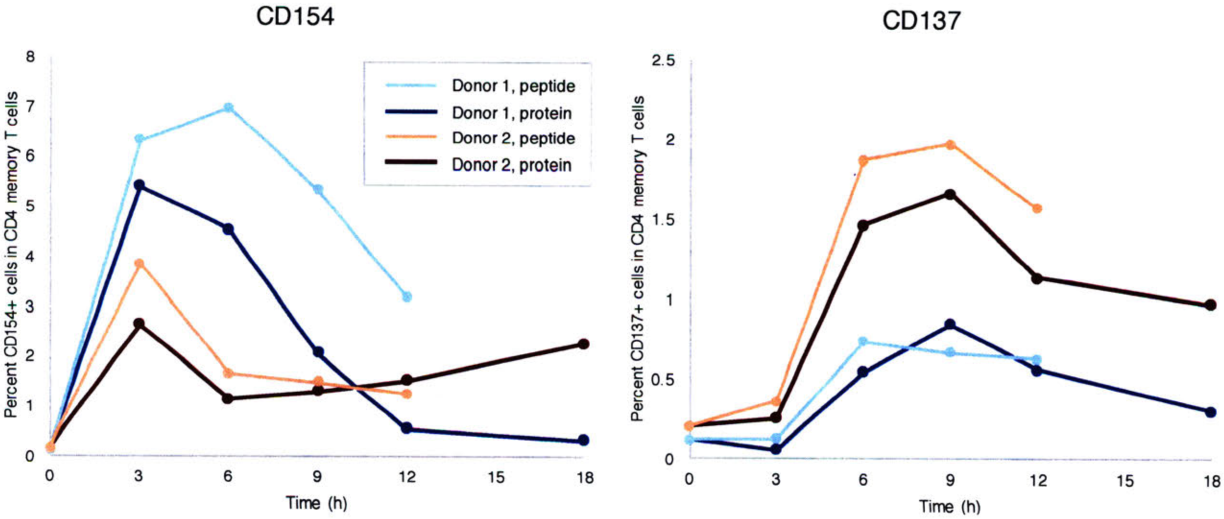


Figure A3-2. Dynamics of CD154 and CD137 expression in response to CMV stimulation. Frequency of CD154+ (left) and CD137+ (right) cells among CD4 memory T cells as a function of stimulation time with CMV antigen. "Peptide" refers to an overlapping peptide pool for CMVpp65 and "protein" refers to CMVpp65 whole protein. (Note that any CD154+CD137+ cells, which are rare, are counted in both graphs.)

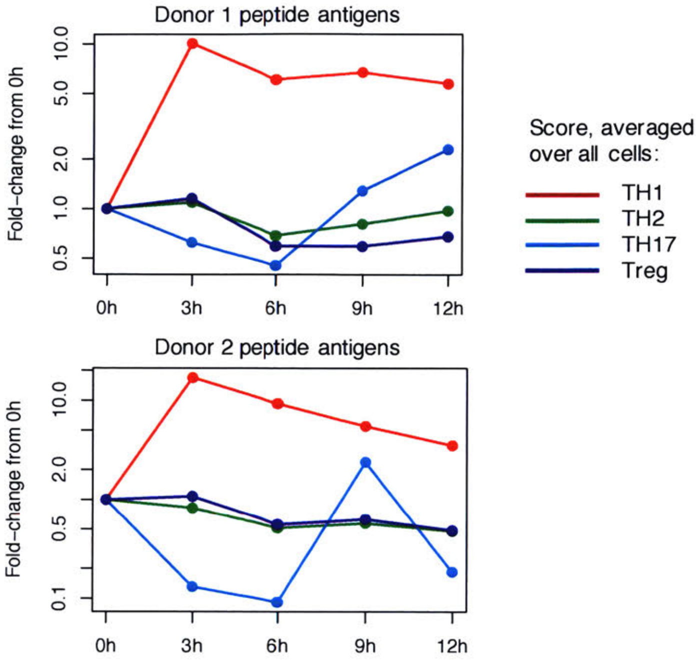


Figure A3-3. Patterns of T-helper gene expression as a function of time. T-helper expression levels were assessed for all peptide-stimulated cells, and fold-change of mean score with respect to 0h is plotted for cells from each donor.

## Appendix A4. Transient suppression of TH2 clonotypes in peanut OIT

Table A4-1. Patient demographics and baseline characteristics.

| Patient ID | Treatment Group | Age | Gender | Race | Peanut IgE (kU/L) | Peanut IgG4 (kU/L) | Total IgE (kU/L) | Skin prick test, adjusted (mm) |
|---|---|---|---|---|---|---|---|---|
| 105 | Treatment | 22 | Female | White | 44.4 | 0.49 | 109 | 28 |
| 106 | Treatment | 32 | Female | White | 4.5 | 0.29 | 40.8 | 10 |
| 111 | Treatment | 22 | Female | White | 20.9 | 0.16 | 169 | 5 |
| 33 | Treatment | 16 | Female | White | 84.1 | 0.54 | 216 | 10 |
| 90 | Treatment | 9 | Male | White | 159 | 0.85 | 338 | 14.5 |
| 93 | Treatment | 11 | Male | White | 40.9 | 1.86 | 208 | 7.5 |
| 69 | Treatment | 15 | Male | White | 11.2 | 0.16 | 141 | 13 |
| 95 | Treatment | 8 | Male | White | 451 | 1.71 | 1524 | 21 |
| 97 | Treatment | 36 | Female | Asian | 2.6 | 0.62 | 339 | 13.5 |
| 84 | Placebo | 22 | Male | White | 61.4 | 0.09 | 174 | 11 |
| 96 | Placebo | 10 | Female | White | 39.1 | 0.18 | 151 | 10 |
| 107 | Placebo | 22 | Female | White | 27.3 | 0.37 | 88.1 | 22 |

