JCI insight

Identification of antigen-specific TCR sequences based on biological and statistical enrichment in unselected subjects

Neal P. Smith, … , J. Christopher Love, Wayne G. Shreffler

JCI Insight. 2021. https://doi.org/10.1172/jci.insight.140028.

Resource and Technical Advance In-Press Preview **Immunology**

Recent advances in high-throughput T cell receptor (TCR) sequencing have allowed for new insights into the human TCR repertoire. However, methods for capturing antigen-specific repertoires remain an area of development. Here, we describe a potentially novel approach that utilizes both a biological and statistical enrichment to define putatively antigen-specific complementarity-determining region 3 (CDR3) repertoires in unselected individuals. The biological enrichment entails fluorescence-activated cell sorting of in vitro antigen-activated memory CD4+ T cells, followed by TCRβ sequencing. The resulting TCRβ sequences are then filtered by selecting those that are statistically enriched when compared to their frequency in the autologous resting T cell compartment. Applying this method to define putatively peanut protein-specific repertoires in 27 peanut-allergic individuals resulted in a library of 7345 unique CDR3β amino acid sequences that had similar characteristics to other validated antigen-specific repertoires in terms of homology and diversity. In-depth analysis of these CDR3βs revealed 36 public sequences that demonstrated high levels of convergent recombination. In a network analysis, the public CDR3βs unveiled themselves as core sequences with more edges than their private counterparts. This method has the potential to be applied to a wide range of T cell-mediated disorders, and to yield new biomarkers and biological insights.

Title

Identification of antigen-specific TCR sequences based on biological and statistical enrichment in unselected subjects

Authors

Neal P. Smith[1]*, Bert Ruiter[1,2]*, Yamini V. Virkud[1, 2, 3], Ang A. Tu[4], Brinda Monian[4], James J. Moon[1, 2 ,5], J. Christopher Love[4], Wayne G. Shreffler[1, 2, 3]

Affiliations

[1] Center for Immunology & Inflammatory Diseases, Massachusetts General Hospital, Boston, MA.

[2] Harvard Medical School, Boston, MA.

[3] Food Allergy Center, Massachusetts General Hospital, Boston, MA.

[4] Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA.

[5] Division of Pulmonary and Critical Care Medicine, Massachusetts General Hospital, Boston, MA

*These authors contributed equally to this work

Corresponding author:

Neal P. Smith
Massachusetts General Hospital
Center for Immunology & Inflammatory Diseases
Room 8.215
149 13th Street
Boston, MA 02129
nsmith19@mgh.harvard.edu

Conflict of interest statement

JCL has interests in Sunflower Therapeutics PBC, Pfizer, Honeycomb Biotechnologies, OneCyte Biotechnologies, SQZ Biotechnologies, Alloy Therapeutics, QuantumCyte, Amgen, and Repligen. JCL's interests are reviewed and managed under Massachusetts Institute of Technology's (MIT's) policies for potential conflicts of interest. JCL receives sponsored research support at MIT from Amgen, the Bill & Melinda Gates Foundation, Biogen, Pfizer, Roche, Takeda, and Sanofi. The spouse of JCL is an employee of Sunflower Therapeutics PBC.

WGS has interests in Aimmune Therapeutics, Allergy Therapeutics, FARE, DBV Biotechnologies, Merck, and Sanofi. WGS's interests are reviewed and managed under Mass General Brigham's policies for potential conflicts of interest. WGS receives sponsored research support at MGH from Aimmune, DBV, The Demarest Lloyd Foundation, The Food Allergy Science Initiative at the Broad Institute, Regeneron, and Vedanta.

Abstract

Recent advances in high-throughput T cell receptor (TCR) sequencing have allowed for new insights into the human TCR repertoire. However, methods for capturing antigen-specific repertoires remain an area of development. Here, we describe a potentially novel approach that utilizes both a biological and statistical enrichment to define putatively antigen-specific complementarity-determining region 3 (CDR3) repertoires in unselected individuals. The biological enrichment entails fluorescence-activated cell sorting of *in vitro* antigen-activated memory $CD4^+$ T cells, followed by TCRβ sequencing. The resulting TCRβ sequences are then filtered by selecting those that are statistically enriched when compared to their frequency in the autologous resting T cell compartment. Applying this method to define putatively peanut protein-specific repertoires in 27 peanut-allergic individuals resulted in a library of 7345 unique CDR3β amino acid sequences that had similar characteristics to other validated antigen-specific repertoires in terms of homology and diversity. In-depth analysis of these CDR3βs revealed 36 public sequences that demonstrated high levels of convergent recombination. In a network analysis, the public CDR3βs unveiled themselves as core sequences with more edges than their private counterparts. This method has the potential to be applied to a wide range of T cell-mediated disorders, and to yield new biomarkers and biological insights.

Introduction

T cells are defined by their antigen-specific T cell receptor (TCR), and the collection of all TCRs in a human, which is dynamic and comprises approximately 10^{10} unique TCRs at a given time, is known as their TCR repertoire (1). TCRs are dimeric proteins comprising of an α and β chain that are both generated through the process of genomic rearrangement of germline V (Variable), D (Diversity) and J (Joining) genes concurrent with random nucleotide insertions and deletions in the VDJ junction. This junction, known as the CDR3 (complementarity-determining region 3), interacts most closely with an epitope during antigen presentation, and is therefore the primary focus of studies aiming to elucidate the mechanisms of TCR specificity (2).

To date, *ex vivo* analysis of antigen-specific TCRs has largely used selection of T cells by peptide-MHC (pMHC)-multimer (e.g. tetramer) binding and fluorescence-activated cell sorting (FACS). Of the 78,701 listed antigen-specific TCRs in the VDJdb (vdjdb.cdr3.net), 73,964 (94%) were reported to be isolated via pMHC-multimer selection (3). In addition, new methodologies aimed at defining features that confer antigen-specificity have been benchmarked on various anti-viral TCRs selected by pMHC-tetramers (2, 4). The advantage of tetramer selection is the inherent functional validity it provides, as it labels cells specific for a single well-defined epitope. However, it comes with a major drawback in the form of a limited scope of analysis. Most immune responses are poly-antigenic, and current T cell epitope mapping information of most antigens is incomplete. Additionally, pMHC-tetramers can only be used with T cell donors that have genetically matched HLA alleles. Thus, tetramer-selection with only known epitopes in specific genetic backgrounds will likely provide an incomplete understanding of the antigen-specific TCR repertoire and T cell-mediated immune responses.

Public TCRs defined on the basis of identical amino acid sequences across multiple individuals have been described in many contexts since the early 1990s. Their existence has been intriguing given the vast number of possible recombination events, which suggests the prevalence of public TCRs would be much lower than what has been observed (5, 6). One major mechanism that shapes an antigen-specific public TCR repertoire is convergent recombination. Herein, selective pressure such as that seen in response to dominant antigenic epitopes across individuals, gives rise to the presence of multiple unique nucleotide sequences producing the same 'public' amino acid CDR3, due to genetic code degeneracy and homologous gene segments (7). Defining the public TCR repertoire for antigen-specific responses can help in the development of diagnostics and therapeutics in T cell driven disorders.

Peanut allergy is a rising public health concern, and currently affects more than 1% of the US population. Compared with other food allergies, peanut allergy is less frequently outgrown and more often presents with severe symptoms (8). Reactions to allergens are mediated by activation of mast cells and basophils through the high-affinity IgE receptor, occurring when receptor-bound specific IgE is cross-linked by binding to peanut allergens. Production of this high affinity allergen-specific IgE is T cell-dependent, and a peanut-specific transcriptional profile characterized by increased expression of *IL4*, *IL5*, *IL9* and *IL13* has been observed in $CD4^+$ T cell subsets from subjects with peanut allergy (9-11). Multiple forms of immunotherapy for the treatment of peanut allergy are currently in clinical trials, of which oral immunotherapy is the most common. The effects of these treatments on the peanut-specific $CD4^+$ T cell response and their correlation with clinical success are an active area of investigation(12, 13).

In this work, we investigated the antigen-specific TCR repertoires of peanut-allergic individuals. We isolated peanut protein-activated and resting memory $CD4^+$ T cells and

sequenced the CDR3 of the TCR β-chain (CDR3β). The most enriched sequences in the peanut-activated compartment were identified as putatively peanut-specific CDR3βs (ps-CDR3s). These biologically and statistically enriched sequences exhibited properties associated with antigen-specific populations, such as increased homology, decreased diversity, and instances of convergent recombination. Within our pool of ps-CDR3s, we found a subset of sequences enriched in multiple subjects (i.e. public ps-CDR3s), suggesting the existence of public T cell responses among peanut-allergic individuals with diverse HLA genotypes. This work describes a potentially novel method to study antigen-specific TCR repertoires without restriction to known T cell epitopes or the availability of matched MHC reagents. This method is potentially applicable to studies of various T cell-mediated diseases beyond the scope of allergy.

Results

Selection of putatively peanut-specific CDR3β sequences

With the goal of enriching for antigen-specific memory $CD4^+$ T cells, peripheral blood mononuclear cells (PBMCs) from 27 peanut-allergic individuals were cultured for 20 hours with a peanut protein extract. Demographic and serological data for the subjects in this study can be found in Table 1. The activation marker CD154 (CD40L) was used to discriminate between peanut-activated ($CD154^+$) and resting ($CD154^-CD69^-$) memory T cells, and TCR CDR3β sequences were generated from both populations (Fig. 1A, Suppl. Fig. 1, Suppl. Table 1) (10, 14). Flow-cytometric analysis revealed that stimulation with peanut protein increased the frequency of CD154-expressing $CD4^+$ memory T cells as compared to unstimulated cultures (median[IQR] $CD154^+$ cells per million $CD4^+$ T cells: 2504[1328, 4105] stimulated vs. 172[117, 319] unstimulated; $P < 0.0001$) (Fig. 1B). To filter out sequences that were likely to be present in the $CD154^+$ population due to bystander activation, we developed a statistical enrichment strategy to define putatively peanut-specific CDR3β sequences (ps-CDR3s). First, a *G*-test of independence was applied to all CDR3β sequences in the peanut-activated compartment and adjusted P-values (q-values) were generated based on a FDR of $q < 0.05$. Sequences were further filtered to include only those with read counts ≥ 2 in the activated compartment, and those with a higher proportion in the activated compartment than in the resting compartment. This approach defined 7345 out of the 53205 unique amino acid sequences in the peanut-activated compartment (14%) across our 27 peanut-allergic individuals as ps-CDR3s, which met all three selection criteria. While the vast majority (n = 7309) of ps-CDR3s were private, 36 unique amino acid sequences (0.49% of ps-CDR3s) were significantly enriched in more than one individual (i.e., public). Of these public sequences, 33 were found in only two individuals.

However, one ps-CDR3 (CASSFRFLSGRALNEQFF) was statistically enriched in 8 of the 27 subjects, suggesting strong selective pressure for entrance of this sequence into the peanut-specific memory T cell repertoire (Suppl. Fig. 2). The CDR3β amino acid and nucleotide sequences that met our ps-CDR3 enrichment criteria only represented a minor fraction of the entire CDR3β pool from peanut-activated T cells (amino acid: median 14.1%, nucleotide: median 14.5%), indicating a high level of stringency with this statistical approach. The median frequency of ps-CDR3s was 117 per million $CD4^+$ T cells (IQR = 46.8, 173.15), which corresponds well to published frequencies of whole allergen-specific T cells in allergic individuals, suggesting the level of stringency is appropriate (Fig. 1C) (11, 15, 16).

We compared this strategy for TCR enrichment to the established approach of selecting antigen-specific TCRs by pMHC-tetramer binding. PBMCs from one of the 27 subjects were cultured with the major peanut allergen Ara h 1 for 20 hours to isolate $CD154^+$ and $CD154^-$ T cells by FACS, or for 2 weeks to expand the Ara h 1-specific T cell population and isolate Ara h 1 (AA 415-426) $tetramer^+$ and $tetramer^-$ T cells (Suppl. Fig. 3A). TCR CDR3β sequences were generated, and our method of statistical enrichment was applied to both populations, comparing $CD154^+$ to $CD154^-$ CDR3βs and $tetramer^+$ to $tetramer^-$ CDR3βs respectively, to filter out bystander-activated clones or those present due to non-specific binding of tetramer. We selected 458 ps-CDR3s from the $CD154^+$ population (accounting for 34.1% of total $CD154^+$ reads, and 17.9% of unique $CD154^+$ CDR3βs) and 175 ps-CDR3s from the $tetramer^+$ population (82.2% of total $tetramer^+$ reads, 17.3% of unique $tetramer^+$ CDR3βs), and detected 11 clones that were present in both populations (Suppl. Fig. 3B). Interestingly, these 11 ps-CDR3s were among the most expanded clones in the $tetramer^+$ population, indicating that our approach captures the most dominant peanut-specific TCRs (Suppl. Fig. 3C). Moreover, the relatively low frequency of

tetramer$^+$ clones present in the CD154$^+$ population (11 out of 458, or 2.4%) highlights the capacity of this strategy to identify ps-CDR3s beyond those specific for the single epitope used in the pMHC-tetramer.

ps-CDR3s are more similar and less diverse than unselected CDR3βs from both activated and resting T cells

To assess the efficacy of our statistical enrichment method for selecting those sequences most likely to be peanut-specific, we examined the similarity of the ps-CDR3s using Hamming and Levenshtein distances. For each ps-CDR3, the minimum Hamming and Levenshtein distances were measured by determining the minimum number of amino acid differences compared to its next closest ps-CDR3. As a control, the minimum Hamming and Levenshtein distances of equal-sized random repeated samplings of resting (CD154$^-$) and activated (CD154$^+$) CDR3βs were also measured (Fig. 2A, B). There were significantly more sequences with a minimum Hamming distance of 0-2 and Levenshtein distance of 0-1 in the ps-CDR3s than in either the activated or resting CDR3βs ($P < 0.01$), indicating enhanced similarity among the statistically enriched sequences. Importantly, these metrics showed no differences between unselected activated vs. resting CDR3βs, supporting our hypothesis that there is a substantial number of non-peanut-specific sequences in the CD154$^+$ population, and that our statistical enrichment strategy is effective in removing them.

