

Cross-reactive public TCR sequences undergo positive selection in the human thymic repertoire

Mohsen Khosravi-Maharlooei,¹ Aleksandar Obradovic,¹ Aditya Misra,¹ Keshav Motwani,² Markus Holzl,¹ Howard R. Seay,² Susan DeWolf,¹ Grace Nauman,^{1,3} Nichole Danzl,¹ Haowei Li,¹ Siu-hong Ho,¹ Robert Winchester,⁴ Yufeng Shen,⁵ Todd M. Brusko,² and Megan Sykes^{1,3,6}

¹Columbia Center for Translational Immunology, Department of Medicine, Columbia University Medical Center, Columbia University, New York, New York, USA. ²Department of Pathology, Immunology and Laboratory Medicine, University of Florida, Gainesville, Florida, USA. ³Department of Microbiology and Immunology, Columbia University Medical Center, Columbia University, New York, New York, USA. ⁴Division of Rheumatology, Department of Medicine, ⁵Center for Computational Biology and Bioinformatics, and ⁶Department of Surgery, Columbia University Medical Center, Columbia University, New York, New York, USA.

We studied human T cell repertoire formation using high-throughput T cell receptor β (TCR β) complementarity-determining region 3 (CDR3) sequencing in immunodeficient mice receiving human hematopoietic stem cells (HSCs) and human thymus grafts. Replicate humanized mice generated diverse and highly divergent repertoires. We observed repertoire narrowing and increased CDR3 β sharing during thymocyte selection. Whereas hydrophobicity analysis implicated self-peptides in positive selection of the overall repertoire, positive selection favored shorter shared sequences that had reduced hydrophobicity at positions 6 and 7 of CDR3 β s, suggesting weaker interactions with self-peptides than were observed with unshared sequences, possibly allowing escape from negative selection. Sharing was similar between autologous and allogeneic thymi and occurred between different cell subsets. Shared sequences were enriched for allo-cross-reactive CDR3 β s and for type 1 diabetes-associated autoreactive CDR3 β s. Single-cell TCR sequencing showed increased sharing of CDR3 α s compared with CDR3 β s between mice. Our data collectively implicate preferential positive selection for shared human CDR3 β s that are highly cross-reactive. Although previous studies suggested a role for recombination bias in producing “public” sequences in mice, our study is the first to our knowledge to demonstrate a role for thymic selection. Our results implicate positive selection for promiscuous TCR β sequences that probably evade negative selection, given their low affinity for self-ligands, in the abundance of “public” human TCR β sequences.

Introduction

A functional TCR repertoire requires great diversity to recognize a wide range of pathogens. Extrapolations of early sequencing results on a few hundred T cell receptors (TCRs) led to an estimate of approximately 10^6 different TCR β chains in human blood, each pairing, on average, with at least 25 different α chains (1). Recent studies have used next-generation TCR sequencing to capture the diversity of TCR repertoires. The first such study estimated that normal human blood contains 3×10^6 to 4×10^6 unique TCR β chains (2), while a later study provided a minimal estimate of 100×10^6 unique TCR β sequences in naive CD4⁺ and CD8⁺ T cell repertoires of young adults (3). Another study estimated that there are 40×10^6 to 70×10^6 unique TCR β sequences and 60×10^6 to 100×10^6 TCR α sequences in human pediatric thymi (4). However, a surpris-

ing degree of repertoire overlap has been observed between individuals in studies of peripheral T cells (5). In silico modeling suggested a role for recombination bias in generating shared sequences in mice (6). In a human system, we have investigated the role of thymic selection in the generation of shared sequences.

Diversity of the TCR repertoire is formed at different levels. First, different V β , D β , and J β genes recombine to generate TCR β chains. Similarly, different V α and J α genes recombine to generate TCR α chains (7). Exonuclease removal of exposed residues and addition of random nucleotides at the junctional sites mediated by the enzyme terminal deoxynucleotidyl transferase (TdT) further diversifies the TCR repertoire (8). Finally, α and β chains combine to form functional TCRs. Studies in mice have shown that thymocytes undergo different selection processes that shape the TCR repertoire after the formation of functional TCRs (9, 10). This repertoire is further altered in the periphery by the expansion and deletion of certain clones (11). Our knowledge of the formation of the human TCR repertoire is limited by access to human thymus samples and the inability to manipulate variables in vivo. Humanized mice generated by transplanting immunodeficient mice with human thymus and hematopoietic stem cells (HSCs) provide a unique opportunity to investigate the factors involved in the formation of a human

► Related Commentary: p. 2195

Authorship note: MKM and AO contributed equally to this work.

Conflict of interest: The authors have declared that no conflict of interest exists.

Copyright: © 2019, American Society for Clinical Investigation.

Submitted: August 20, 2018; **Accepted:** March 21, 2019; **Published:** May 13, 2019.

Reference information: *J Clin Invest.* 2019;129(6):2446–2462.

<https://doi.org/10.1172/JCI124358>.

TCR repertoire. The few studies in the literature that have used humanized mice to investigate the human thymus T cell repertoire (12–15) are limited to V-J gene usage and complementarity-determining region 3 (CDR3) length distribution analysis (spectratyping). Here, we performed high-throughput and single-cell TCR immunosequencing of human thymocyte subsets generated in human thymi and the periphery of replicate humanized mice to determine whether exposure to the same antigens for positive and negative selection in thymus tissue from the same human donor would lead to the formation of similar TCR repertoires in different individuals. We also used this model to investigate the impact of positive and negative selection on the human TCR repertoire and to examine the factors that determine V-J pairing. Although initial repertoire formation was highly stochastic and differed markedly among biological replicates, we demonstrate preferential selection of a set of “public” sequences that can be selected by disparate HLA alleles. Comparison with known allo-cross-reactive and type 1 diabetes-associated (T1D-associated) autoreactive TCRs support the interpretation that these are highly cross-reactive TCRs that are preferentially selected and shared between different individuals.

Results

Kinetics of human cell development in humanized mice and the histology of grafted thymi. To study the formation of the human thymic and peripheral TCR repertoire, we generated 3 batches of mice. As shown in Figure 1A, the first experiment consisted of 3 mice (1autoA, 1autoB, and 1autoC) that were generated by transplantation with the same fetal liver HSCs and autologous fetal thymus. Therefore, they had the same genetic background, and selection took place in the same thymus (Figure 1A). The second batch consisted of 6 mice that were generated by transplantation with the same fetal liver HSCs (different from experiment 1). Mice designated 2autoA, 2autoB, and 2autoC received an autologous fetal thymus, while the mice designated 2alloA, 2alloB, and 2alloC received an allogeneic fetal thymus, so that thymic selection occurred in a different thymus, whereas the thymocyte genetic backgrounds were the same as those of the other 3 mice (Figure 1B). Experiment 3 consisted of 2 mice (2autoA and 2autoB) that were thymectomized and transplanted with the same fetal liver HSCs and autologous fetal thymus. We also analyzed and sequenced peripheral CD4⁺ and CD8⁺ cells in these mice, in addition to thymic single-positive CD4⁺ (SP-CD4) and SP-CD8 cells. The mice in the first experiment were euthanized 14 weeks after transplantation, whereas the mice in the second and third experiments were euthanized 20 and 22 weeks after transplantation, respectively. Supplemental Figure 1A (supplemental material available online with this article; <https://doi.org/10.1172/JCI124358DS1>) shows the gross appearance of the spleen, lymph nodes (LNs), and the grafted thymus under the kidney capsule of a representative humanized mouse at the time of harvest. Supplemental Figure 1B shows H&E staining of a representative grafted thymus and a thymus from a 13-year-old child. Cortical (hypercellular) and medullary (hypocellular) areas and Hassall’s corpuscles in the medullary areas are noticeable in the H&E stains. Immunofluorescence staining of a representative grafted thymus and a thymus from a 13-year-old child stained for HLA-DR, cytokera-

tin 8 (CK8) and CK14 is shown in Supplemental Figure 1C. HLA-DR⁺ cells that were not stained for CKs were HSC-derived antigen-presenting cells (APCs) that were mainly concentrated in medullary areas. We further characterized the APCs in grafted thymi by flow cytometric analysis (FCM). B cells (CD19⁺), monocytes (CD14⁺), and DCs (CD11c⁺) collectively constituted approximately 30% of the double-negative (CD4⁻CD8⁻) cells in grafted thymi (Supplemental Figure 1D).