Table A4-2. Clinical outcomes.

| Patient ID | Treatment Group | Cumulative dose consumed at DBFC2 | DBFC2 outcome | Cumulative dose consumed at DBFC3 | DBFC3 outcome | Adverse event count | Therapeutic outcome |
|---|---|---|---|---|---|---|---|
| 105 | Treatment | 4443 | Pass | 4440 | Pass | 605 | Tolerance |
| 106 | Treatment | 4443 | Pass | 4440 | Pass | 269 | Tolerance |
| 111 | Treatment | 4443 | Pass | 4440 | Pass | 61 | Tolerance |
| 33 | Treatment | 4443 | Pass | 4440 | Fail | 306 | Partial tolerance |
| 90 | Treatment | 4440 | Pass | 4440 | Fail | 60 | Partial tolerance |
| 93 | Treatment | 4443 | Pass | 4440 | Fail | 177 | Partial tolerance |
| 69 | Treatment | 943 | Fail | 40 | Fail | 101 | Treatment failure |
| 95 | Treatment | 4443 | Fail | 440 | Fail | 26 | Treatment failure |
| 97 | Treatment | 289.6 | Fail | 1440 | Fail | 497 | Treatment failure |
| 84 | Placebo | 443 | Fail | -- | -- | 81 | Treatment failure |
| 96 | Placebo | 143 | Fail | -- | -- | 42 | Treatment failure |
| 107 | Placebo | 943 | Fail | -- | -- | 25 | Treatment failure |