We performed uniform resampling of autologous sets of ps-CDR3s, activated and resting CDR3β sequences to measure Hill's diversity index across different diversity orders to capture both the true richness (low q value) and abundance (high q value) of the samples (Fig. 2C) (17). For the 25 subjects with > 30 ps-CDR3s, the general diversity index was significantly lower for

the ps-CDR3s than for either the activated or resting sequences at diversity orders 0-2 ($P < 0.01$). Consistent with biological enrichment, the diversity of the activated sequences was significantly lower than that of the resting samples at diversity orders 0-2 for 22 of the 25 subjects ($P < 0.01$). Furthermore, the difference in diversity between the ps-CDR3s and either the activated or resting CDR3βs was significantly greater than the difference observed between the activated and resting CDR3βs ($P < 0.001$), emphasizing the utility in our statistical filtering (Fig. 2D).

Public ps-CDR3s are closer to germline and show evidence of convergent recombination

Antigen-specific public TCRs from individuals with different HLA genotypes can provide valuable insights into shared immune responses within a diverse population. For this reason, we wanted to further examine the characteristics of our public ps-CDR3s. We observed that the public ps-CDR3s contained fewer total N-nucleotide insertions than the private ps-CDR3s (median public: 3 insertions, median private: 7 insertions; $P < 0.01$). Similar patterns were observed when looking specifically at V-D insertions (median public: 1 insertion, median private: 3 insertions; $P < 0.01$) and D-J insertions (median public: 1 insertion, median private: 3 insertions; $P < 0.01$), which is consistent with prior literature demonstrating that public TCRs are closer to germline (Fig. 3A) (7, 18-20).

Convergent recombination of TCRs (i.e. multiple TCR nucleotide sequences translating to the same amino acid sequence) is a characteristic of antigenic selection (7). We searched the ps-CDR3s for this phenomenon and observed that convergence occurred in 467 out of 7345 unique ps-CDR3 amino acid sequences (6.4%). Interestingly, convergence was very common among the public ps-CDR3s, occurring in 33 out of 36 public amino acid sequences (91.7%), which suggests a strong selective pressure for these sequences' entrance into the peanut-specific

memory T cell pool. Additionally, the number of nucleotide sequences encoding each public ps-CDR3 was significantly higher than the number for each private ps-CDR3 (median public: 2 sequences, median private: 1 sequence; $P < 0.001$) (Fig. 3B). Analysis of the 3 most convergent public ps-CDR3s (≥ 5 unique nucleotide sequences) demonstrated convergence derived from germline (V-gene) and non-germline (N-nucleotide insertions) changes (Fig. 3C).

To obtain full TCR sequences of the public ps-CDR3s, PBMCs from 12 of the 27 peanut-allergic subjects from the initial analysis were stimulated with peanut protein extract for 20 hours and single-cell TCRα and TCRβ capture and sequencing was performed by Seq-Well on FACS-sorted peanut-activated $CD154^{+}$ memory T cells (21). Of the 36 public ps-CDR3 TCRβ sequences, 15 were found in at least 1 cell (41.7%), with 10 having matched TCRα and TCRβ sequences (27.8%) (Table 2).

Analysis of ps-CDR3 amino acid sequences reveals public, dominant motifs

Given the ps-CDR3 sequences demonstrated enhanced overall similarity, a motif analysis was performed according to Glanville et al. (2), to find dominant patterns among the putatively peanut-specific pool of sequences (Suppl. Fig. 4A). Briefly, for all ps-CDR3s, the region most likely to be in contact with an antigenic peptide was broken into continuous 3mer, 4mer and 5mer amino acid sequences. In addition, discontinuous 4mers and 5mers were generated to allow for the detection of motifs with gaps. Nmers that were at least 10-fold enriched in the ps-CDR3s as compared to the resting sequences and found in at least 3 unique ps-CDR3s from at least 3 individuals, were considered dominant motifs (Fig. 4A-B, Suppl. Fig. 4B-D). These criteria were met by 148 of the 104520 unique nmer sequences (0.14%), and the proportion of dominant motifs in the total motif pool was fairly similar for each nmer size (0.06 - 0.16%)

(Table 3). With these strict criteria, 399 of the 7345 ps-CDR3s (5.4%) contained at least one dominant motif. These motifs likely represent sequences specific for the most immunogenic T cell epitopes in peanut protein. Indeed, there were significantly more public ps-CDR3s (16.6%) than private ps-CDR3s (5.3%) that contained a dominant motif, supporting the hypothesis that these motifs are associated with the most common epitopes ($P < 0.05$) (Fig. 4C).

Network analysis confirms that public ps-CDR3s are core sequences

Utilizing the measured metrics of similarity, the dominant motifs and evidence of convergent recombination, we performed a network analysis to group highly homologous sequences (Fig. 5A). Similar to previously published approaches (2, 22, 23), each ps-CDR3 amino acid sequence was treated as a node and an edge was created between nodes if the ps-CDR3s were within a Levenshtein distance of 1 or shared a dominant motif. In addition, self-edges were created to represent additional nucleotide sequences for a corresponding amino acid sequence. Overall, there were 1759 edges created among 1324 ps-CDR3s. Interestingly, the majority (n = 23; 63.9%) of public ps-CDR3s were present in the network, whereas a much smaller fraction of private ps-CDR3s (n = 1301; 17.7%) were found in the graph ($P < 0.001$) (Fig. 5B). Moreover, the node degree (the number of edges per node) of the public sequences was higher than that of the private sequences (public median node degree: 4, private median node degree: 2; $P < 0.001$) (Fig. 5C). To assess the amount of structure observed among the ps-CDR3s, we compared the number of edges in the ps-CDR3 graph to the median number of edges created when performing the same network analysis on equal-sized random repeated samplings of activated or resting CDR3βs. Indeed, there were significantly more edges created between the ps-CDR3s than between all activated ($CD154^+$) CDR3βs (median 1234 edges) or resting

(CD154$^-$) CDR3βs (median 1091 edges) ($P < 0.001$) (Fig. 5D). We hypothesized that specific clusters of CDR3βs would be associated with individual HLA genotypes, which are shown in Suppl. Table 2, but no correlations could be found. Taken together, these data suggest that public ps-CDR3s represent "core" sequences that likely bind dominant T cell epitopes in peanut protein, which are recognized by multiple peanut-allergic individuals.

The frequency of ps-CDR3s is highest in the T_H17 subset, but the clonal expansion of ps-CDR3s is highest in the T_H2 subset

To better understand the phenotypes of the CD4$^+$ T cells from which ps-CDR3s were derived, TCRβ sequencing was performed on bulk T_H1 (CXCR3$^+$CCR5$^+$), T_H2 (CRTH2$^+$CCR4$^+$), T_H17 (CD161$^+$CCR6$^+$) and T_{FH} (CXCR5$^+$) cells from 8 of the 27 subjects used in the initial analysis (Suppl. Fig. 5, Suppl. Table 1). These bulk TCRβ sequences were then probed for ps-CDR3 amino acid sequences derived from the corresponding subjects, and the proportion of ps-CDR3 reads was observed to be highest in the T_H17 subset (Fig. 6A). However, ps-CDR3s detected in the T_H2 subset showed a higher level of clonal expansion than ps-CDR3s in the other subsets, indicating that the most dominant peanut-specific T cell clones have a T_H2 phenotype ($P < 0.01$) (Fig. 6B). Although ps-CDR3s were predominantly detected within a single T cell subset, we also observed shared sequences across the subsets, and these were more common among the public ps-CDR3s than private ps-CDR3s. In addition, shared sequences were most often found between the T_H1 and T_H17 phenotypes (Fig. 6C). These data suggest either shared epitope recognition by T cells from these subsets, or plasticity between T_H1 and T_H17 cell states, supporting recently published data on colonic T helper subsets in the gut (24).

Discussion

Selection of antigen-specific T cells based on their ability to bind an MHC-peptide complex (pMHC-multimer selection) is a widely used technique. This method, however, depends on a priori knowledge of T cell epitopes in subjects with specific HLA backgrounds, and is therefore limited in scope. Here, we expand upon a method not constrained by these limitations, utilizing a biological enrichment and augmenting it with a statistical enrichment to isolate likely antigen-specific CDR3βs that are not derived from bystander-activated T cells. This approach generated a library of ps-CDR3s that exhibited properties of an antigen-specific pool in terms of diversity, homology and convergence. Using these features in a network analysis to cluster similar sequences, we were able to show that public ps-CDR3s tended to have more neighbors (i.e. were more similar) than private ps-CDR3s, which corresponds with observations made by Madi et al. (23). When using the ps-CDR3s to probe the T_H-cell subset repertoires from a selection of the peanut-allergic individuals, we found that ps-CDR3s were most frequently present within the T_H17 subset, but those ps-CDR3s within the T_H2 subset were most clonally expanded. These data substantiate previous observations from our group, as we found that peanut-activated memory $CD4^+$ T cells from peanut-allergic patients highly express both T_H17- and T_H2-related genes. However, only T_H2-related gene expression and cytokine production distinguished patients with high clinical sensitivity to peanut from those with low sensitivity (10).

The use of CD154 as a marker to isolate antigen-activated T cells was first described in 2005, when Frentsch et al. demonstrated the utility of this method with an array of microbial and allergenic antigens (25) . Since then, numerous other groups have successfully applied this method with allergens and pathogen-derived proteins to better describe the phenotypes of

373 antigen-specific $CD4^+$ T cell populations (11, 14, 26-29). However, utilizing this methodology
374 to better characterize antigen-specific TCR repertoires is in its infancy. Glanville et al. single-
375 cell sorted and sequenced the TCRs of $CD154^+$ T cells activated with an *M. tuberculosis* lysate
376 and successfully validated antigen-specific motifs with an *in vitro* binding assay, demonstrating
377 that CD154 up-regulation is a valid method to study antigen-specific repertoires (2). Given the
378 concern of bystander activation of non-antigen-specific T cells, we chose to use the CD154 up-
379 regulation assay in conjunction with a statistical enrichment approach to define our library of ps-
380 CDR3s. This statistical enrichment has the distinct advantage of utilizing autologous resting
381 memory TCR repertoires as a reference when selecting the ps-CDR3s, therefore controlling for
382 the presence of clones in the $CD154^+$ compartment due to bystander activation of common
383 clones. The ps-CDR3s exhibited characteristics associated with antigen-specific TCR pools,
384 such as increased similarity and decreased diversity, when compared to both resting as well as
385 total activated CDR3βs. These findings emphasize the utility of applying the biological (CD154
386 up-regulation) and statistical (*G*-test of independence) enrichments in combination when
387 determining the most relevant TCRs.

388 The application of this method for defining antigen-specific ps-CDR3s has the potential
389 to answer a wide array of biological questions. As described here, antigen-specific ps-CDR3
390 libraries can be used to probe phenotypically defined subsets of $CD4^+$ T cells to better
391 understand the phenotype and physiology of antigen-specific T cells. Recently, our group has
392 successfully used a similar approach to identify differences in the peanut-specific $CD4^+$ T cell
393 repertoire and phenotypes between peanut-allergic individuals with high and low clinical
394 sensitivity (10). We found that high clinical sensitivity correlates with an expanded and more
395 diverse peanut-specific effector T cell population. Our strategy could also be extended to track

antigen-specific ps-CDR3s across a time-course or therapy. Examining the clones that expand or contract in response to an intervention could lead to insights regarding efficacy or susceptibility to side-effects.

A major advantage of our approach is that it is unbiased in terms of MHC restriction, representing the complete $CD4^+$ T cell response to one or multiple antigenic proteins, which is unlikely to be captured with pMHC-multimer-based sorting. In a previous study, Ryan et al. used a pMHC-dextramer approach to select Ara h 2-specific T cells from peanut-allergic individuals with a defined HLA background (*HLA-DRB1*1501, HLA-DRB4*) (12). While this method was successful in identifying phenotypic changes in Ara h 2-specific T cells over the course of peanut oral immunotherapy, the authors were only able to define 13 Ara h 2-specific TCRs, providing limited insights into the peanut-specific repertoire. Of note, we did not find any of the 13 reported Ara h 2-specific TCRs within our ps-CDR3s. In the present study, we observed that the most expanded CDR3β sequences captured by using an Ara h 1 (AA 415-426) pMHC-tetramer were also present among ps-CDR3s from Ara h 1-activated $CD154^+$ T cells of the same patient (Suppl. Fig. 3). These data indicate that our approach based on biological and statistical enrichment detects the most relevant CDR3βs specific for known T cell epitopes. Furthermore, the ps-CDR3s from Ara h 1-activated T cells included hundreds of additional antigen-enriched sequences, suggesting our epitope-agnostic approach is capturing more of the antigen-specific repertoire.