The kinetics of the peripheral appearance of human immune cells (hCD45⁺), B cells (CD19⁺), and T cells (CD3⁺), as well as the T cell naive/memory phenotype are shown in Supplemental Figure 1, E–H. The majority of T cells in peripheral blood at weeks 14–16 were naive.

Our method for constructing humanized mice included several measures to eliminate preexisting thymocytes and their progeny from the transplanted fetal thymic tissue. These measures included freezing and thawing the thymus tissues as described previously (16), pipetting up and down to physically release thymocytes, and injecting 2 weekly doses of a depleting anti-CD2 antibody as described previously (16). To assess the role of cells carried in the thymic tissue in producing peripheral and intrathymic T cell populations in this model, we generated a batch of mice with allogeneic fetal HSCs and thymus tissue. The fetal thymic cells were HLA-A3⁻, whereas the fetal HSCs were HLA-A3⁺. Twenty-four weeks after transplantation, we euthanized the animals and evaluated the origin of T cells in grafted thymi and peripheral lymphoid tissues. Approximately 3% of double-positive (DP) and SP-CD8 thymocytes and 2% of SP-CD4 cells were thymus graft derived (HLA-A3⁻) (Supplemental Figure 1I). Approximately 0.5% of CD4⁺ and CD8⁺ cells in the spleen were thymus graft derived (Supplemental Figure 1J). Therefore, the majority of T cells in the grafted thymi and spleens of these animals were derived from the HSCs that were given intravenously.

Effect of selection on diversity. The cell counts of grafted thymi in addition to the sorted cell numbers are summarized in Supplemental Table 1. For each sample, we obtained template counts, clonality scores, and unique clone counts at the nucleotide level (for both productive rearrangements and nonproductive rearrangements that include frame shifts or premature stop codons) and the aa level. These data are shown in Supplemental Table 1. Template counts for CD69⁻ DP cells were lower than expected from the number of cells, probably reflecting the rearrangement of TCRβ after acquisition of the DP phenotype in a significant fraction of cells (17). Clonality (a normalized measure of inverse diversity based on CDR3β sequences) in all thymic samples was very low, demonstrating production and selection of a highly diverse repertoire in the human thymus grafts. Clonality scores are typically much higher for both CD4⁺ and CD8⁺ T cells in human peripheral blood, most markedly for CD8⁺ T cells, presumably reflecting antigen-driven expansions (18). Accordingly, clonality of peripheral CD4⁺ and CD8⁺ cells was markedly higher than that of thymic SP-CD4 and SP-CD8 cells in experiment 3 (Figure 1F). Although only some differences achieved statistical significance, all thymocyte subsets (CD8⁺ SP, CD4⁺ SP non-Tregs, CD4⁺ Tregs) showed increased clonality scores for aa compared with nucleotide sequences (Figure 1, D and E, and Supplemental Table 1). Collectively, these results show the effect of selection on narrowing the TCR repertoire, since

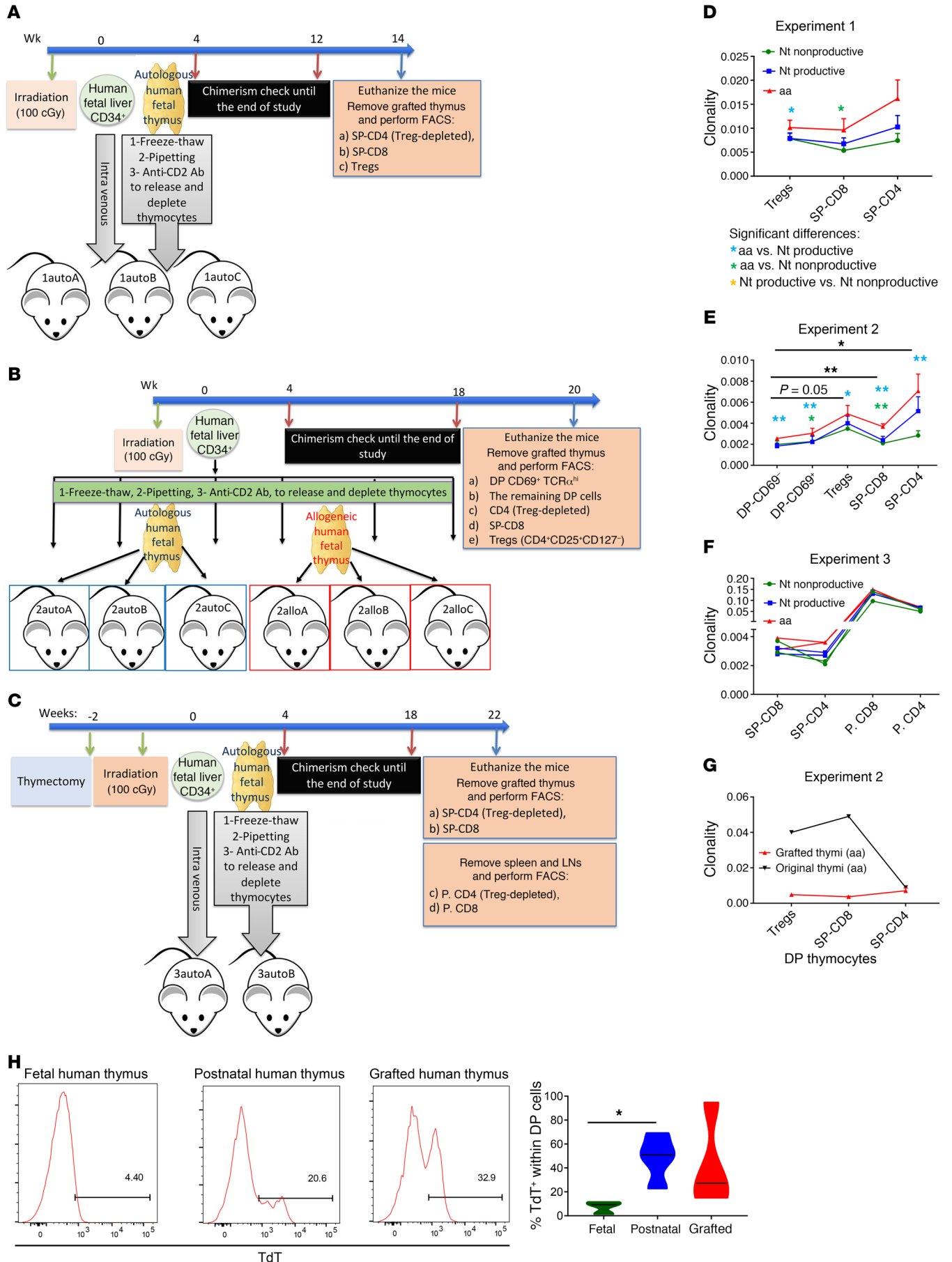


Figure 1. Experimental design and clonality scores. (A–C) Generation of humanized mice for experiments 1, 2, and 3. Cell populations were sorted for sequencing at 14, 20, and 22 weeks after transplantation, respectively. **(D–F)** Clonality scores for cell populations in experiments 1, 2, and 3 at the nucleotide nonproductive (Nt nonproductive), Nt productive (Nt productive), and aa levels (mean \pm SEM, except for experiment 3, which shows individual animals). Paired *t* tests were performed to compare the clonality of each sequence set within each cell population. Paired *t* tests with Bonferroni's multiple testing correction were performed to compare different cell populations in experiment 2. **P* < 0.05 and ***P* < 0.01, by paired *t* test (paired by mouse, with Bonferroni's multiple-testing correction). P. CD8, peripheral CD8⁺; P. CD4, peripheral CD4⁺. **(G)** Scores for aa clonality of grafted thymi and the original autologous thymus in experiment 2. **(H)** Expression of TdT in DP thymocytes of fetal (*n* = 3, gestational ages of 17, 20, and 21 weeks), postnatal (*n* = 4, age 4 months, 6 months, 13 years, and 17 years), and grafted human thymi in humanized mice (*n* = 3, at 18, 26, and 33 weeks after transplantation). **P* < 0.05, by unpaired *t* test.

selection is applied to TCR protein and multiple productive nucleotide sequences can produce the same peptide sequence.