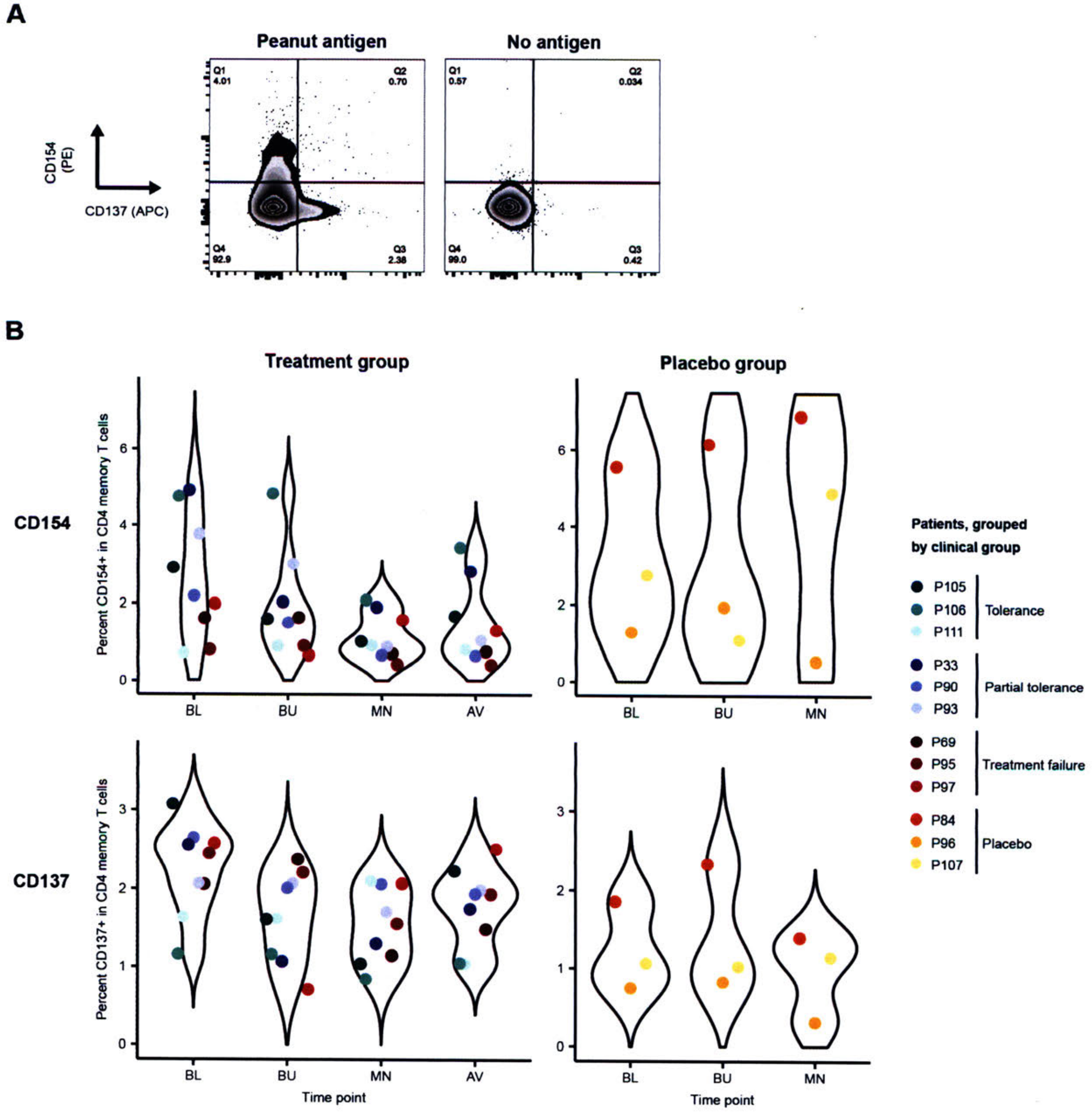


Figure A4-1. Summary of flow-based enrichment of CD154+ and CD137+ peanut-reactive T cells. A) Representative flow plot of CD154 and CD137 expression in cells stimulated with peanut antigen (left) or no antigen (right). B) Percent of CD4 memory T cells at each time point that are CD154+ (top) or CD137+ (bottom), within patients of the treatment group (left) or placebo group (right).

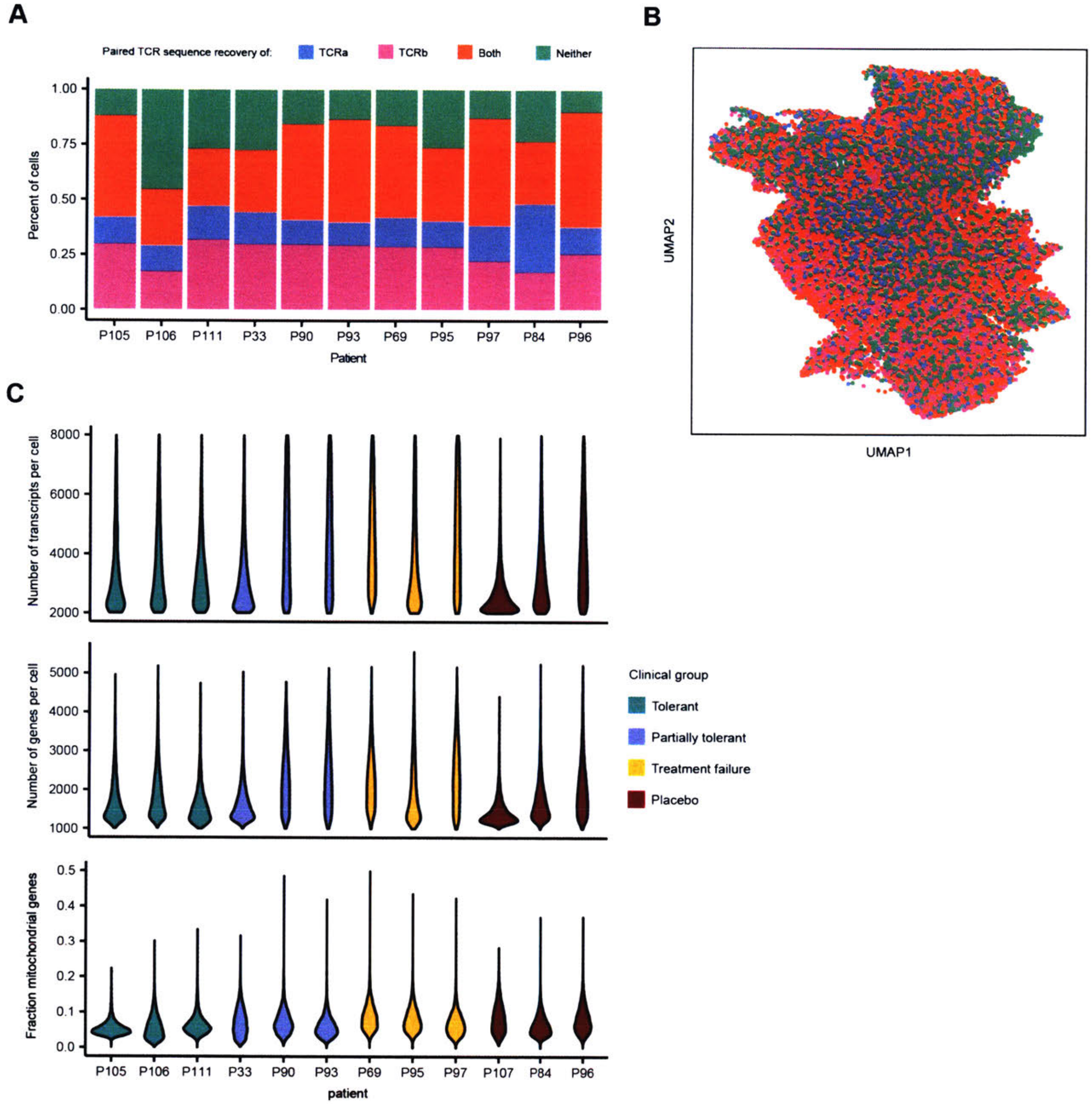


Figure A4-2. Quality of single-cell libraries recovered. A) Distribution of recoveries of paired TCR sequences for all cells, grouped by patient. B) Overlay of TCR recovery status of each cell onto the UMAP visualization. C) Quality control metrics of single-cell RNA sequencing libraries: number of UMIs per cell (top), number of genes detected per cell (middle), and fraction of genes detected that are mitochondrial for each cell (bottom), grouped by patient and colored by clinical group.