Using a method agnostic to both HLA genotype and known epitopes can help to better define the TCR landscape of a given immune response. For example, our study enabled us to describe the public nature of 36 unique antigen-specific CDR3βs. This would have been difficult to capture with methods based on pMHC-multimers, given the very low numbers of antigen-

specific T cells detected by these methods and the sparsity of these T cells in the overall T cell repertoire. The propensity of the public ps-CDR3s to demonstrate convergent recombination further exemplifies the utility of our methodology, as convergence suggests antigenic selection (7, 30, 31). Through the use of single-cell sequencing technologies, we were able to elucidate the matching TCRα and TCRβ chains for 10 of the public ps-CDR3s, demonstrating that these two methodologies in conjunction can help resolve the complete sequences that are most important for the development of diagnostics and/or therapeutics.

In sum, by using a combination of biological and statistical enrichment of CDR3βs from antigen-activated $CD4^+$ T cells, we were able to identify private and public CDR3βs that show evidence of antigenic selection. This method holds promise for application in a wide range of T cell-mediated disorders, as it effectively interrogates the antigen-specific TCR repertoire in unselected subjects, and potentially yields biomarkers for disease state and response to various treatments.

Methods

Subjects

The subjects included in this study participated in a peanut oral immunotherapy trial (NCT01750879) conducted at the Food Allergy Center at Massachusetts General Hospital. Criteria to be screened for the study included having a previous diagnosis of peanut allergy, a history of peanut-induced reactions consistent with immediate hypersensitivity, and peanut-specific serum IgE levels greater than 5 kU/L (ImmunoCAP; Thermo Fisher). Subjects who met these criteria underwent a double-blind placebo-controlled food challenge to confirm peanut allergy. Increasing peanut protein doses were administered every 20 minutes up to a maximum dose of 300 mg according to the following schedule: 3, 10, 30, 100, and 300 mg. All 27 individuals included in this study had an objective allergic reaction during the food challenge.

Cell culture and sorting of peanut-activated and resting memory $CD4^+$ T cells

Patient blood samples were collected at baseline (before the food challenge and the start of peanut oral immunotherapy), and PBMCs were isolated by means of density gradient centrifugation (Ficoll-Paque Plus; GE Healthcare). Fresh PBMCs were cultured in AIM V medium (Gibco) for 20 hours at a density of 5×10^6 in 1 mL medium per well in 24-well plates, and were left unstimulated or cultured with 100 µg/mL peanut protein extract (15×10^6 PBMCs per variable). The peanut extract was prepared by agitating defatted peanut flour (Golden Peanut and Tree Nuts) with PBS, centrifugation, and sterile-filtering. PE-conjugated anti-CD154 (clone TRAP1; BD Biosciences) was added to the cultures (20 µL/well) for the last 3 hours. After harvesting, the cells were labeled with AF700-conjugated anti-CD3 (clone UCHT1), APC-Cy7–conjugated anti-CD4 (RPA-T4), FITC–conjugated anti-CD45RA (HI100), PE-conjugated anti-

CD154 (all from BD Biosciences), AF647-conjugated anti-CD69 (FN50; BioLegend), and Live/Dead Fixable Violet stain (L34955; Thermo Fisher). Live $CD3^+CD4^+CD45RA^-$ activated $CD154^+$ and resting $CD154^-CD69^-$ T cells were sorted with a FACSAria II instrument (BD Biosciences). Sorted T cells were lysed in Buffer RLT Plus (Qiagen) with 1% β-mercaptoethanol (Millipore-Sigma), and stored at -80ºC, before total RNA and genomic DNA were isolated using the AllPrep DNA/RNA Micro Kit (Qiagen).

Expansion of Ara h 1-specific $CD4^+$ T cells and sorting of pMHC-Tetramer$^+$ T cells

Cryopreserved PBMCs from one of the 27 subjects (patient 107) were thawed and cultured in RPMI 1640 + 2 mM Glutamax (both from Gibco) + 10% human serum (Millipore-Sigma) + 100 U/ml penicillin and 100 µg/ml streptomycin (Thermo Fisher) (complete RPMI) with 50 µg/ml natural Ara h 1 (Indoor Biotechnologies), at a density of 6 x 10^6 cells in 1 ml medium per well in 24-well plates. Complete RPMI + 10 U/mL IL-2 (R&D Systems) was added after 5 days, and cells were cultured for a total of 14 days to expand Ara h 1-specific $CD4^+$ T cells. After harvesting, memory $CD4^+$ T cells were isolated with the EasySep human memory $CD4^+$ T cell enrichment kit (Stemcell Technologies), and labeled with APC-conjugated Ara h 1 (DRB1*03:01, AA 415-426) tetramer (32) (made in-house in James Moon's lab), at a concentration of 10 nM for 1 hour at room temperature. After washing off excess tetramer, the cells were labeled with BUV395-conjugated anti-CD3 (clone UCHT1; BD Biosciences), APC-Cy7-conjugated anti-CD4, FITC-conjugated anti-CD45RA, and Live/Dead Fixable Blue stain (L23105; Thermo Fisher), for 30 minutes at 4ºC. Live $CD3^+CD4^+CD45RA^-$ tetramer$^+$ and tetramer$^-$ T cells were sorted with a FACSAria Fusion instrument (BD Biosciences), and genomic DNA was isolated using the AllPrep DNA/RNA Micro Kit.

TCRβ sequencing

Genomic DNA was used to amplify and sequence the complementarity-determining region 3 (CDR3) regions (immunoSEQ assay; Adaptive Biotechnologies). The immunoSEQ approach generates an 87 basepair fragment capable of identifying the VDJ region spanning each unique CDR3β. Amplicons were sequenced using the Illumina NextSeq platform. Using a baseline developed from a suite of synthetic templates, primer concentrations and computational corrections were used to correct for the primer bias common to multiplex PCR reactions. Raw sequence data were filtered on the basis of TCRβ V, D, and J gene definitions provided by the IMGT database (www.imgt.org) and binned using a modified nearest-neighbor algorithm to merge closely related sequences and remove both PCR and sequencing errors.

Selection of ps-CDR3s

TCRβ sequencing data from activated ($CD154^+$) and resting ($CD154^-CD69^-$) memory $CD4^+$ T cell samples of all 27 peanut-allergic individuals were parsed using the *tcR* package (version 2.2.3) in R (33). We applied a statistical method to define putatively peanut-specific CDR3βs (ps-CDR3s), i.e. sequences that are likely to be antigen-specific and not detected due to bystander activation. This methodology's workflow starts with performing a G-test of independence on every unique CDR3β found in an individual's activated $CD154^+$ T cell sample, to determine if the proportion of a given sequence is higher in the activated CDR3βs than the resting CDR3βs. A G-test (also known as a likelihood ratio test) can be described by the formula:

$$G = 2\sum_i O_i * \ln\left(\frac{O_i}{E_i}\right)$$

where O_i is the observed read count of a given CDR3β in a particular population (activated or resting) and E_i is the expected count based on the proportion of this CDR3β in the entire memory T cell population (pooled activated and resting CDR3βs). With the G value for each CDR3β, the probability that the proportion of a CDR3β in the activated compartment is derived by chance from the proportion in the resting compartment can be calculated. All G-tests were calculated using the *GTest* function from the R package *Desctools* (version 0.99.34) (34). Given the sizable number of CDR3βs that were analyzed from each individual's data, a false discovery rate (FDR) correction was used to generate adjusted P-values (q-value) for each sequence, and those that met a cutoff of $q < 0.05$ were considered significant for this study. To enhance the stringency and reduce type 1 error, we also filtered out those CDR3βs with a read count of only 1 in the activated compartment (n = 24520, 46%) and those with a proportion in the activated compartment being less than that in the resting compartment (n = 4, 0.008%). All sequences selected after filtering were deemed putatively peanut-specific (ps-)CDR3s.

Minimum Hamming and Levenshtein distances

To determine global levels of similarity, the minimum Hamming distance (number of amino acid differences among CDR3βs of same length) and minimum Levenshtein distance (minimum number of insertions/deletions/substitutions between CDR3βs) of each ps-CDR3 against all other ps-CDR3s was determined using R (version 3.5.1) with the package *stringdist* (version 0.9.5.1). The percentage of ps-CDR3s at each minimum Hamming or Levenshtein distance was calculated. As a control, the minimum Hamming and Levenshtein distances of 100 equal-sized random resamplings of CDR3βs from the total activated and resting T cell pools

were determined. The median percentage of sequences at each minimum Hamming and Levenshtein distance for the 100 resamplings was calculated.

Diversity measurements

Diversity curves that measured Hill's diversity metric across diversity orders 0-4 were created using the R package *alakazam* (version 0.3.0) with the *alphaDiversity* function (17). Hill's diversity metric can be described by the following formula:

$$D = (\sum_{j=1}^{N} w_j \sum_{i=1}^{S} P_{i|j}^{q})^{\frac{1}{(1-q)}}$$

Where i represents a unique CDR3β, j is the sample size, $P_{i|j}$ is the proportional abundance of the i^{th} CDR3β in the j^{th} sample, W_j is the proportional abundance of the j^{th} sample relative to the entire dataset and q is a tuning parameter. q controls the influence of abundant over rare species on the metric (as q increases, the abundance of dominant sequences is more heavily weighted). Therefore, examining Hill's diversity metric across different values of q captures both the richness and abundance of species. The fold-change of Hill's diversity metric of all pairwise combinations of ps-CDR3s, total activated CDR3βs and resting CDR3βs was determined and the differences in these distributions were compared.

Paired single-cell TCRα/β sequencing

In parallel experiments, peanut protein-activated $CD154^+$ memory $CD4^+$ T cells from 12 of the 27 subjects were processed for single-cell RNA sequencing using the Seq-Well platform as previously described (35). A portion of each cDNA library was reserved for paired TCRα/β enrichment. Paired TCR sequencing was performed according to Tu et al (21). 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 (21). 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.

Motif analysis

CDR3β amino acid sequences were trimmed to the region most likely to be in contact with antigenic peptide (IMGT positions 106-117), as the stem positions of CDR3βs are not predicted to be involved with antigen binding (2). The trimmed sequences were then broken into all possible continuous nmers of size 3, 4, and 5 amino acids. In addition, discontinuous motifs were generated of size 4 and 5, allowing for gaps in the sequences so long as there were still 3 conserved residues. The proportion of each nmer in the ps-CDR3s was calculated by dividing the number of reads containing that nmer by the total number of ps-CDR3 reads. To determine fold-enrichment, this proportion of ps-CDR3 reads with a given nmer was divided by the proportion of reads in the resting CDR3βs with the nmer.

Network analysis

Each unique ps-CDR3 sequence was represented as a node and edges between nodes were made if there was a Levenshtein distance of 1 between them or if they shared an enriched motif. Levenshtein distances were determined using the R package *stringdist* (version 0.9.5.1). A self-edge was created on a node for every additional nucleotide sequence that corresponded to the same amino acid sequence. Network object (gml) files were created in R using the *igraph* package (version 1.0.1) and network visualization was performed with Cytoscape 3.7.0, using a force-directed open-CL layout. To evaluate the amount of structure present among ps-CDR3s, the number of edges in ps-CDR3s was compared with the median number of edges in 100 equal-sized random resamplings of resting and activated CDR3β sequences, by creating edges between the sequences of each resampling if they were within a Levenshtein distance of 1, had an enriched motif, or if sequences demonstrated convergence.

T cell subset sorts and probing for ps-CDR3 sequences

Cryopreserved PBMCs from 8 of the 27 subjects (Suppl. Table 1) were thawed, and memory $CD4^+$ T cells were isolated with the EasySep human memory $CD4^+$ T cell enrichment kit. Memory T cells were labeled with APC-Cy7-conjugated anti-CD4, PerCP-Cy5.5-conjugated anti-CRTH2 (clone BM16; BD Biosciences), BV605-conjugated anti-CCR4 (L291H4; BioLegend), AF647-conjugated anti-CXCR5 (J252D4; BioLegend), PE-Cy7-conjugated anti-PD1 (eBioJ105; eBioscience), FITC-conjugated anti-CXCR3 (G025H7; BioLegend), BUV395-conjugated anti-CCR5 (2D7/CCR5; BD Biosciences), eFluor450-conjugated anti-CD161 (HP-3G10; eBioscience), BV711-conjugated anti-CCR6 (11A9; BD Biosciences), and Live/Dead Fixable Blue stain. Live $CD4^+$ $CXCR3^+CCR5^+$ T_H1 cells, $CRTH2^+CCR4^+$ T_H2 cells,

601 $CD161^+CCR6^+$ T_H17 cells and $CXCR5^+$ T_{FH} cells were sorted with a FACSAria Fusion
602 instrument. T_H2 cells were gated out first, then T_{FH} cells, then T_H1 cells, and finally T_H17 cells.
603 Sorted T cell subsets then underwent the TCRβ sequencing protocol described above. These T
604 cell subset libraries were probed for the ps-CDR3 amino acid sequences from the corresponding
605 subjects. To determine the degree of overlap between the ps-CDR3s detected in each subset, a
606 Jaccard index was calculated between the unique ps-CDR3 sequences found in each pairwise
607 combination of subsets. The Jaccard index was calculated using the R package *OmicsMarkeR*
608 (version 1.16.0) and the resulting matrix was visualized with the R package *ComplexHeatmap*
609 (version 2.0.0) (36).
610
611 **Statistics**
612 A p-value less than 0.05 was considered significant for this study. In all boxplots shown
613 in this manuscript, the lower and upper hinges correspond to the first and third quartiles (the 25^{th}
614 and 75^{th} percentiles). The upper whisker extends from the hinge to the largest value no further
615 than 1.5 x IQR from the hinge (where IQR is the inter-quartile range, or distance between the
616 first and third quartiles). The lower whisker extends from the hinge to the smallest value at most
617 1.5 x IQR from the hinge. Data beyond the end of the whiskers ("outlying" points) are plotted
618 individually. All comparisons made with matched pairs of values from individual subjects were
619 tested with a Wilcoxon matched-pairs signed rank test. Such comparisons include frequencies of
620 $CD154^+$ T cells per million $CD4^+$ T cells in cultures and differences in the ratios of diversities in
621 resting, activated and ps-CDR3 pools. Evaluation of enrichment of ps-CDR3s at minimum
622 Hamming/Levenshtein distances utilized a 2-sided Fisher's exact test, comparing the median
623 frequencies of activated or resting CDR3βs to the observed frequency in the ps-CDR3s. To

assess differences in diversity, the empirical cumulative distribution function of bootstrap delta distribution was used as described by Stern et al. (37). Comparisons of public and private ps-CDR3 repertoire features, such as the number of nucleotide insertions, total rearrangements and node degree in the ps-CDR3 network were done with a Mann-Whitney U test. A 2-sided Fisher's exact test was used to compare the proportion of public ps-CDR3s with an enriched motif to private ps-CDR3s with a motif. Similarly, a 2-sided Fisher's exact test was used to compare the proportion of public ps-CDR3s with an edge in the ps-CDR3 network to that of the private ps-CDR3s. A z-test was used to evaluate the number of edges created within the ps-CDR3 network to that of the distribution of edges created by 100 equal-sized random samplings of resting or activated CDR3βs. Associations between $CD4^+$ T cell subset ps-CDR3s and clonal expansion were evaluated with logistic regression using the *glm* function in R. For each subset, ps-CDR3 membership was predicted with the sequences' read counts.