In experiment 2, in which CD69⁻ and CD69⁺ DP thymocytes were sequenced in addition to the 3 SP subsets, each animal showed very low clonality scores for the CD69⁻ DP cell population. With the exception of 1 animal (2alloB), we were unable to detect a positive selection-induced increase in clonality in the CD69⁻ to CD69⁺ transition. However, a comparison of CD69⁻ DP cell populations and all 3 SP cell subsets (SP-CD4 non-Treg [referred to hereafter as SP-CD4 for simplicity], SP-CD8, and CD4⁺ Tregs [referred to hereafter as Tregs]) revealed an increase in aa sequence clonality (Figure 1E and Supplemental Table 1). Collectively, these data demonstrate a narrowing of the T cell repertoire due to thymic selection.

Compared with the original fetal thymus, clonality scores were lower in the grafted thymi for SP-CD8 and Treg populations (Figure 1G), demonstrating greater diversity of the thymocytes generated from human HSCs used to construct humanized mice than in the original fetal thymus, which had a gestational age of 17 weeks, when thymic development and generation of a fully diversified repertoire is not complete. It has been previously reported that the TCR repertoire of mouse neonates (day 1 after birth) is much narrower than that of adult mice because of the lack of random nucleotide insertions in the CDR3s (19). As shown in Figure 1H, TdT was not expressed in DP cells of fetal human thymus. However, it was expressed in DP cells of postnatal thymi as well as grafted human thymus in humanized mice, thus explaining the greater diversity of TCR repertoires in grafted human thymi in our study compared with that in the fetal human thymus.

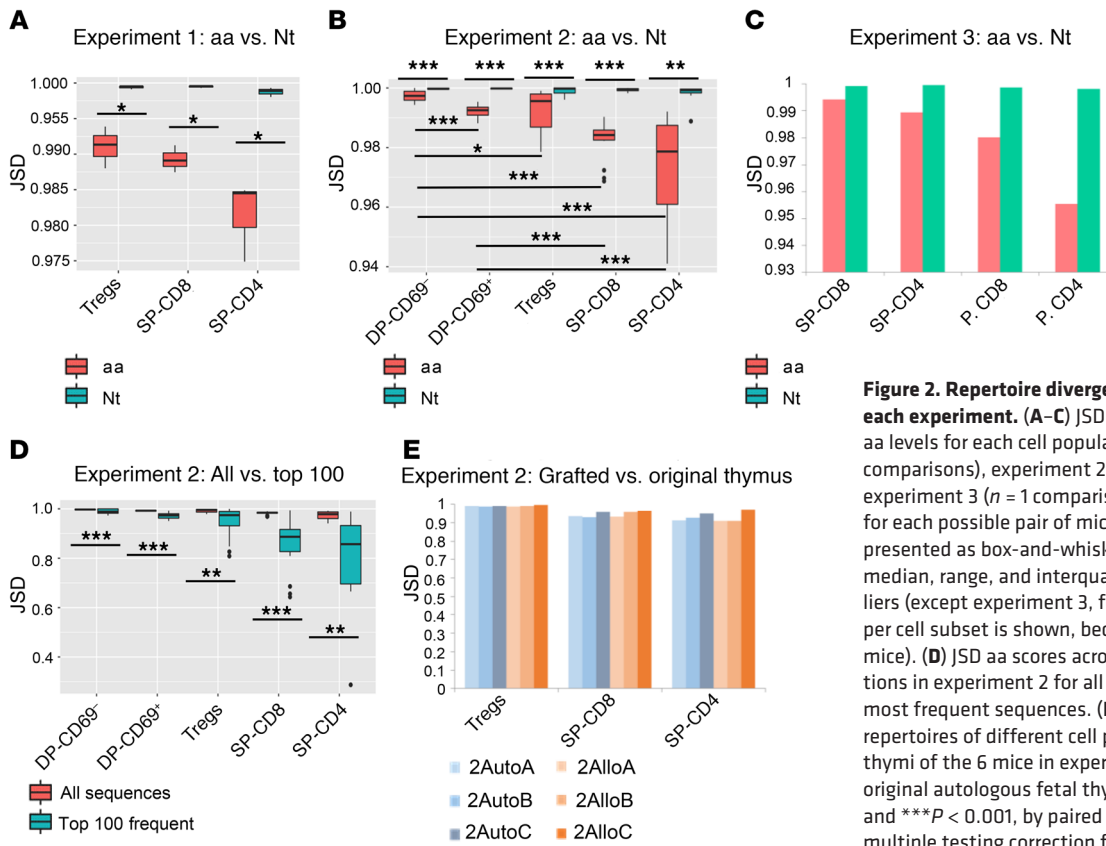
Role of stochastic rearrangement and selection in TCR repertoire formation. To obtain an understanding of the impact of stochastic TCR rearrangement versus background genetics on the TCR repertoire, we compared repertoires generated under the same conditions from the same progenitor pool across identical as well as allogeneic, extensively HLA-mismatched (Supplemental Table 2) thymi by measuring the Jenson-Shannon divergence (JSD) and the number of shared CDR3 β TCR sequences, as the shared CDR3 β fraction quantifies sharing of unique sequences and JSD additionally accounts for the frequency of shared sequences. In experiment 1, even though all 3 mice received the same HSCs and thymus from the same human fetal donor, their TCR repertoires at the level of

CDR3 β were highly divergent at both the nucleotide and aa levels (Figure 2A). In experiment 2, in which 6 mice received the same HSCs, with 2autoA, 2autoB, and 2autoC mice receiving autologous thymus tissues and 2alloA, 2alloB, and 2alloC mice receiving allogeneic thymus tissues, we observed a similarly high divergence among all thymi in different cell populations (Figure 2B). Furthermore, there was no difference in divergence between pairs of mice whose T cells developed in the same thymus versus those whose T cells developed in the allogeneic thymi (Supplemental Table 3). In addition, the observed divergence between mice for both aa and nucleotide repertoires was significantly higher than the baseline generated from repeated undersamplings of identical repertoires for all thymic subpopulations, as determined by both JSD (Supplemental Figure 3A) and the shared CDR3 β fraction (Supplemental Figure 3B). All of these findings emphasize the highly stochastic nature of TCR repertoire formation at the level of CDR3 β .

In all experiments, and in every thymocyte and peripheral subset, the divergence was lower at the aa level compared with the nucleotide level, and the JSD decreased for selected (CD69⁺ DP and SP cell populations) compared with unselected (CD69⁻ DP) cell populations at the aa but not the nucleotide level (Figure 2, A–C). This finding suggested that, despite the stochastic nature of repertoire formation, thymic selection in identical thymi results in selection of some shared sequences between individuals.