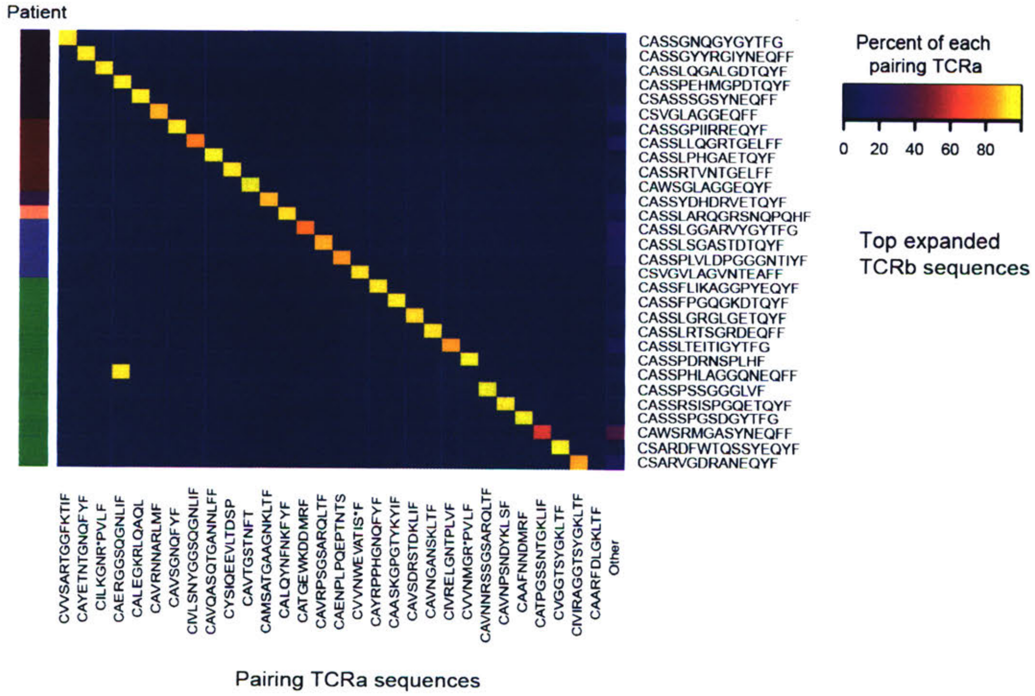


Figure A4-3. TCRa pairing for top expanded TCRb sequences. Heatmap of TCRa pairing sequences (columns) found in cells with the top expanded TCRb sequences (rows). Within each TCRb clonotype, the percent of cells mapping to each TCRa is plotted. Rows are annotated with the majority patient in which the TCRb clonotype was detected.