Data Availability

The TCRβ sequencing datasets used for this project have been deposited at the National Center for Biotechnology Information database of Genotypes and Phenotypes (dbGaP) under accession phs001897.v1.p1.

Code availability

The source code used in this study will be available at https://github.com/nealpsmith/peanut_pscdr3.

Study approval

All subjects were recruited with informed consent prior to sample collection, and the study was approved by the Institutional Review Board of Partners Healthcare (protocol no. 2012P002153). The study was conducted according to the principles of the Declaration of Helsinki.

Author contributions

NPS, BR, WGS and JCL designed the study. NPS and BR performed all of the lab work and generated all of the bulk TCRseq data. AAT and BM generated the single-cell TCRseq data. NPS, YVV and WGS performed the data analysis. NPS and BR wrote the manuscript. JJM developed the pMHC-tetramers used in this study. All authors reviewed and approved the final version of the manuscript.

Acknowledgements

We would like to thank our patients and their families who generously gave their time and participation, as well as Lauren Tracy, Colby Rondeau, Christine Elliot, and Leah Hayden, the clinical research coordinators who were instrumental in obtaining the samples used in the study. We would also like to thank the staff at the MGH Department of Pathology Flow and Image Cytometry Research Core who helped with the FACS. The Flow Core obtained funding from the National Institutes of Health Shared Instrumentation program (grant nos. 1S10OD012027-01A1, 1S10OD016372-01, 1S10RR020936-01, and 1S10RR023440-01A1). Additionally, we would like to thank Alexandra-Chloé Villani for her assistance with the TCRβ sequencing used in the study and Andre Han for technical assistance with the tetramer production. This work was supported by the NIH/NIAID (grant U19AI095261 to WGS, JJM and Andrew D. Luster, and K23AI130408 to YVV) and the Food Allergy Science Initiative (FASI).

References

1. Lythe G, Callard RE, Hoare RL, and Molina-Paris C. How many TCR clonotypes does a body maintain? *J Theor Biol.* 2016;389:214-24.
2. Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, et al. Identifying specificity groups in the T cell receptor repertoire. *Nature.* 2017;547(7661):94-8.
3. Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, Dolton G, et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* 2018;46(D1):D419-D27.
4. Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature.* 2017;547(7661):89-93.
5. Burrows SR, Silins SL, Moss DJ, Khanna R, Misko IS, and Argaet VP. T cell receptor repertoire for a viral epitope in humans is diversified by tolerance to a background major histocompatibility complex antigen. *J Exp Med.* 1995;182(6):1703-15.
6. Li H, Ye C, Ji G, and Han J. Determinants of public T cell responses. *Cell Res.* 2012;22(1):33-42.
7. Venturi V, Kedzierska K, Price DA, Doherty PC, Douek DC, Turner SJ, et al. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc Natl Acad Sci U S A.* 2006;103(49):18691-6.
8. Sicherer SH, and Sampson HA. Food allergy: A review and update on epidemiology, pathogenesis, diagnosis, prevention, and management. *J Allergy Clin Immunol.* 2018;141(1):41-58.
9. Gowthaman U, Chen JS, Zhang B, Flynn WF, Lu Y, Song W, et al. Identification of a T follicular helper cell subset that drives anaphylactic IgE. *Science.* 2019;365(6456).
10. Ruiter B, Smith NP, Monian B, Tu AA, Fleming E, Virkud YV, et al. Expansion of the CD4(+) effector T-cell repertoire characterizes peanut-allergic patients with heightened clinical sensitivity. *J Allergy Clin Immunol.* 2020;145(1):270-82.
11. Wambre E, Bajzik V, DeLong JH, O'Brien K, Nguyen QA, Speake C, et al. A phenotypically and functionally distinct human TH2 cell subpopulation is associated with allergic disorders. *Sci Transl Med.* 2017;9(401).
12. Ryan JF, Hovde R, Glanville J, Lyu SC, Ji X, Gupta S, et al. Successful immunotherapy induces previously unidentified allergen-specific CD4+ T-cell subsets. *Proc Natl Acad Sci U S A.* 2016;113(9):E1286-95.
13. Wang W, Lyu SC, Ji X, Gupta S, Manohar M, Dhondalay GKR, et al. Transcriptional changes in peanut-specific CD4+ T cells over the course of oral immunotherapy. *Clin Immunol.* 2020;219:108568.
14. Chattopadhyay PK, Yu J, and Roederer M. Live-cell assay to detect antigen-specific CD4+ T-cell responses by CD154 expression. *Nat Protoc.* 2006;1(1):1-6.
15. Renand A, Farrington M, Whalen E, Wambre E, Bajzik V, Chinthrajah S, et al. Heterogeneity of Ara h Component-Specific CD4 T Cell Responses in Peanut-Allergic Subjects. *Front Immunol.* 2018;9:1408.
16. Akdis M, Verhagen J, Taylor A, Karamloo F, Karagiannidis C, Crameri R, et al. Immune responses in healthy and allergic individuals are characterized by a fine balance between allergen-specific T regulatory 1 and T helper 2 cells. *J Exp Med.* 2004;199(11):1567-75.

17. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, and Kleinstein SH. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics.* 2015;31(20):3356-8.
18. Khosravi-Maharlooei M, Obradovic A, Misra A, Motwani K, Holzl M, Seay HR, et al. Crossreactive public TCR sequences undergo positive selection in the human thymic repertoire. *J Clin Invest.* 2019;129(6):2446-62.
19. Madi A, Shifrut E, Reich-Zeliger S, Gal H, Best K, Ndifon W, et al. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res.* 2014;24(10):1603-12.
20. Venturi V, Chin HY, Asher TE, Ladell K, Scheinberg P, Bornstein E, et al. TCR beta-chain sharing in human CD8+ T cell responses to cytomegalovirus and EBV. *J Immunol.* 2008;181(11):7853-62.
21. Tu AA, Gierahn TM, Monian B, Morgan DM, Mehta NK, Ruiter B, et al. TCR sequencing paired with massively parallel 3' RNA-seq reveals clonotypic T cell signatures. *Nat Immunol.* 2019;20(12):1692-9.
22. Ritvo PG, Saadawi A, Barennes P, Quiniou V, Chaara W, El Soufi K, et al. High-resolution repertoire analysis reveals a major bystander activation of Tfh and Tfr cells. *Proc Natl Acad Sci U S A.* 2018;115(38):9604-9.
23. Madi A, Poran A, Shifrut E, Reich-Zeliger S, Greenstein E, Zaretsky I, et al. T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. *Elife.* 2017;6.
24. James KR, Gomes T, Elmentaite R, Kumar N, Gulliver EL, King HW, et al. Distinct microbial and immune niches of the human colon. *Nat Immunol.* 2020;21(3):343-53.
25. Frentsch M, Arbach O, Kirchhoff D, Moewes B, Worm M, Rothe M, et al. Direct access to CD4+ T cells specific for defined antigens according to CD154 expression. *Nat Med.* 2005;11(10):1118-24.
26. Archila LD, Chow IT, McGinty JW, Renand A, Jeong D, Robinson D, et al. Ana o 1 and Ana o 2 cashew allergens share cross-reactive CD4(+) T cell epitopes with other tree nuts. *Clin Exp Allergy.* 2016;46(6):871-83.
27. Chiang D, Chen X, Jones SM, Wood RA, Sicherer SH, Burks AW, et al. Single-cell profiling of peanut-responsive T cells in patients with peanut allergy reveals heterogeneous effector TH2 subsets. *J Allergy Clin Immunol.* 2018;141(6):2107-20.
28. Stonier SW, Herbert AS, Kuehne AI, Sobarzo A, Habibulin P, Dahan CVA, et al. Marburg virus survivor immune responses are Th1 skewed with limited neutralizing antibody responses. *J Exp Med.* 2017;214(9):2563-72.
29. Weissler KA, Rasooly M, DiMaggio T, Bolan H, Cantave D, Martino D, 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.* 2018;141(5):1699-710 e7.
30. Joachims ML, Leehan KM, Lawrence C, Pelikan RC, Moore JS, Pan Z, et al. Single-cell analysis of glandular T cell receptors in Sjogren's syndrome. *JCI Insight.* 2016;1(8).
31. Ruggiero E, Nicolay JP, Fronza R, Arens A, Paruzynski A, Nowrouzi A, et al. High-resolution analysis of the human T-cell receptor repertoire. *Nat Commun.* 2015;6:8081.
32. DeLong JH, Simpson KH, Wambre E, James EA, Robinson D, and Kwok WW. Ara h 1-reactive T cells in individuals with peanut allergy. *J Allergy Clin Immunol.* 2011;127(5):1211-8 e3.

33. Nazarov VI, Pogorelyy MV, Komech EA, Zvyagin IV, Bolotin DA, Shugay M, et al. tcR: an R package for T cell receptor repertoire advanced data analysis. *BMC Bioinformatics.* 2015;16:175.
34. Andri Sea. 2021.
35. Gierahn TM, Wadsworth MH, 2nd, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat Methods.* 2017;14(4):395-8.
36. Gu Z, Eils R, and Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics.* 2016;32(18):2847-9.
37. Stern JN, Yaari G, Vander Heiden JA, Church G, Donahue WF, Hintzen RQ, et al. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci Transl Med.* 2014;6(248):248ra107.

Figures

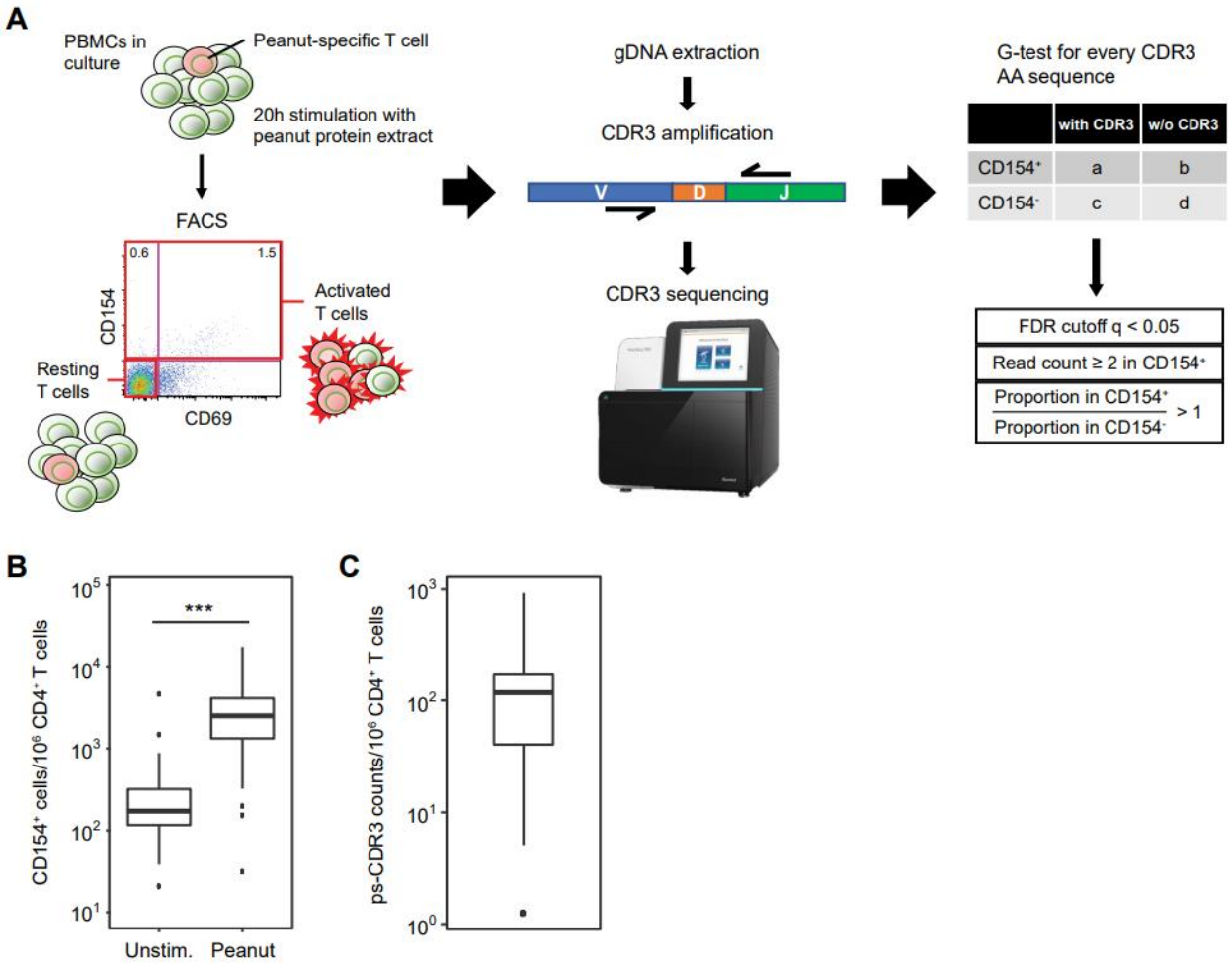


Figure 1 : Selection of putatively peanut-specific CDR3β sequences (ps-CDR3s). A, A general schema for the selection of ps-CDR3s (see methods). B, The frequency of $CD154^+$ T cells per million $CD4^+$ T cells in peanut protein-stimulated PBMC cultures is higher than in unstimulated cultures (n = 27, *** P < 0.001, Wilcoxon matched-pairs signed rank test). C, The frequency of ps-CDR3s per million $CD4^+$ T cells in peanut protein-stimulated cultures (n = 27).