We compared the fraction of CDR3 β s that were shared between every possible pair of mice for each thymocyte population. As shown in Figure 3, the fraction of shared CDR3 β s between paired mice in all 3 experiments was less than 4% of all thymic TCR β s for each mouse. The proportion of shared CDR3 β s was highest at the aa level and also was higher at the productive/nucleotide level compared with the nonproductive/nucleotide level (Figure 3, A–C, and Supplemental Table 4). In addition, the proportion of shared CDR3 β s increased significantly during transition from the CD69⁻ DP to the CD69⁺ DP stage, indicating positive selection for these shared CDR3 β s. The proportion of shared CDR3 β s further increased for the sorted mature (CD3^{hi}CD5^{hi}) SP-CD4 and -CD8 cell populations, which had completed positive selection and partially undergone negative selection, compared with the positively selected CD69⁺ DP cell population (Figure 3B). CDR3 β sharing was even higher in peripheral CD4⁺ and CD8⁺ cell subsets compared with thymic SP-CD4 and SP-CD8 samples in experiment 3 (Figure 3C), possibly due to repertoire narrowing via completion of negative selection and post-thymic selection and expansion of certain clones. Pairwise divergence analyses by JSD yielded similar results, with significant decreases in DP-CD69⁺ compared with DP-CD69⁻ cells and further decreases in the mature SP cell populations (SP-CD4, SP-CD8) (Figure 2B) and peripheral CD4⁺ and CD8⁺ cell populations (Figure 2C). Together, these findings demonstrate that both positive and negative selection of human thymocytes increases CDR3 β overlap between individual T cell repertoires. However, the highest proportion of sequences shared between 2 replicate thymic repertoires found in the SP-CD4 subset accounted for only 3.5% of the repertoire.

As another readout of the effect of selection, we compared aa sequence convergence of nucleotide sequences for shared and unshared CDR3 β s. For each pair of mice in experiment 2, we measured the number of unique CDR3 β nucleotide sequences



corresponding to each aa sequence shared between the same cell populations in both mice (shared) compared with the number of unique nucleotide sequences corresponding to each aa sequence present in at least 1 of the mice but not shared between both (unshared). Although the average nucleotide-per-aa sequence ratio was close to 1 for unshared sequences, it was significantly higher for the shared CDR3 β s in all cell populations, indicating preferential selection of shared aa sequences (Figure 3G and Supplemental Table 5). Within the population of shared CDR3 β s, this ratio was significantly higher in DP-CD69 $^{+}$ cells compared with that in DP-CD69 $^{-}$ cells and in SP cell populations (except Tregs) compared with the ratio in DP-CD69 $^{+}$ cells, indicating selection for the shared sequences (Supplemental Table 6).

The JSD between different mice was lower among the 100 most frequent sequences compared with all sequences for each of the 5 selected and nonselected cell populations, indicating greater overlap among the more abundant sequences (Figure 2D). Consistently, the fraction of CDR3 β sequences overlapping between animals was greater among the top 100 sequences compared with the entire cell population (Figure 3D). Although this finding may reflect the greater likelihood of detecting abundant sequences in general, it is also consistent with the possibility that the shared CDR3 β s are preferentially selected.

Surprisingly, for the 5 different selected and nonselected cell populations, the proportion of shared CDR3 β s was not different between mice with allogeneic versus autologous thymi (Figure 3E). Furthermore, we detected no dramatic increase in shared CDR3 β s among mice within an experiment compared with those between

experiments, despite the different genetic backgrounds of the HSCs and thymi used to generate the T cells in each experiment (Figure 3F). In addition, different cell subsets in allogeneic and autologous thymi in experiment 2 had similar divergences compared with the original autologous fetal thymus (Figure 2E). Supplemental Tables 7 and 8 show the numbers of unique and shared CDR3 β s between the 3 mice that were generated by transplantation with the same thymus and HSCs in experiment 1 and the 6 mice transplanted with allogeneic and autologous thymi in experiment 2, respectively. Supplemental Table 9 shows the total number of shared and nonshared CDR3 β s for each cell population in each experiment at both aa and nucleotide levels. Consistent with the results described above, the number of overlapping CDR3 β s increased as the selection progressed, and we detected dramatically larger numbers of overlapping CDR3 β s at the aa sequence level compared with numbers at the nucleotide sequence level.

In order to address variable template counts across samples, we validated the results of the repertoire divergence analysis by randomly subsampling each sample to the same template count and then repeating the analysis. With 3 subsamples of 1000 templates each, we observed the same trends as in whole-sample comparisons with regard to the shared CDR3 β fraction (Supplemental Figure 3C). Specifically, we observed consistently increased sharing at the aa level compared with the nucleotide level and increased sharing in DP-CD69 $^{+}$ samples compared with DP-CD69 $^{-}$ samples, and in SP compared with DP samples. Therefore, our results are stable across random subsamples of the data, regardless of the variable sample sizes (Supplemental Figure 3C).