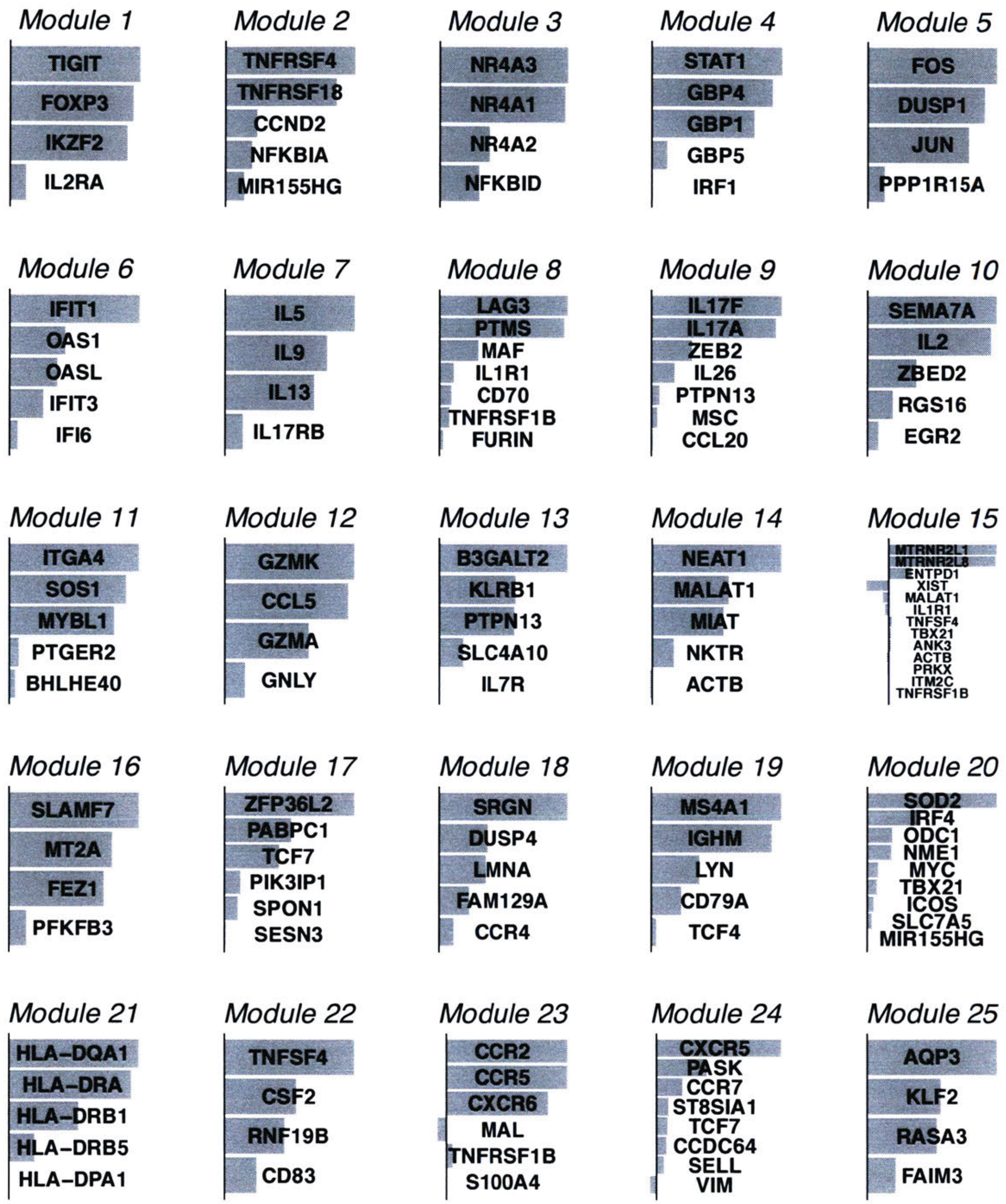


Figure A4-4. Top 25 gene modules identified by unsupervised sparse PCA approach. Modules are sorted by percent variance explained. The magnitude and direction of each bar represent the weight and sign of each gene in each component. Gene expression values are scaled.

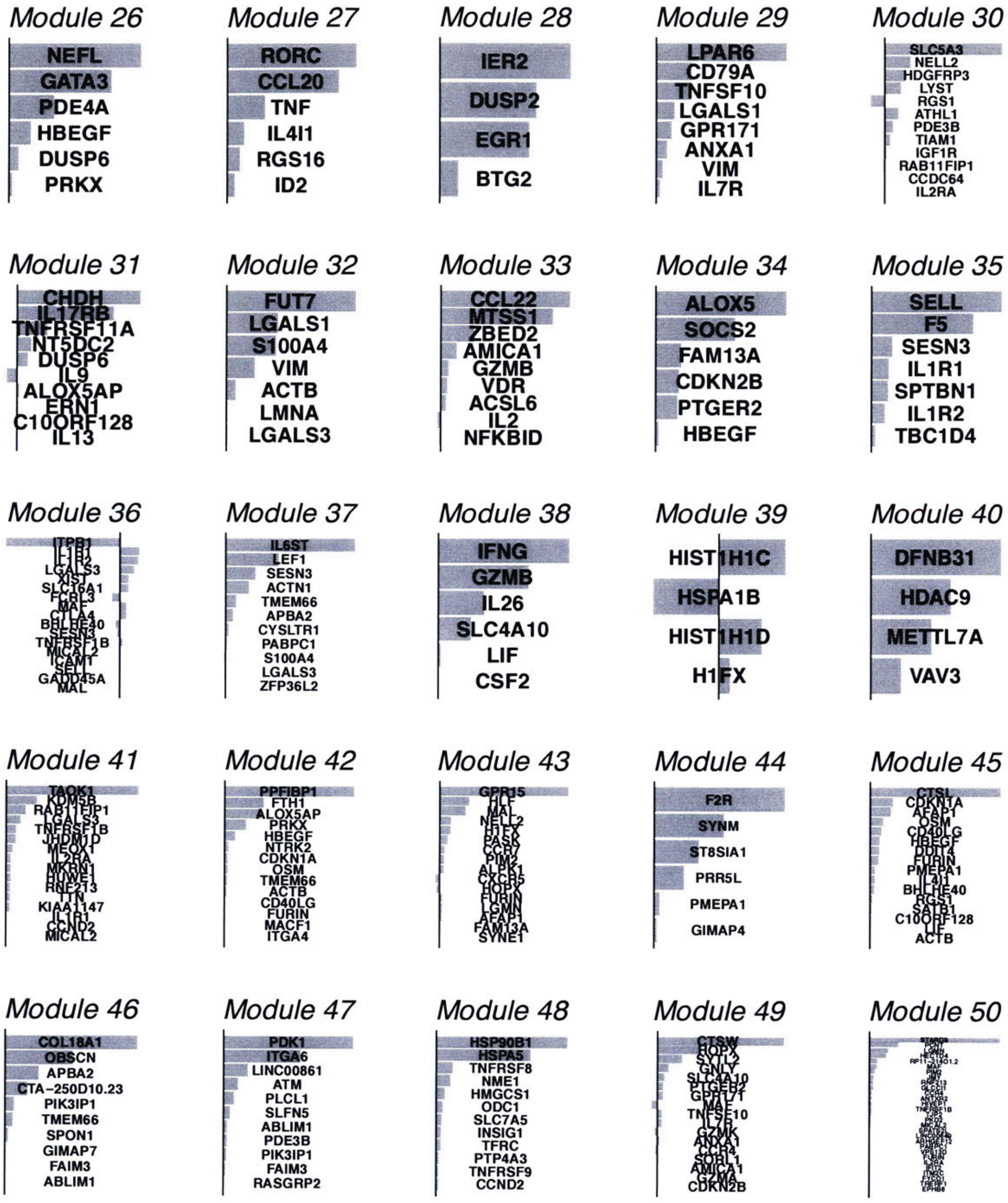


Figure A4-5. Top 26 through 50 gene modules identified by unsupervised sparse PCA approach. Modules are sorted by percent variance explained. The magnitude and direction of each bar represent the weight and sign of each gene in each component. Gene expression values are scaled.