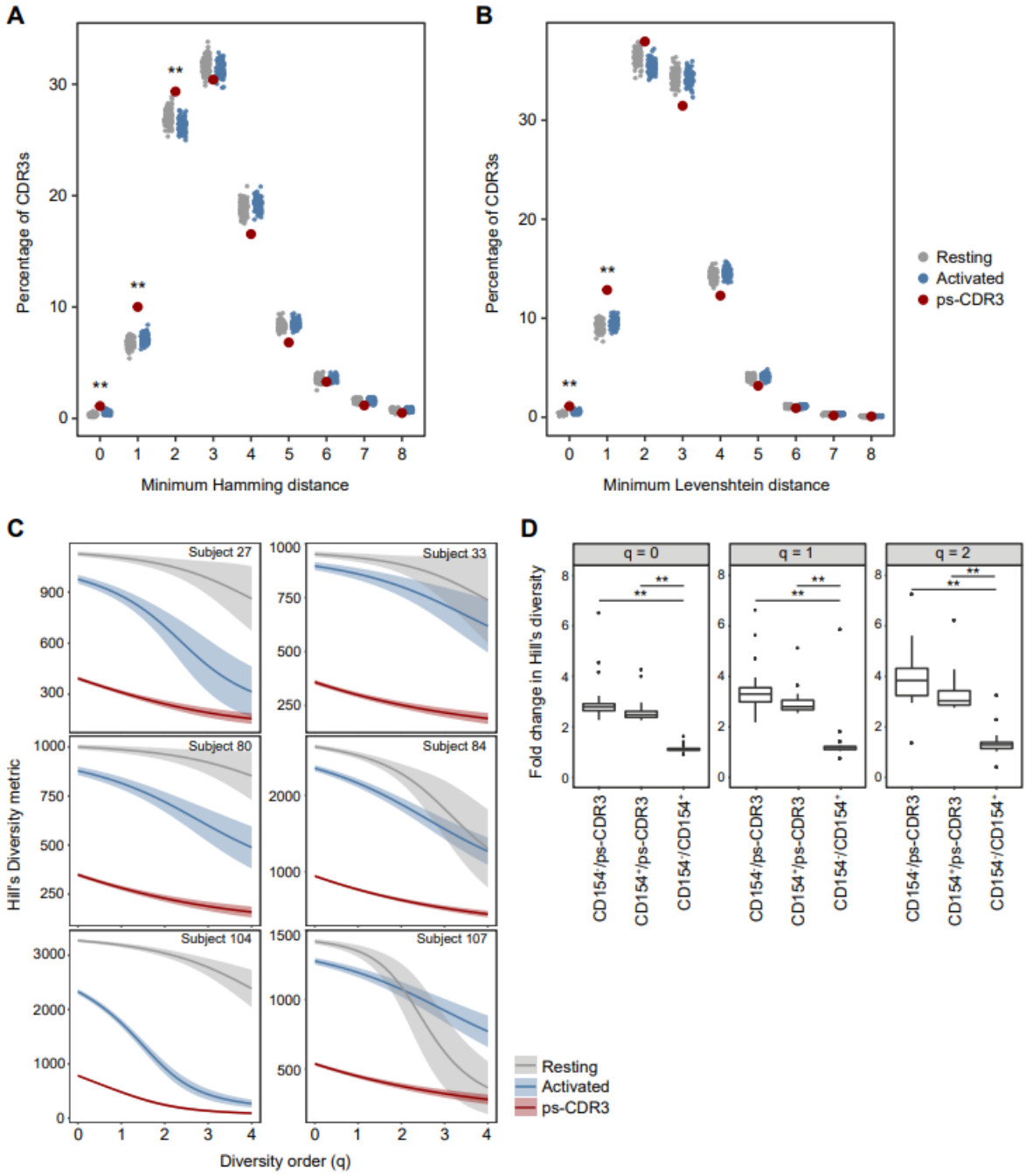
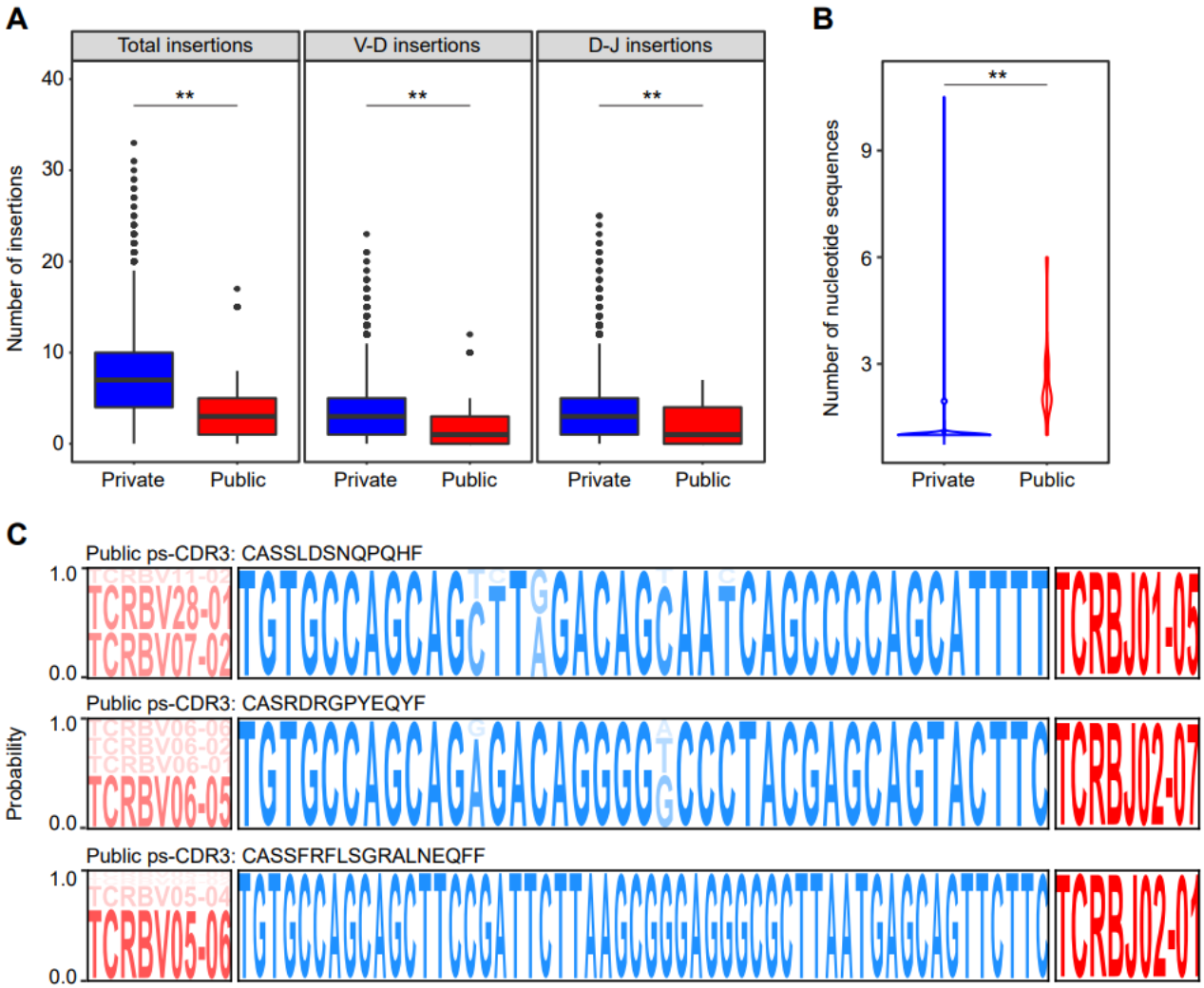


Figure 2: ps-CDR3s demonstrate increased similarity and decreased diversity when compared to total activated and resting CDR3βs. A-B, Minimum Hamming and Levenshtein distance of ps-CDR3s compared to 100 equal-sized random resamplings of total activated and resting CDR3βs (** P < 0.01, ps-CDR3 vs. resting and ps-CDR3 vs. activated, Fisher's exact test). C, Smoothed Hill's diversity curve at diversity orders 0-4 for 6 subjects. Diversity of ps-CDR3s was significantly lower than that of activated and resting CDR3βs at all diversity orders (P < 0.01, empirical cumulative distribution function of bootstrap delta distribution, see Methods). D, Fold-change of Hill's diversity metric of resting/ps-CDR3, activated/ps-CDR3, and resting/activated CDR3βs. Distributions represent values from 25 subjects with > 30 unique ps-CDR3s (** P < 0.01, Wilcoxon matched-pairs signed rank test).



881
882 **Figure 3: Public ps-CDR3s are characterized by being closer to germline and exhibiting**
883 **convergent recombination.** A, Public ps-CDR3s are closer to germline than private ps-CDR3s.
884 Distribution of total, V-D and D-J insertions of public and private ps-CDR3s (** $P < 0.01$,
885 Mann-Whitney U test). B, Public ps-CDR3s demonstrate more convergent recombination than
886 private ps-CDR3s. Violin plots represent the number of unique nucleotide sequences responsible
887 for each unique public and private ps-CDR3 (** $P < 0.01$, Mann-Whitney U test). C,
888 Convergence of public ps-CDR3s occurs in both germline (V-gene) and non-germline (CDR3β)
889 regions. Logo plots represent the probability of either specific gene segments (red) or CDR3β
890 nucleotides (blue) to be encoding for their corresponding amino acid sequence.

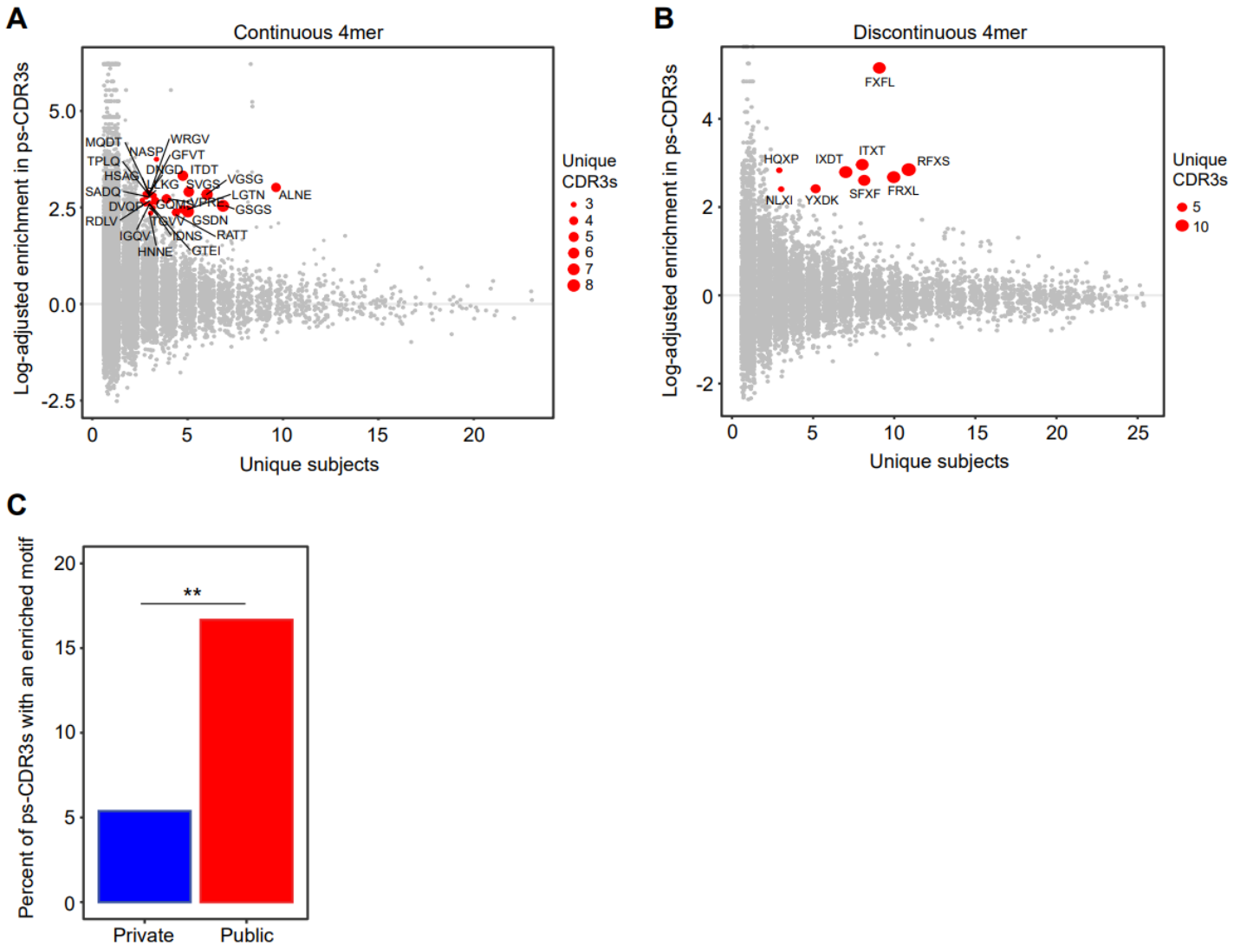


Figure 4: ps-CDR3s contain public, dominant motifs. A-B, The ps-CDR3s contain shared, enriched motifs as compared to resting CDR3βs. Red points represent the most dominant motifs, found to be ≥ 10-fold enriched in ps-CDR3s as compared to resting CDR3βs, and present in ≥ 3 unique ps-CDR3s derived from ≥ 3 subjects. C, Public ps-CDR3s are more likely to have a dominant motif than private ps-CDR3s (** $P < 0.01$, Fisher's exact test).