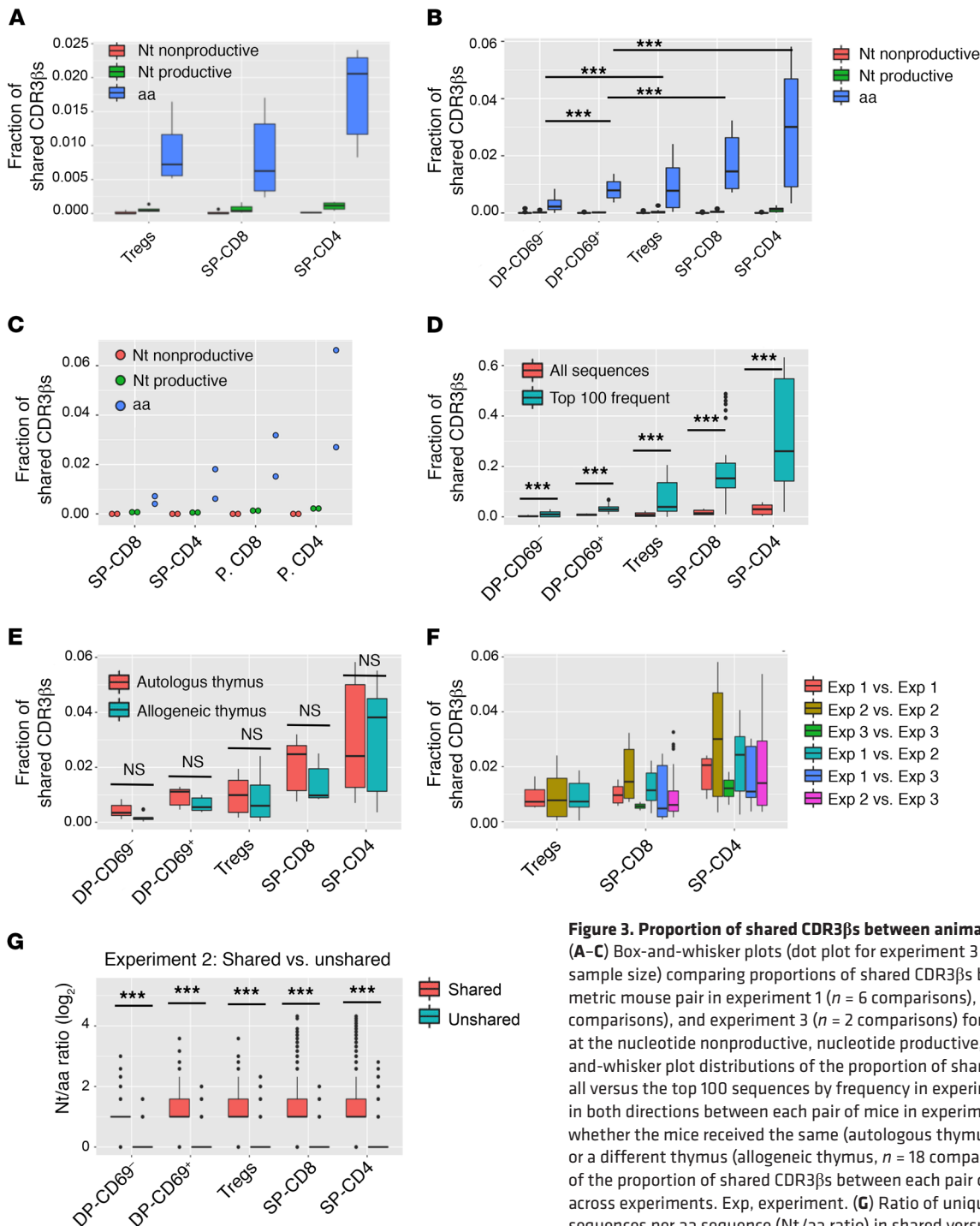


Figure 3. Proportion of shared CDR3βs between animals and experiments. (A–C) Box-and-whisker plots (dot plot for experiment 3 because of the smaller sample size) comparing proportions of shared CDR3βs between each asymmetric mouse pair in experiment 1 ($n = 6$ comparisons), experiment 2 ($n = 30$ comparisons), and experiment 3 ($n = 2$ comparisons) for each cell population at the nucleotide nonproductive, nucleotide productive, and aa levels. (D) Box-and-whisker plot distributions of the proportion of shared CDR3βs comparing all versus the top 100 frequent by frequency in experiment 2. (E) Comparisons in both directions between each pair of mice in experiment 2, depending on whether the mice received the same (autologous thymus, $n = 12$ comparisons) or a different thymus (allogeneic thymus, $n = 18$ comparisons). (F) Distributions of the proportion of shared CDR3βs between each pair of mice within and across experiments. Exp, experiment. (G) Ratio of unique CDR3β nucleotide sequences per aa sequence (Nt/aa ratio) in shared versus unshared sequences for each pair of mice in experiment 2. Supplemental Table 5 shows the mean Nt/aa ratio for each subset and P values comparing different subsets. Box-and-whisker plots show the median, range, interquartile range, and outliers. *** $P < 0.001$, by paired t test with Bonferroni’s multiple testing correction.

Shared CDR3βs have a shorter length due to fewer N insertions than do unique CDR3βs and often use different V genes. Further characterization of the CDR3βs that were shared between any 2 thymi versus those that were detected in only 1 thymus (unique sequences) revealed that the shared CDR3βs were significantly shorter than the unique CDR3βs. The shared CDR3βs had an

average length of approximately 40 nucleotides, whereas the unique CDR3βs had an average CDR3β length of approximately 44 nucleotides (Figure 4A). The number of inserted nucleotides at V-D and D-J junctions was significantly lower for the shared CDR3βs compared with numbers for the unshared CDR3βs (Figure 4B). As the number of V and J nucleotide deletions was slightly

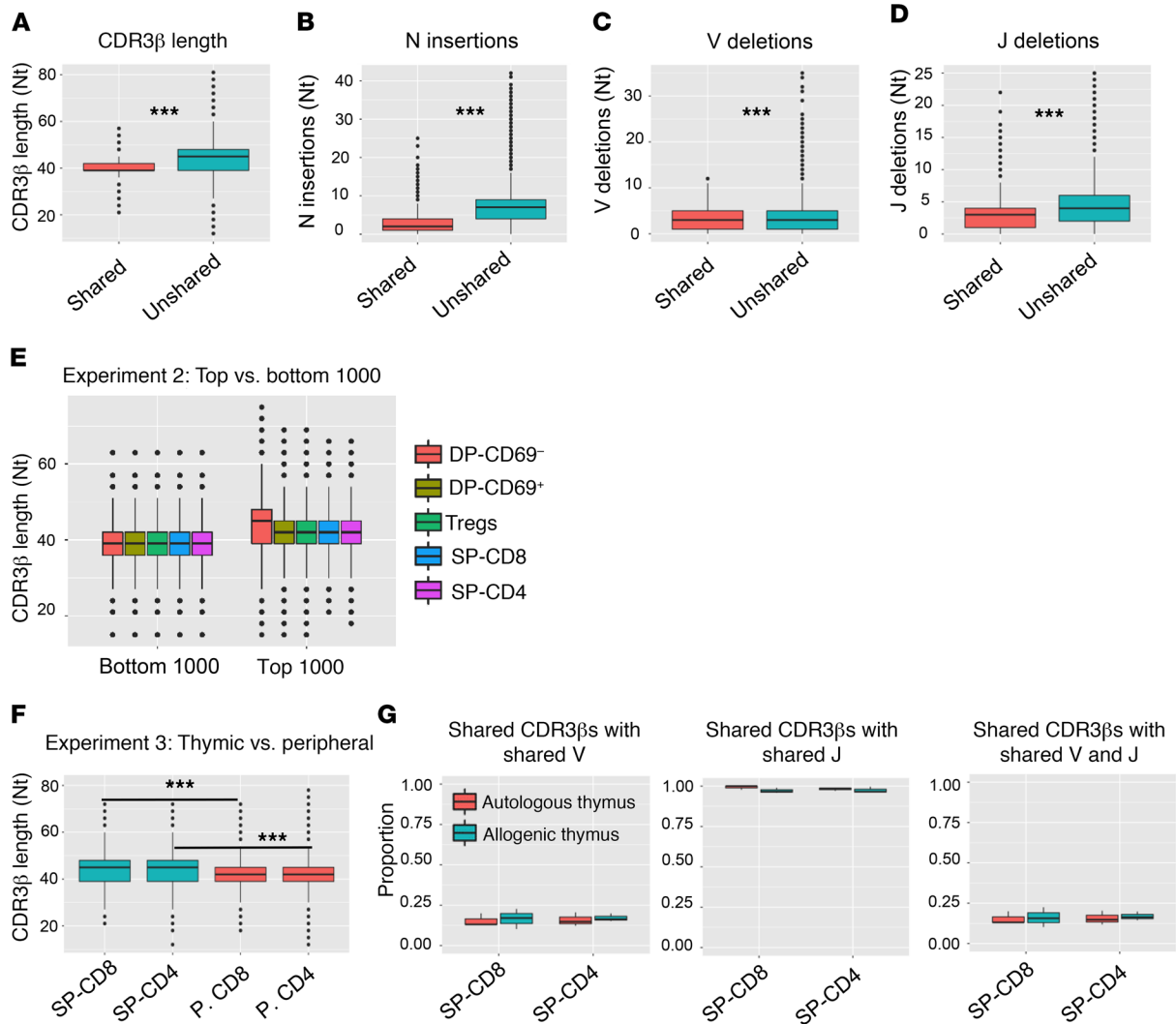
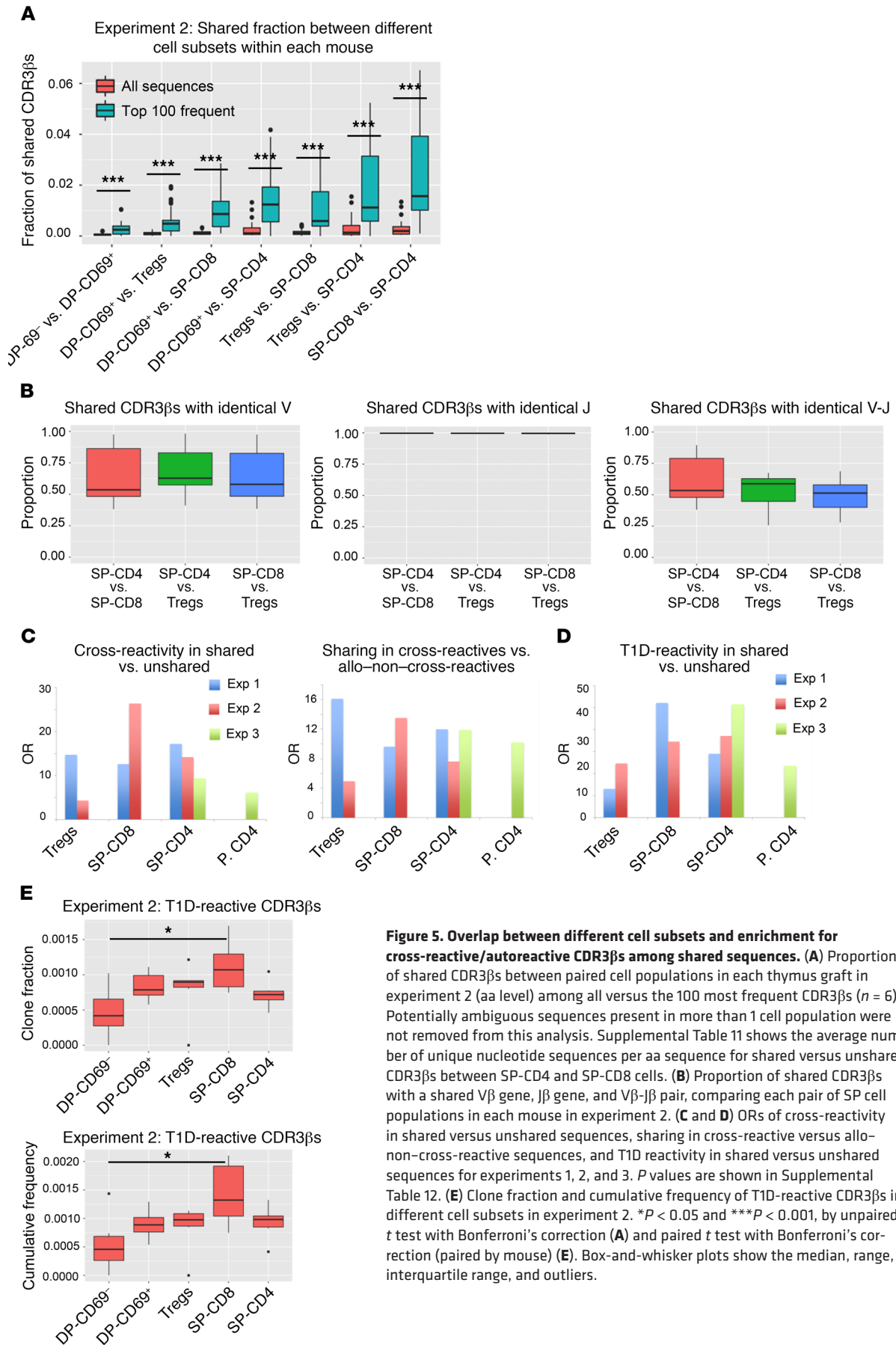