## Appendix A5. Correlates of clinical reactivity to peanut allergen

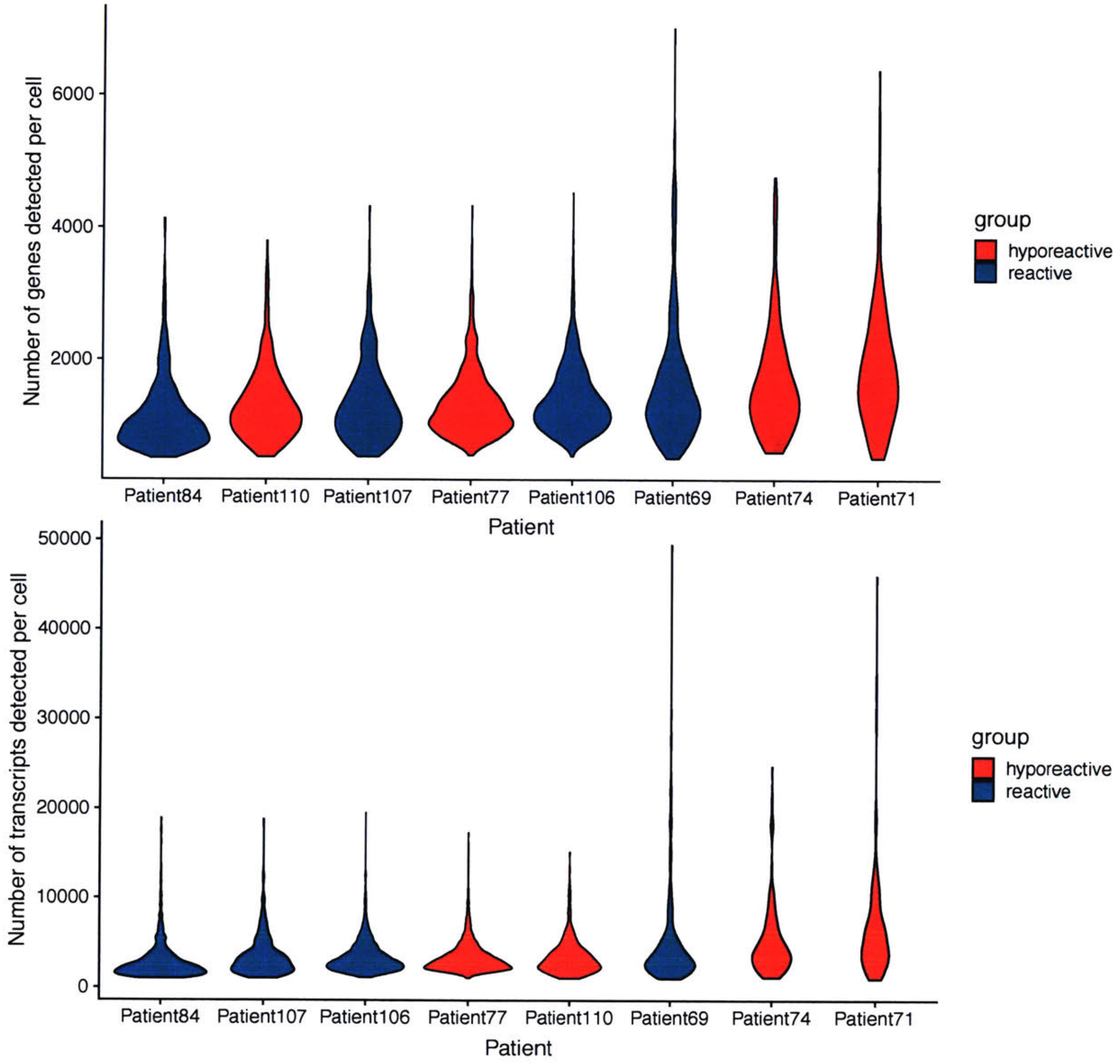
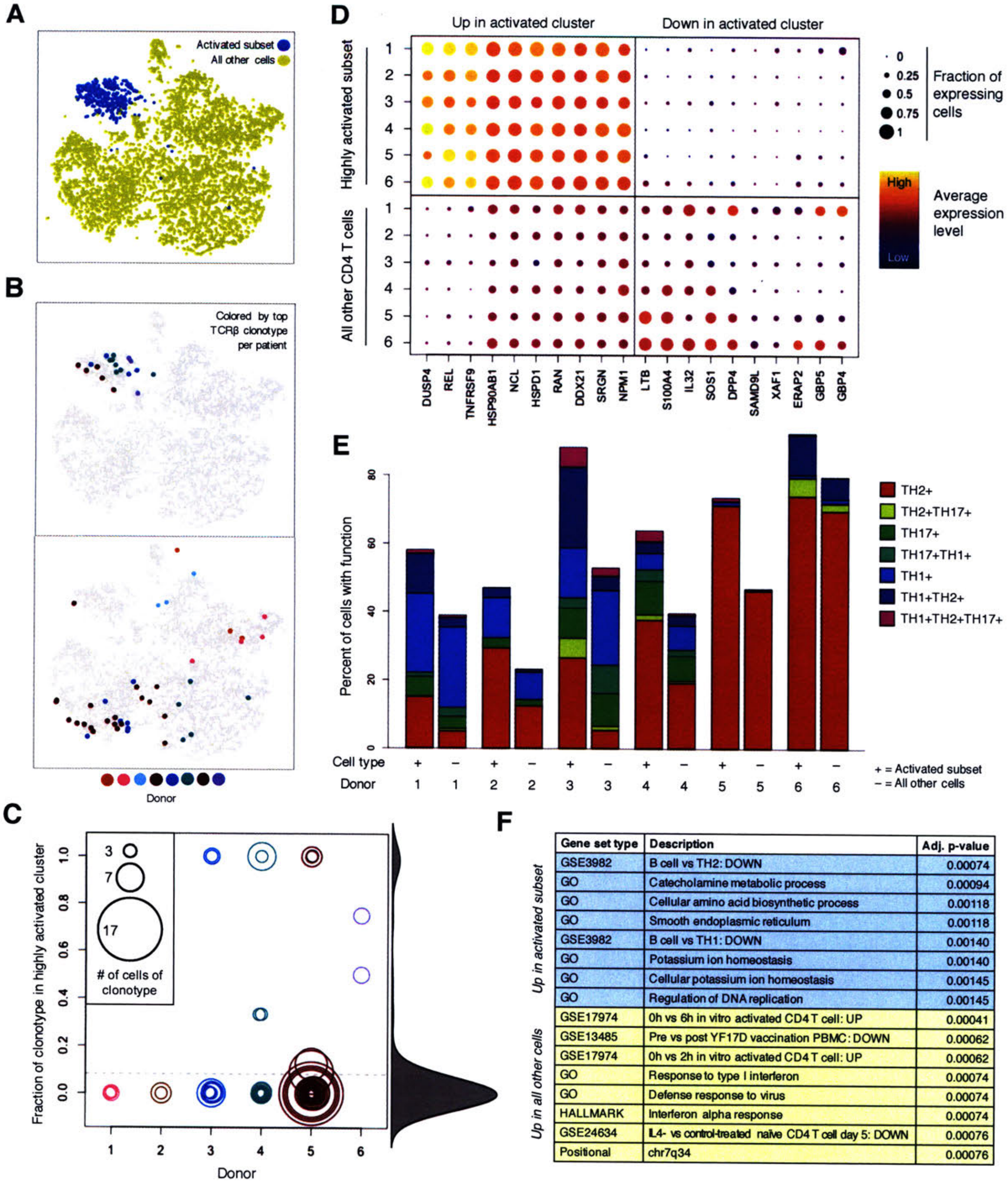