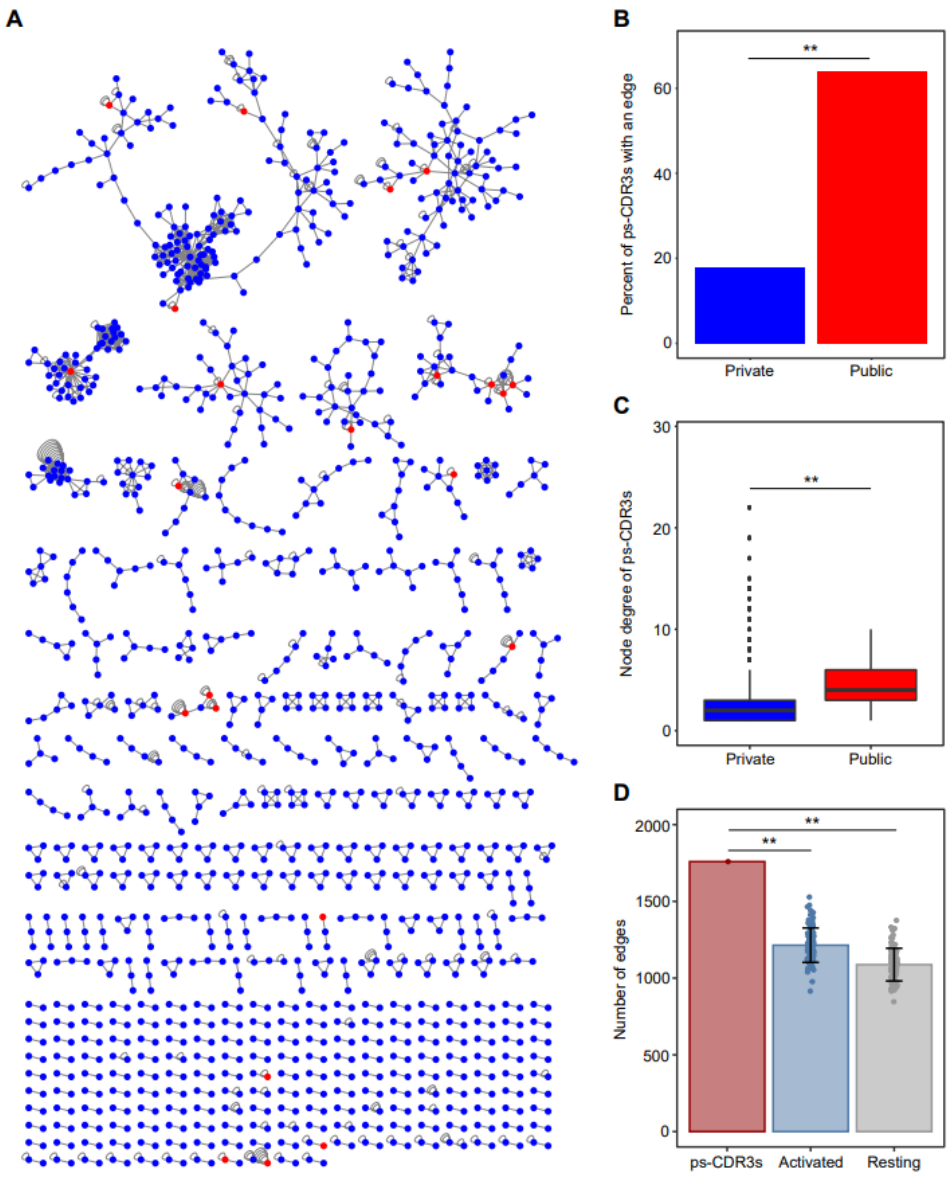


Figure 5: Network analysis reveals that public ps-CDR3s are core sequences with more structure than activated and resting CDR3βs. A, Network analysis of ps-CDR3s, where each node represents a unique ps-CDR3 amino acid sequence. Edges were created between the nodes if these were within a Levenshtein distance of 1 or if the nodes shared an enriched motif. Self-edges were created on nodes to represent additional unique nucleotide sequences encoding the same amino acid sequence (i.e. convergence). Blue nodes represent private ps-CDR3s and red nodes represent public ps-CDR3s. B, The percent of public ps-CDR3s with an edge to another ps-CDR3 (as shown in panel A) was higher than that of private ps-CDR3s (** $P < 0.01$, Fisher's exact test). C, The overall node degree (number of edges per node) of public ps-CDR3s was higher than that of private ps-CDR3s. Boxplots represent distribution of the node degree of all private (blue) and public (red) ps-CDR3s in the graph (** $P < 0.01$, Mann-Whitney U test). D, The number of total edges created among the ps-CDR3s compared to the median number of edges created among 100 equal-sized random resamplings of total activated and resting CDR3βs. Error bars represent standard deviation (** $P < 0.01$, z-test).

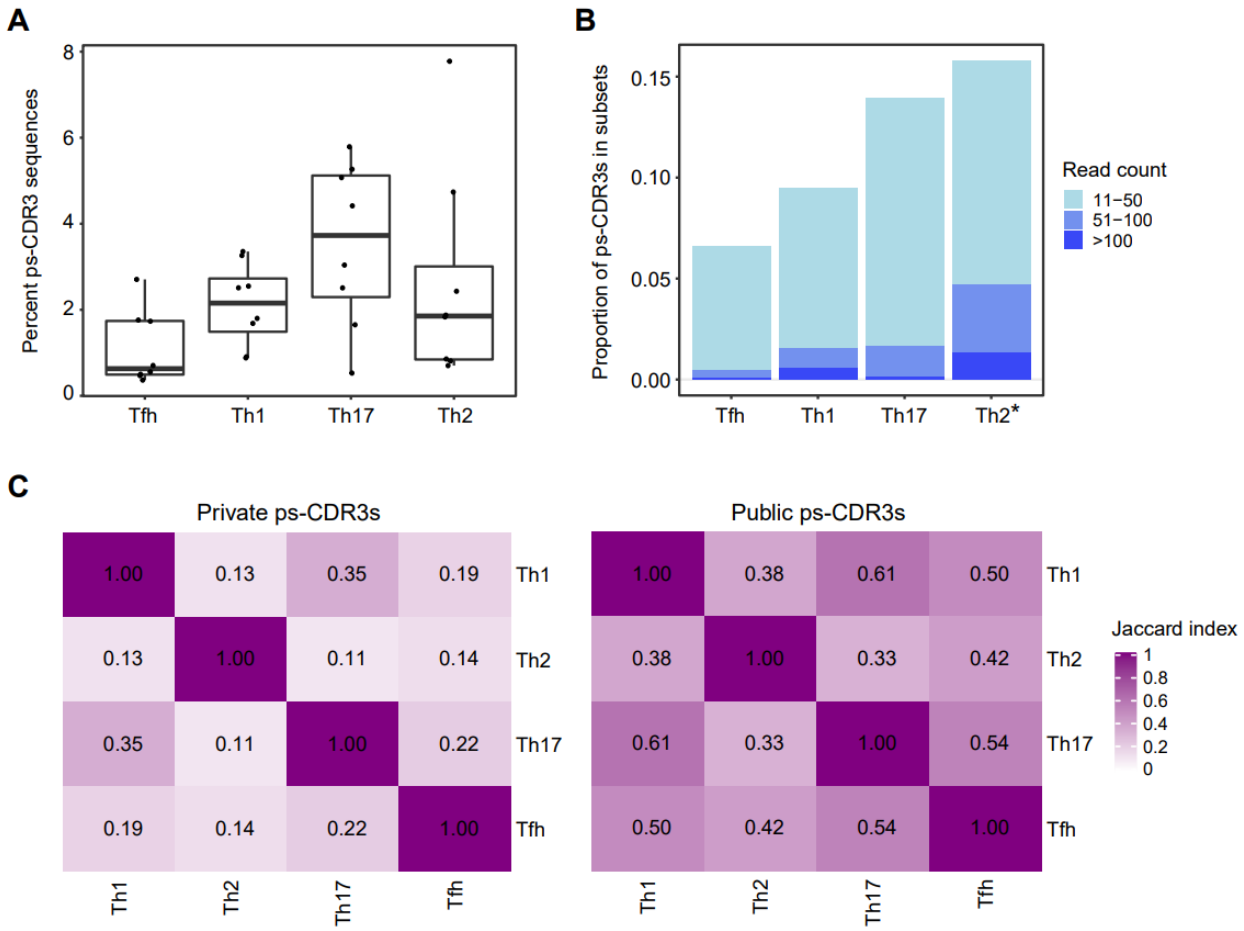
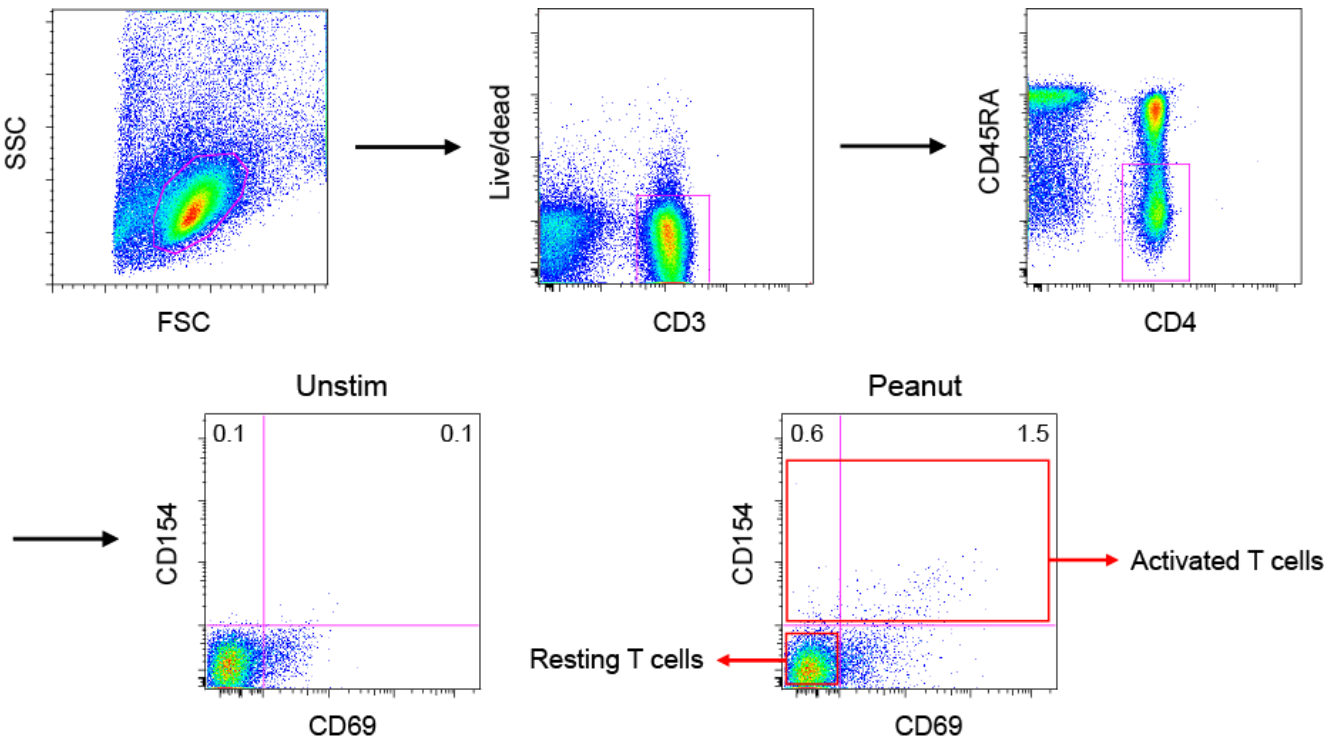
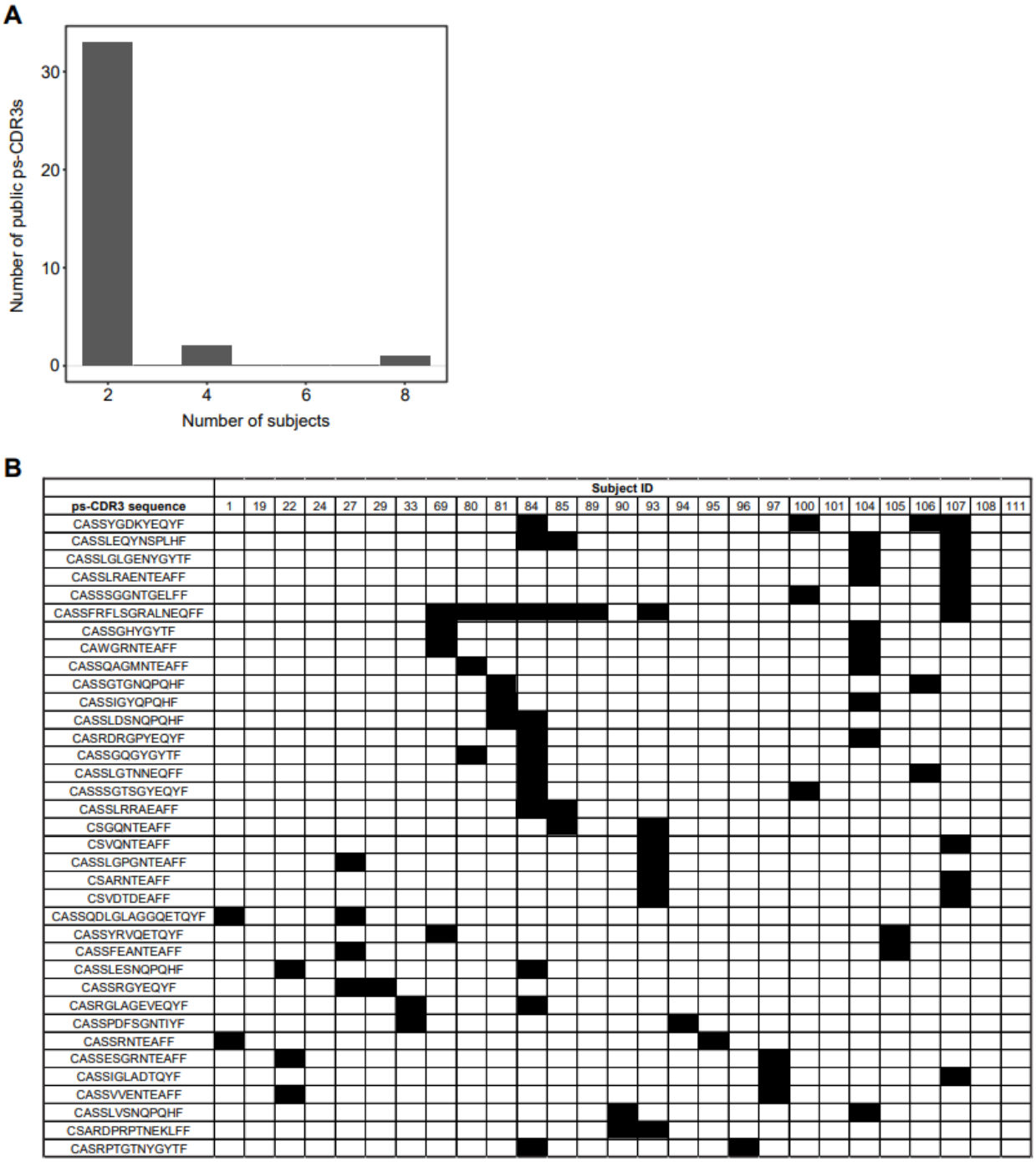