Figure 4. Characteristics of shared versus unshared CDR3 β s. (A) Nucleotide length distribution of shared versus unshared SP-CD4 CDR3 β s in experiment 2. (B) Number of nontemplate nucleotide insertions at V-D plus D-J junctions for shared versus unshared SP-CD4 CDR3 β s. (C and D) Number of nucleotides that are deleted from the 3' end of V genes and the 5' end of J genes at V-D and D-J junctions of SP-CD4 CDR3 β s, respectively. (E) Distribution of combined (all 6 mice in experiment 2) CDR3 β length for the 1000 most frequent CDR3 β s and the 1000 CDR3 β s with the lowest frequencies across different thymic cell populations. The Supplemental Table 6 shows *P* values comparing different cell subsets (unpaired *t* test). (F) Nucleotide CDR3 β length for all thymic and peripheral T cell subsets in experiment 3. (G) Proportion of shared CDR3 β s (aa level) using the same V β gene, J β gene, and V β -J β pair for SP-CD4 and SP-CD8 T cell populations, comparing mice with the same (autologous) versus allogeneic thymus in experiment 2. ****P* < 0.001, by paired *t* test with Bonferroni's multiple testing correction. Box-and-whisker plots show the median, range, interquartile range, and outliers.

higher in unshared CDR3 β s (Figure 4, C and D), the shorter length of shared CDR3 β s was thus attributable to the lower number of nucleotide insertions in these sequences. The shorter length of shared CDR3 β s did not simply reflect the fact that they tended to be relatively abundant, as the average CDR3 length of the 1000 most abundant CDR3 β s overall was significantly greater than that of the 1000 least abundant CDR3 β s (Figure 4E and Supplemental Table 6). However, each animal showed CDR3 β shortening as selection progressed from the CD69⁺ DP to the SP stage among the 1000 most abundant sequences, but not among the least abundant sequences (Figure 4E and Supplemental Table 10), indicating a selective preference for shorter shared CDR3 β s. Shortening of CDR3 β s continued further in the transition from thymic SP to peripheral CD4⁺ and CD8⁺ cells (Figure 4F).

We characterized the V and J gene usage of shared CDR3 β s among SP-CD4 cells. Only approximately 20%–25% of CDR3 β s shared between different thymi used the same V gene, whereas almost all shared CDR3 β s used the same J gene. The percentage of shared CDR3 β s that used the same V-J pair and hence the same TCR β chain is therefore 20%–25% (Figure 4G). These percentages were not different when allogeneic versus autologous thymi were compared.

TCR β chain overlap between different cell subsets in individual human thymus grafts. To better understand the selection of shared CDR3 β sequences, we compared CDR3 β sequences of SP T cell populations within each mouse in experiment 2. As shown in Figure 5A, there was overlap in CDR3 β sequences between SP thymocyte populations from each mouse, especially among the 100



most frequent sequences. Among the sequences with identical CDR3 β s in different mature thymocyte subsets, approximately 60% used the same V gene within individual mice (Figure 5B), whereas approximately 40% used different V genes. Almost all of the shared CDR3 β s were associated with the same J gene, so approximately 60% of TCRs with shared CDR3 β s used the same V-J pair and shared the entire TCR β chain (Figure 5B). Since the HLAs that SP-CD4 and SP-CD8 cells are selected on are different, these results suggest that cross-reactive TCR β chains can be selected on different MHCs. CDR3 β s that were shared in both SP-CD4 and SP-CD8 cells of each mouse in experiment 2 had an average nucleotide-per-aa sequence ratio of approximately 2, whereas the CDR3 β s that were not shared had an average ratio close to 1, pointing to preferential selection of the shared CDR3 β s in both T cell subsets (Supplemental Table 11).

Shared CDR3 β s are more likely to be cross-reactive than are unshared sequences. The data above suggested that shared CDR3 sequences might be highly cross-reactive against disparate specificities. To address this possibility, we compared the repertoires of shared and unshared CDR3 β s from SP cell populations in both experiments with a list of cross-reactive CDR3 β s defined by a greater than 2-fold frequency expansion in mixed lymphocyte reactions of a human peripheral blood sample against 2 different allogeneic donors sharing no HLA alleles. Among 100,112 and 29,033 alloreactive CDR3 β sequences, 1,019 sequences expanded to both stimulators and were therefore identified as cross-reactive. Fisher's exact test revealed a highly significant increase (Supplemental Table 12) in the rate at which shared versus unshared CDR3 β sequences from experiments 1, 2, and 3 were cross-reactive against 2 different sets of alloantigens (Figure 5C). Conversely, we observed a highly significant increase in the odds of allo-cross-reactive sequences compared with alloreactive but non-cross-reactive sequences being shared between the mice in experiments 1 and 2 (Figure 5C). *P* values by Fisher's exact test for the OR of cross-reactivity in shared versus unshared sequences as well as for the OR of sharing in cross-reactive sequences versus allo-non-cross-reactive sequences are listed in Supplemental Table 12. These data demonstrate that shared CDR3 sequences are more cross-reactive than are unshared sequences.

Selection of autoreactive TCRs. In view of the evidence for cross-reactivity of shared sequences selected between disparate thymi and subsets, we hypothesized that shared sequences might be enriched for autoreactivity. We interrogated a previously described list of 1655 T1D-associated autoreactive CDR3 β s (20), along with some newer unique CDR3 β aa sequences (total of 2208 sequences) associated with T1D, largely from peripheral blood but also found in pancreas, LNs, and spleen of T1D donors from the network for Pancreatic Organ donors with Diabetes (nPOD) program (21). These sequences were derived from a number of assays including sequencing of T cells following FACS proliferation of dye-labeled responding T cells harvested following culture with autoantigens (22), direct MHC tetramer isolation of autoreactive T cells (22–25), or following isolation and examination of peptide reactivities from islet-infiltrating T cells (26). T1D reactivity for these sequences was defined as reactivity to islet antigens such as GAD65 and insulin as described previously (21).