Figure A5-1. Quality metrics for single-cell RNA sequencing libraries from eight peanut-allergic patients. Violin plot of number of genes detected per cell (top) and number of transcripts detected per cell (bottom), by patient, and patients are sorted by mean gene or transcript recovery. Violins are colored by patient's clinical reactivity to peanut.



| | Gene set type | Description | Adj. p-value |
|---|---|---|---|
| Up in activated subset | GSE3982 | B cell vs TH2: DOWN | 0.00074 |
| | GO | Catecholamine metabolic process | 0.00094 |
| | GO | Cellular amino acid biosynthetic process | 0.00118 |
| | GO | Smooth endoplasmic reticulum | 0.00118 |
| | GSE3982 | B cell vs TH1: DOWN | 0.00140 |
| | GO | Potassium ion homeostasis | 0.00140 |
| | GO | Cellular potassium ion homeostasis | 0.00145 |
| | GO | Regulation of DNA replication | 0.00145 |
| Up in all other cells | GSE17974 | 0h vs 6h in vitro activated CD4 T cell: UP | 0.00041 |
| | GSE13485 | Pre vs post YF17D vaccination PBMC: DOWN | 0.00062 |
| | GSE17974 | 0h vs 2h in vitro activated CD4 T cell: UP | 0.00062 |
| | GO | Response to type I interferon | 0.00074 |
| | GO | Defense response to virus | 0.00074 |
| | HALLMARK | Interferon alpha response | 0.00074 |
| | GSE24634 | IL4- vs control-treated naïve CD4 T cell day 5: DOWN | 0.00076 |
| | Positional | chr7q34 | 0.00076 |

Figure A5-2. Identification of a highly-activated, clonotypically distinct subset of peanut-reactive T cells. A) tSNE projection of all cells, colored by membership in highly activated cluster. B) Top TCRβ clonotype from each patient for which the majority is present in the highly activated cluster (top) or the majority is present outside highly activated cluster (bottom). Analysis was restricted to expanded clonotypes with 3+ cells. C) Fraction of each expanded clonotype (3+ cells) present in the highly activated cluster. Each clonotype is a circle, with the size of the circle indicating the number of cells in the clonotype. At right, a density curve shows the distribution of all fractions on

a per-cell basis. Analysis was restricted to six of the eight donors, for whom there were at least 20 cells detected in the highly activated subset. D) Plot of top 20 most discriminating genes (using a ROC test) upregulated inside and outside the activated cluster. Analysis was restricted to the same six donors as before. Each dot represents all cells from a donor either inside or outside the highly activated subset, with the color representing the mean gene expression level and the size representing the percent of cells expressing the gene. E) T-helper function inside and outside the highly activated cluster. Cells were scored for TH1, TH2, and TH17 function using the sum of expression levels of the genes IFNG and TNF (TH1), IL13, IL5, IL4 and IL9 (TH2), and IL17A and IL17F (TH17). Cells were scored as having a function if the score was greater than zero. Each bar is normalized to sum to 100%, with the cells not expressing any function not shown. F) GSEA results for top eight gene sets enriched in cells inside and outside the highly activated subset. Adjusted p-value (adjusted by GSEA's permutation approach) is shown.