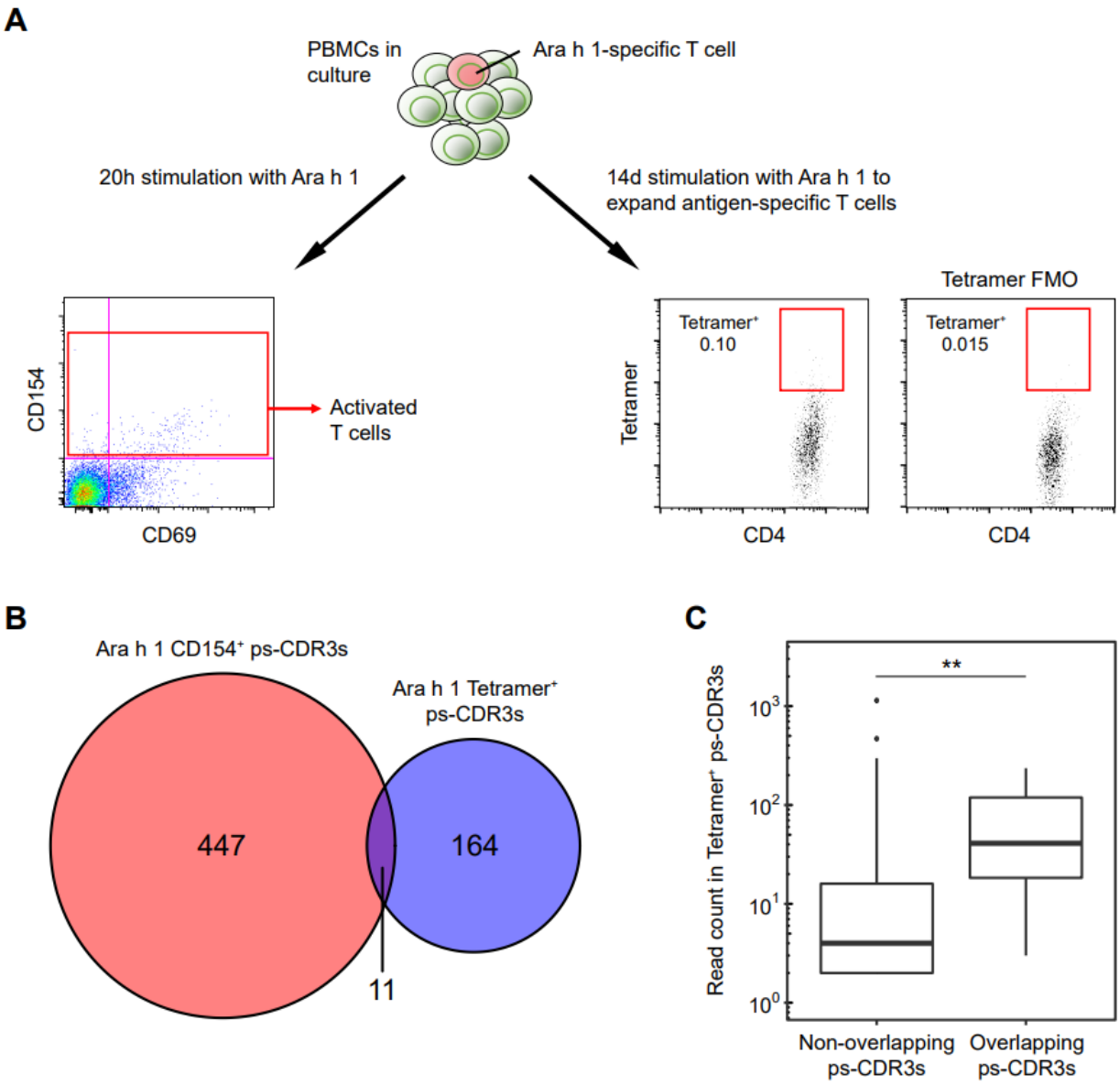
Figure 6: The frequency of ps-CDR3s is highest in the T_H17 subset, but the clonal expansion of ps-CDR3s is highest in the T_H2 subset. A, The percent of total read counts corresponding to a ps-CDR3 sequence was highest in the T_H17 subset (n = 8). B, T_H2-derived ps-CDR3s were more clonally expanded than ps-CDR3s in the other subsets. Bar represents the proportion of ps-CDR3s in a given subset based on their read count (* $P < 0.01$, logistic regression). C, Jaccard index showing the degree of overlap in public and private ps-CDR3s between the T cell subsets.



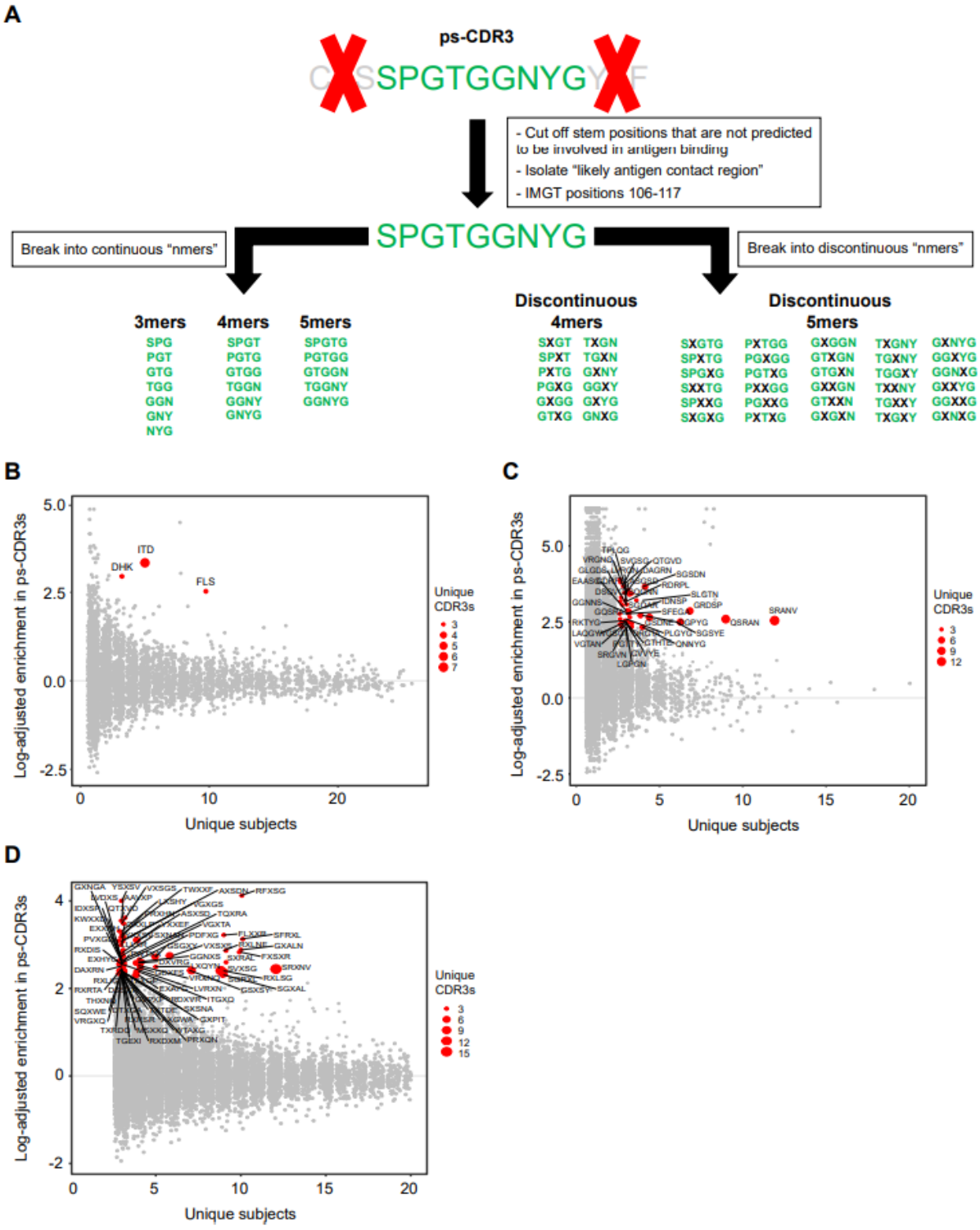
Supplemental figure 1. FACS gating strategy for isolating peanut-activated memory $CD4^+$ T cells ($CD154^+$) and resting memory $CD4^+$ T cells ($CD154^-CD69^-$) from the PBMCs of peanut-allergic individuals.