Comparison of these autoreactive TCRs with the TCR repertoires of grafted thymi in experiment 2 revealed a significant increase in both the cumulative frequency and clone fraction of T1D-associated sequences in SP-CD8 versus DP-CD69⁺ cell populations (Figures 5E). Remarkably, the odds that a CDR3 β shared between SP subsets in any 2 mice in experiments 1, 2, or 3 was T1D reactive was highly significantly greater than that for nonshared CDR3 β s (Figure 5D), suggesting that shared CDR3s were enriched for autoreactivity. The *P* values for the odds of T1D reactivity in shared versus unshared sequences are listed in Supplemental Table 12.

CDR3 α and TCR sharing from single-cell sequencing. To determine the extent to which CDR3 β sharing was associated with sharing of the entire TCR, including the α chain, we performed single-cell TCR sequencing of thymic SP-CD4 cells from the same mice whose cells were bulk-sequenced in experiment 2 (except mouse 2autoA, due to a technical failure). Comparing each pair of mice, we found that the level of CDR3 α sharing was significantly higher than that of CDR3 β sharing (Figure 6A). However, the level of sharing for paired CDR3 α -CDR3 β was near zero and significantly lower than for either TCR chain on its own (Figure 6A), showing that the TCRs were almost always different among clones with a shared CDR3 α or β sequence. Consistent with the findings from bulk sequencing, the levels of shared CDR3s were not different between mice with allogeneic versus autologous thymi, either for TCR α , TCR β , or paired TCR α -TCR β (Figure 6B). The number of unique CDR3 α s, CDR3 β s and paired CDR3 α -CDR3 β s, the fraction of cells with a β chain that have at least 1 paired α chain or 2 paired α chains and the fraction of cells with an α chain that have a paired β chain is shown in Supplemental Table 13.

Sub-sequence features are conserved in shared CDR3 β s. Methods from Greiff et al. (27), which successfully distinguished between public and private antibody repertoires, were applied to this data set to determine whether sub-sequence-level features can distinguish between shared and unshared sequences. This method uses a normalized gapped k-mer (2 sub-sequences of length *k*, separated by a gap of up to *m* aa) count as an input to a support vector machine (SVM) to determine whether a shared or unshared status can be predicted. Optimal parameters determined by Greiff et al. (*k* = 1, *m* = 1, and cost = 100) were used for SVM analysis, and 10-fold cross validation was performed to assess the performance of the classifier, using balanced accuracy (mean of sensitivity and specificity) as a performance metric. This was repeated on 100 length-matched shared and unshared sequence data sets generated as described above. As shown in Supplemental Figure 4A, these features can be used to predict a shared or unshared status of sequences with a median balanced accuracy of approximately 62% to 78% for all cell subsets, in which 50% would be equivalent to a random classifier. The frequency of gapped k-mers in shared sequences plotted against the frequency in unshared sequences further supported the hypothesis that there are sub-sequence features that are conserved in shared sequences (Supplemental Figure 4B). We also found a notable enrichment in the "CASSL" motif at the 5' end of shared CDR3 β s relative to unshared sequences, even in the unselected CD69⁺ DP cell population (Supplemental Figure 5), though that motif was highly represented in both shared and unshared sequences (Supplemental Figure 5).

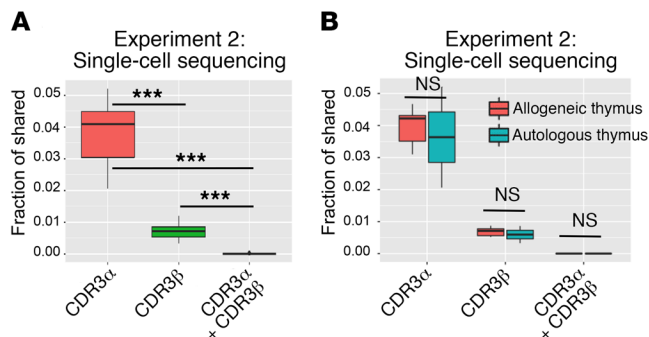


Figure 6. Fraction of shared CDR3 α s, CDR3 β s, and paired CDR3 α -CDR3 β s revealed by single-cell T cell sequencing. (A) Fraction of shared CDR3 α s, CDR3 β s, and paired CDR3 α -CDR3 β s for SP-CD4 cells between each pair of mice in experiment 2 (except 2autoA mice) at the aa level (comparisons in both directions, $n = 20$ comparisons). *** $P < 0.001$, by unpaired t test. (B) Comparisons in both directions between each pair of mice, depending on whether the mice received the same (autologous thymus, $n = 8$ comparisons) or a different thymus (allogeneic thymus, $n = 12$ comparisons).

Evidence suggesting a role for self-peptides in human thymocyte selection. In preselection murine thymocytes, TCR β CDR3 interfacial hydrophobicity at position 6 and position 7 (P6 and P7), the residues that interface with the peptide and MHC, correlated with the ability to be activated by self-peptide and MHC (28). Stadinski et al. developed a self-reactivity index based on the hydrophobicity of aa at CDR3 β P6 and P7 and showed that this index correlates well with increased and decreased self-reactivity during positive and negative selection, respectively. We performed a similar analysis on human thymocyte and peripheral T cell subsets from experiments 2 and 3, focusing on CDR3 β lengths of the greatest frequency in all thymocyte subsets (13–16 aa). For each mouse, we analyzed P6 and P7 aa frequencies in the thymic (DP-CD69 $^-$, DP-CD69 $^+$, SP-CD4 $^+$, SP-CD8 $^+$, and Treg) and peripheral (CD4 $^+$ and CD8 $^+$) cell populations, and normalized the frequencies within each cell population. For each animal, a fold-change in frequency of P6 and P7 aa residues between cell populations was recorded, and these values were averaged across the mice within each grafted thymus group (i.e., experiment 2, allogeneic versus autologous thymus, and experiment 3). For experiment 2 samples, we compared SP-CD4, SP-CD8, and Tregs against DP-CD69 $^-$ thymocytes to evaluate the entire thymic selection process, DP-CD69 $^+$ thymocytes against DP-CD69 $^-$ thymocytes, and SP-CD4, SP-CD8, and Tregs against DP-CD69 $^+$ thymocytes. For experiment 3 samples, we compared peripheral CD4 $^+$ cells and peripheral CD8 $^+$ cells against SP-CD4 and SP-CD8 cells, respectively. As shown for P6 in Figure 7A and Supplemental Figure 6A and for P7 in Figure 7B and Supplemental Figure 6B, we observed a trend toward enrichment of hydrophobic aa (as defined in Figure 8C in ref. 29) at both positions as thymic selection progressed. Results of the Spearman's nonparametric rank test for the mice in experiments 2 and 3 are shown in Figure 7, A and B. We observed statistically significant correlations between the fold changes of the aa residue at P6 (Figure 7A) or P7 (Figure 7B) and its hydrophobicity during selection from the DP-CD69 $^-$ stage to the mature SP-CD4, CD8 $^+$, and Treg populations. Both autologous and allogeneic thymi showed a similar trend toward increasing hydrophobicity

at P6 and P7 as selection progressed (Supplemental Figure 6). Positive selection from the CD69 $^-$ to the CD69 $^+$ DP stage was associated with significantly increased P6 hydrophobicity, and overall selection from CD69 $^-$ DP to both CD4 $^+$ and CD8 $^+$ SP populations was associated with significantly increased hydrophobicity at P6 and P7 and at P6 for the CD69 $^-$ -to-Treg transition. The CD69 $^+$ DP to SP transition was associated with significantly increased hydrophobicity only for SP-CD8 cells at P6 and for SP-CD8 and Tregs at P7. Overall, we found that increased hydrophobicity with the transition from DP-CD69 $^-$ to SP cells was more pronounced than with the transition from DP-CD69 $^+$ to SP cells. This trend was stopped or reversed in the transition from SP-CD4 and SP-CD8 cells to peripheral CD4 $^+$ and CD8 $^+$ cells, both at P6 and P7 (Figures 7, A and B, and Supplemental Figure 6). In sum, our data demonstrate an increase in hydrophobic aa usage at P6 and P7 in association with selection of human thymocytes (more associated with positive selection) and arrest or reversal of this trend in the transition from SP thymocytes to peripheral T cells, possibly in association with completion of negative selection.