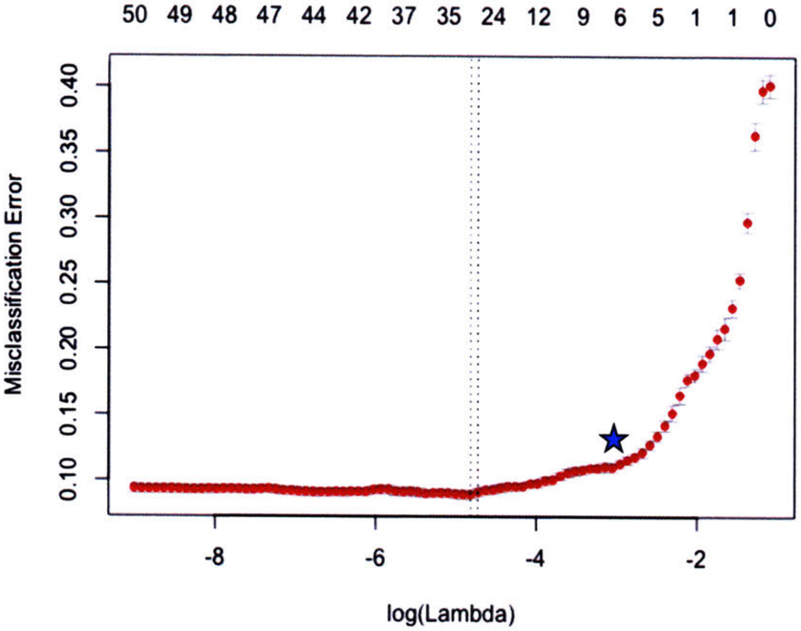
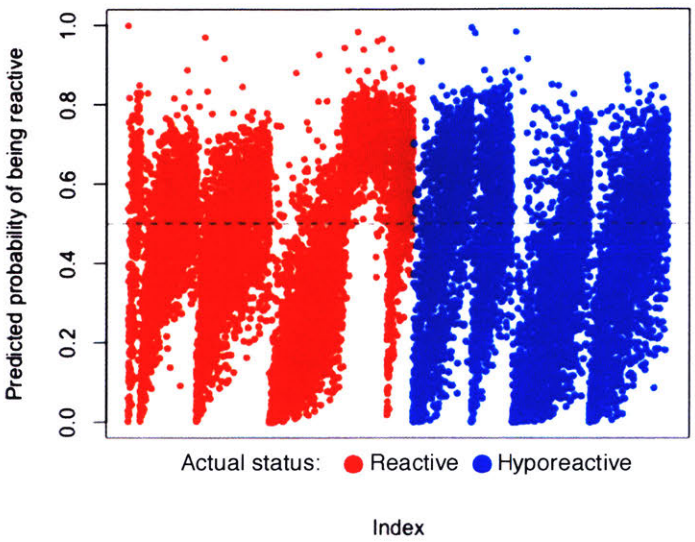


Figure A5-3. Cross-validation of logistic regression model for classifying cells based on clinical reactivity of the patient. Top 50 gene modules were used as input. Shown is the misclassification rate of a holdout set of cells using 10-fold cross-validation, at various settings of lambda (tuning parameter). Across the top are the number of gene modules (or features) retained in the classifier in each case. The dotted lines denote the recommended selection of lambda based on the minimum value, and the star indicates the value of lambda that was actually chosen. The reason for the discrepancy was a desire to produce as parsimonious a model as possible that would not be vulnerable to overfitting, while tolerating a slight increase in misclassification error.



| Donor | Status | Misclass rate |
|---|---|---|
| 1 | Reactive | 66.6% |
| 2 | Reactive | 24.5% |
| 3 | Reactive | 98.3% |
| 4 | Reactive | 53.3% |
| 5 | Reactive | 42.5% |
| 6 | Hyporeactive | 50.1% |
| 7 | Hyporeactive | 80.1% |
| 8 | Hyporeactive | 61.8% |
| 9 | Hyporeactive | 53.7% |
| | **OVERALL:** | **54%** |

Figure A5-4. Testing of classifier in cells from an independent test cohort. Logistic regression model developed earlier was tested on an independent cohort, profiled by single-cell RNA sequencing one year later. The cohort consisted of five reactive and four hyporeactive donors. Left: Classification success for all cells in the test set (ordered by patient on the x-axis). Cells are colored by their true status (reactive or hyporeactive patient) and the y-axis shows the predicted probability of the cells coming from a reactive patient. Cells with a probability of above 0.5 were classified as "reactive", and below as "hyporeactive". The overall misclassification rate was 54%. Right: Misclassification rates for cells in the independent cohort, by patient.