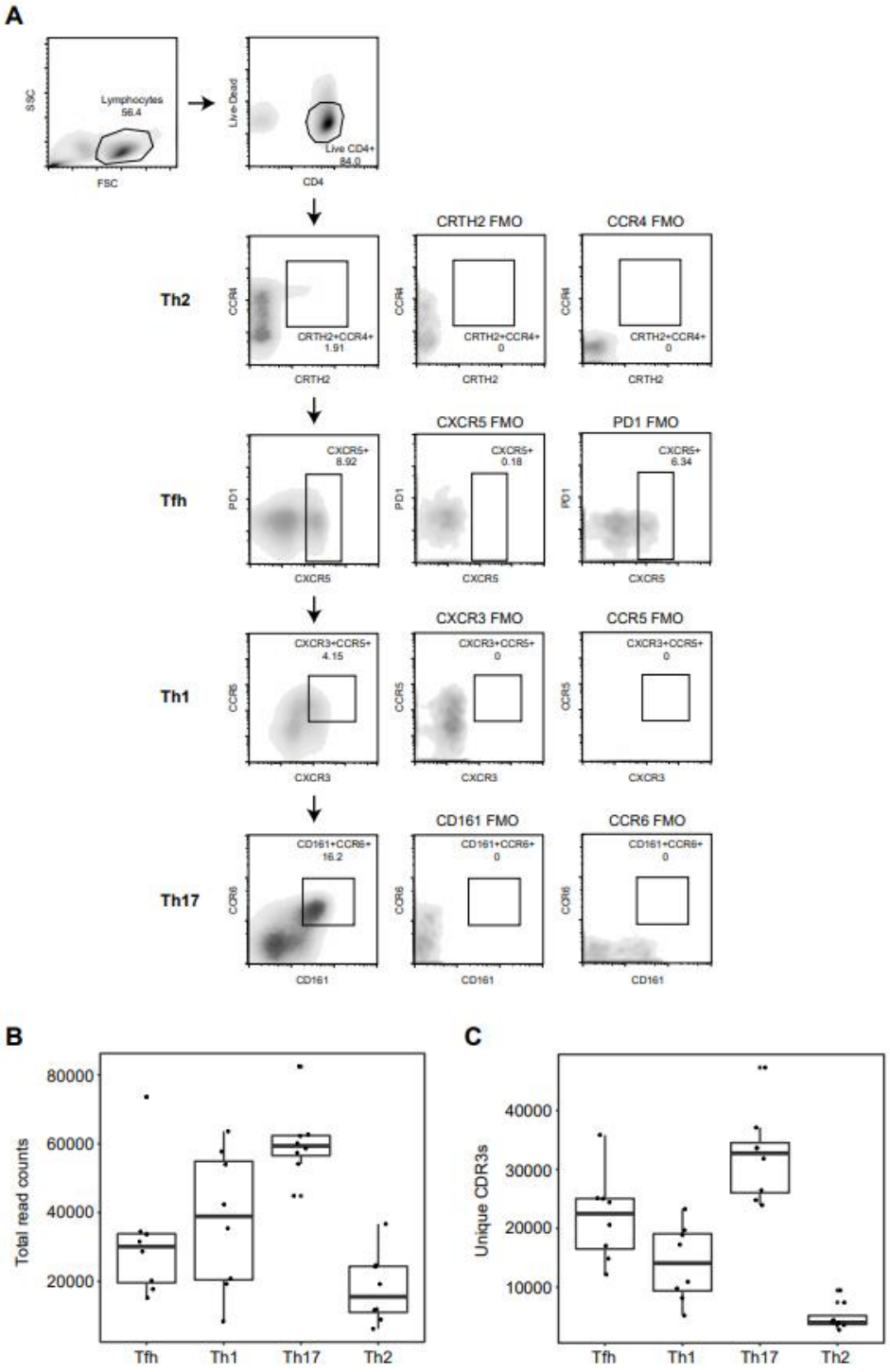
Supplemental Figure 2: Distribution of public ps-CDR3s across the peanut-allergic subjects. A, Distribution of the public ps-CDR3s over the subjects. Shown is the number of public ps-CDR3s present in a given number of individuals. B, Summary of the presence of each unique public ps-CDR3 in each subject.



Supplemental Figure 3: ps-CDR3s from Ara h 1 tetramer$^+$ T cells are present among ps-CDR3s derived from Ara h 1-activated T cells, and are highly expanded. A, General schema of the experimental approach. PBMCs from subject 107 were stimulated with the major peanut allergen Ara h 1 (50 µg/ml) for 20 hours, and ps-CDR3s were derived from CD154$^+$ T cells. In addition, PBMCs from the same subject were cultured for 2 weeks with Ara h 1, and ps-CDR3s were derived from sorted Ara h 1 (DRB1*03:01, AA 415-426) tetramer$^+$ T cells (see Methods). FMO = Fluorescence-minus-one control. B, Eleven ps-CDR3s from Ara h 1 tetramer$^+$ T cells were also present in the ps-CDR3 pool from Ara h 1-activated CD154$^+$ T cells. C, The 11 overlapping ps-CDR3s were more clonally expanded than the other ps-CDR3s derived from Ara h 1 tetramer$^+$ T cells. Boxplots represent distributions of read counts of the corresponding ps-CDR3 sequences (** $P < 0.01$, Mann-Whitney U test).



Supplemental Figure 4: Motif analysis of ps-CDR3s. A, Schema of nmer generation with an example ps-CDR3 (see Methods). B-D, The ps-CDR3s contain shared and enriched 3-mer, 5-mer, and discontinuous 5-mer motifs as compared to resting CDR3s. Red points represent the most dominant motifs, which were ≥ 10-fold enriched, in ≥ 3 unique ps-CDR3s derived from ≥ 3 subjects.



Supplemental Figure 5: Sorting of T cell subsets from peanut-allergic individuals. A, FACS gating strategy for isolating bulk memory CD4$^+$ T_H2 (CRTH2$^+$CCR4$^+$), T_{FH} (CXCR5$^+$), T_H1 (CXCR3$^+$CCR5$^+$), and T_H17 (CD161$^+$CCR6$^+$) cells (see Methods). FMO = Fluorescence-minus-one control. B, Distribution of the total read counts in the T cell subset samples. C, Distribution of unique CDR3β sequences in the T cell subset samples.

Table 1 : Demographics and serological data of the subjects included in the experiments described in this paper

Subject ID	Age (yr)	Gender	Race	Peanut-specific IgE (kU/L)
1	16	Male	Caucasian	66
19	16	Male	Caucasian	65
22	19	Female	Caucasian	148
24	36	Male	Caucasian	25
27	28	Male	Caucasian	7
29	28	Male	Caucasian	22
33	16	Female	Caucasian	84
69	15	Male	Caucasian	11
80	8	Female	African American	320
81	11	Female	Caucasian	246
84	22	Male	Caucasian	61
85	9	Male	Caucasian	151
89	8	Female	Caucasian	88
90	9	Male	Caucasian	159
93	11	Male	Caucasian	41
94	8	Male	Hispanic	74
95	8	Male	Caucasian	451
96	10	Female	Caucasian	39
97	36	Female	Asian	5
100	22	Female	Caucasian	396
101	8	Female	Caucasian	75
104	8	Male	Caucasian	432
105	22	Female	Caucasian	44
106	32	Female	Caucasian	5
107	22	Female	Caucasian	27
108	8	Male	Caucasian	650
111	22	Female	Caucasian	21

Table 2 : Full paired TCRα and TCRβ sequences of 10 public ps-CDR3s as determined by single-cell TCR sequencing.

TCRα		TCRβ	
V gene	**CDR3**	**V gene**	**CDR3**
TRAV41	CAVRGSNYKLTF	TRBV14	CASSQAGMNTEAFF
TRAV5	CAETFTGGGNKLTF	TRBV7-2	CASSLDSNQPQHF
TRAV8-6	CAVSDGGSARQLTF	TRBV12-4	CASSLRRAEAFF
TRAV12-3	CAMSYSDGQKLLFA	TRBV28	CASRGLAGEVEQYF
TRAV41	CAAGNKLTF	TRBV6-5	CASRDRGPYEQYF
TRAV20	CAVQAGGLGKLSF	TRBV7-2	CASSLGTNNEQFF
TRAV35	CAGPQGGSEKLVF	TRBV29-1	CSARNTEAFF
TRAV21	CAVNTGNQFYF	TRBV20-1	CSARDPRPTNEKLFF
TRAV22	CAVERSNFGNEKLTF	TRBV7-9	CASSRGYEQYF
TRAV29	CAASGRSFPSGTYKYIF	TRBV29-1	CSVQNTEAFF

Table 3: Summary of all nmers generated from the ps-CDR3s.

Nmer	3mer	4mer	5mer	Disc. 4mer	Disc. 5mer	Total
Unique	4376	15559	23473	8279	52833	104520
Enriched	3	26	38	9	72	148
% enriched	0.06	0.16	0.16	0.11	0.14	0.14

Supplemental Table 1: Subjects included in each set of experiments described in this paper.

Subject ID	TCRβ-seq CD154+/CD154-	T cell subset sorting	scTCR-seq CD154+	pMHC-Tetramer TCRβ-seq
1	X			
19	X			
22	X	X		
24	X			
27	X	X		
29	X	X		
33	X	X	X	
69	X		X	
80	X			
81	X	X		
84	X		X	
85	X			
89	X			
90	X		X	
93	X		X	
94	X			
95	X		X	
96	X		X	
97	X	X	X	
100	X			
101	X			
104	X			
105	X		X	
106	X	X	X	
107	X	X	X	X
108	X			
111	X		X	

Supplemental Table 2: HLA genotypes of the subjects in this study.

Subject ID	DPA1_allele1	DPA1_allele2	DPB1_allele1	DPB1_allele2	DQA1_allele1	DQA1_allele2	DQB1_allele1	DQB1_allele2	DRB1_allele1	DRB1_allele2	DRB3_allele1	DRB3_allele2	DRB4	DRB5
1	1:03:01	2:02:02	1:01:01	3:01:01	1:03:01	3:03:01 AM	3:01:01	6:03:01	13:01:01	4:01:01	1:01:02		1:03:01	
19	2:01:01	2:07:01	11:01:01	19:01:01	1:03:01	2:01:01	2:02:01	6:03:01	13:01:01	7:01:01	2:02:01		1:01:01	
22	1:03:01	1:03:01	4:01:01	4:01:01	1:02:01	3:01:01	3:02:01	6:02:01	15:01:01	4:02:01			1:03:01	1:01:01
24	2:01:01	2:01:02	1:01:01	17:01:01	2:01:01	5:01:01	2:01:01	2:02:01	3:01:01	7:01:01	1:01:02		1:01:01	
27	1:03:01	2:01:01	3:01:01	10:01:01	1:04:01	3:01:01	3:02:01	5:03:01	14:54:01	4:01:01	2:02:01		1:03:01	
29	1:03:01	2:01:01	4:01:01	13:FNVU	3:01:01	5:05:01	3:01:01	3:02:01	11:04:01	4:02:01	2:02:01		1:03:01	
33	1:03:01	2:01:07	4:01:01	13:FNVU	1:02:01	3:02:01	3:03:02	6:09:01	13:02:01	09:CWA	3:01:01		1:03:02	
69	1:03:01	1:03:01	2:01:02	13:FNVU	1:02:01	2:01:01	2:02:01	6:02:01	15:01:01	7:01:01			1:01:01	1:01:01
80	1:03:01	2	4:01:01	85:01:01	2:01:01	5:01:01	2:01:01	2:02:01	3:01:01	7:01:01	1:01:01		1:03:01	
81	1:03:01	1:03:01	2:01:01	4:01:01	1:02:01	3:01:01	3:02:01	6:02:01	15:01:01	4:02:01			1:03:01	1:01:01
84	1:03:01	2:01:01	4:01:01	10:01:01	1:01:02	1:02:01	5:01:01	6:09:01	1:02:01	13:02:01	3:01:01	3:01:01		
85	1:03:01	1:03:01	4:01:01	4:02:01	5:01:01	5:01:01	2:01:01	2:01:01	3:01:01	3:01:01	1:01:02	1:01:02		
89	1:03:01	2:06	4:02:01	5:01:01	1:02:01	1:02:01	6:02:01	6:02:01	15:01:01	15:01:01				1:01:01
90	1:03:01	1:03:01	4:01:01	4:01:01	5:01:01	5:05:01	2:01:01	3:01:01	3:01	11:01	1:01:02	2:02:01		
93			4:01:01	4:01:01	1:04:01	5:01:01	2:01:01	5:03:01	3:01:01	14:01:01	1:01:02	2:24		
94	1:03:01	1:03:01	2:01:02	4:01:01	1:01:01	1:02:01	5:01:01	6:02:01	1:03:01	15:01:01				1:01:01
95	1:03:01	1:03:01	4:01:01	4:01:01	1:02:01	5:05:01	3:01:01	6:02:01	15:01:01	11:01:01	2:02:01			1:01:01
96	1:03:01	2:01:02	1:01:01	2:01:02	3:03:01	5:05:01	3:01:01	3:02:01	12:JV	4:01:01	2:02:01		1:03:01	
97	1:03:01	1:03:01	2:01:02	48:01:00	3:01:01	5:03:01	3:01:01	3:02:01	13:12:01	4:03:01	2:02:01		1:03:01	
100	1:03:01	1:03:01	2:01:02	4:01:01	5:05:01	5:05:01	3:01:01	3:01:01	11:04:01	11:04:01	2:02:01	2:02:01		
101	1:03:01	1:03:01	3:01:01	4:01:01	1:02:01	1:02:01	6:02:01	6:04:01	15:01:01	13:02:01	3:01:01			1:01:01
104	1:03:01	1:03:01	4:01:01	4:01:01	1:02:01	2:01:01	3:03:02	6:02:01	15:01:01	7:01:01			01:03:01 N	1:01:01
105	1:03:01	2:01:01	2:01:02	5:01:01	3:01:01	5:01:01	2:01:01	3:02:01	3:01:01	4:02:01	1:01:02		1:03:01	
106	1:03:01	1:03:01	4:01:01	104:01:01	1:01:02	2:01:01	3:01:01	5:01:01	1:02:01	7:01:01			1:01:01	
107	1:03:01	1:03:01	4:01:01	20:01:01	3:01:01	5:01:01	2:01:01	3:02:01	3:01:01	4:04:01	1:01:02		1:03:01	
108	1:03:01	1:03:01	3:01:01	4:02:01	1:02:01	5:01:01	2:01:01	6:04:01	3:01:01	13:02:01	1:01:02	3:01:01		
111	1:03:01	1:03:01	2:01:02	4:01:01	1:03:01	5:05:01	3:01:01	6:03:01	13:01:01	13:05:01	1:01:02	2:02:01		