Shared sequences might escape negative selection. To analyze differential usage of aa at each position as defined by the international ImMunoGeneTics (IMGT), we performed a Fisher's exact test for all sequences in each of the 100 length-matched data sets of shared and unshared sequences. Differentially used aa were plotted if the Benjamini-Hochberg-adjusted P value was less than 0.05 for the Fisher's exact test to ensure that differences were significant (Figure 7C). Only aa showing up in at least 75 of the 100 downsamples were annotated. We noted a significant enrichment for the neutral aa G and hydrophilic aa (e.g., Q) and a significant decrease in hydrophobic aa (e.g., W) at P6 and P7 (equal to positions 109 and 110 on the plots, respectively) in shared sequences among CD69 $^+$ DP cells, most thymic SP cell populations, and peripheral cell populations (Figure 7C). We did not observe this pattern for shared sequences among CD69 $^-$ DP cells. The reduced hydrophobicity at P6 and P7 in shared sequences among selected but not unselected cell populations suggests that selected shared sequences may have weaker interactions with self-peptides than unshared sequences and that this may allow them to escape negative selection.

V and J gene usage. As shown in Supplemental Figure 7, the pattern of V and J gene usage for the SP-CD4 cell population was very similar between the mice that received autologous tissue versus those that received allogeneic thymus tissue in experiment 2, with no significant differences in V or J gene usage, arguing against a major role for selection in determining V and J gene usage. Overall, we detected a similar pattern of V and J gene usage for SP-CD4 cells when comparing mice in experiments 1, 2, and 3, which received different HSCs as well as different thymi (Supplemental Figure 7). We also observed a similar pattern of V and J gene usage between thymic SP-CD4 and peripheral CD4 $^+$ cells in the mice in experiment 3 (Supplemental Figure 7). We also observed similar V and J gene usage patterns across all cell populations for the mice in experiment 2 (Supplemental Figure 8). Few statistically significant differences are shown in Supplemental Figure 7 and Supplemental Figure 8. Thus, we detected only small effects of genetic background and/or thymic selection on the overall pattern of V and J gene usage. These findings were confirmed by

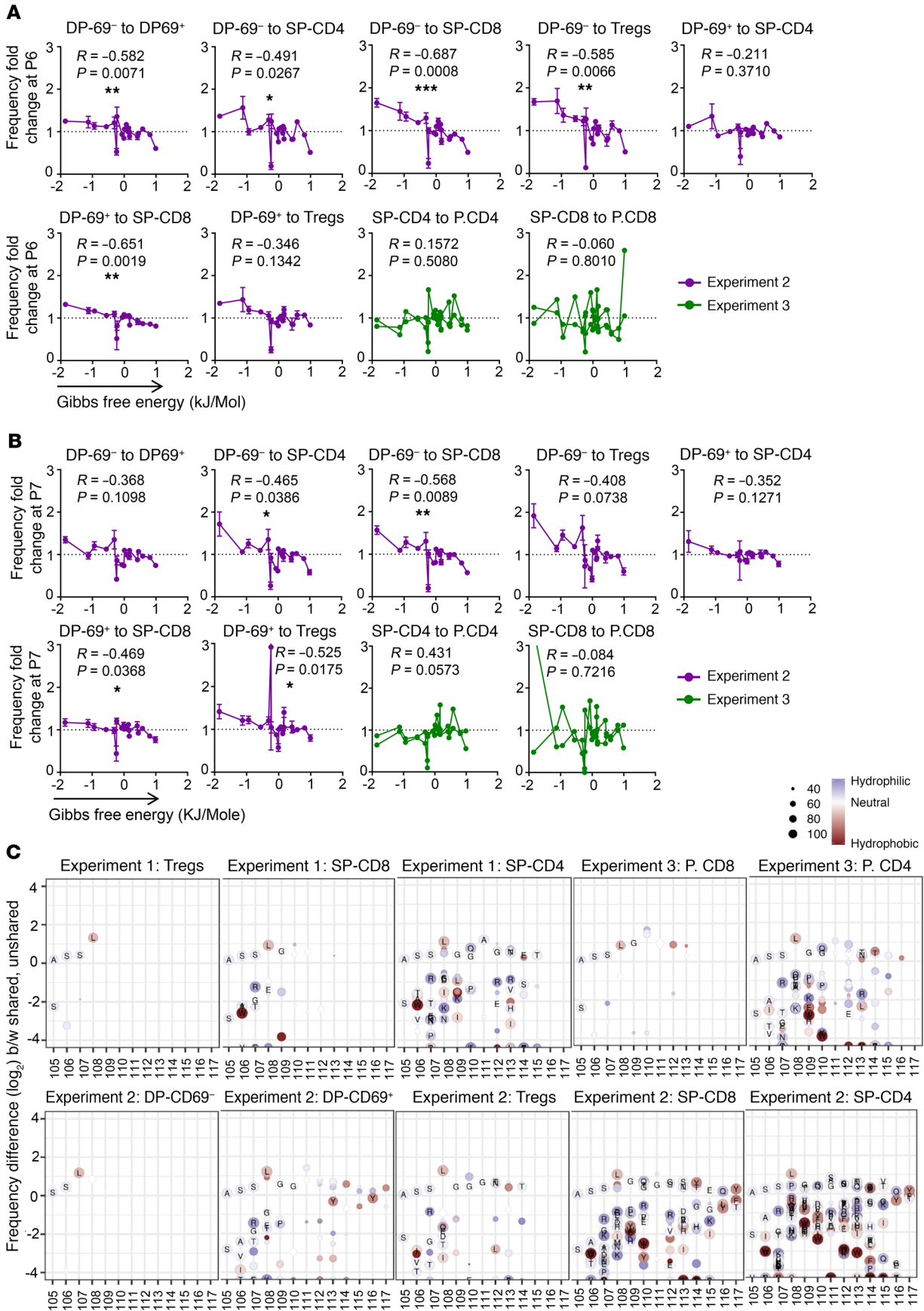


Figure 7. Interaction with self-peptides in the selection of shared and unshared sequences. Fold changes (mean \pm SEM) in the relative aa frequencies versus hydrophobicity of the aa based on Gibbs free energy at P6 (A) or P7 (B) for transition from DP-CD69⁺ to DP-CD69⁺ cells and from there to SP cell subsets in experiment 2, and also in transition from SP-CD8 and SP-CD4 to peripheral CD8⁺ and CD4⁺ cells for experiment 3. Spearman's correlation coefficient *R* and *P* values from the nonparametric Spearman's correlation test are shown. Negative *R* values imply that, as hydrophobicity increases, so does the fold change in the relative aa frequency across the 2 cell populations. **P* < 0.05, ***P* < 0.01, and ****P* < 0.001, by unpaired *t* test. (C) Differential abundance of each aa at each position in CDR3 β , computed by random selection of a length-matched unshared sequence for each shared sequence. Shared sequences are those present in at least 2 mice, and unshared sequences are unique to a single mouse. Only results for aa producing a Benjamini-Hochberg-adjusted *P* value of less than 0.05 by Fisher's exact test are shown. The aa plotted at a frequency of 0 were preferentially used at that position in shared sequences, whereas those with a frequency of less than 0 were preferentially used in unshared sequences.

single-cell TCR-sequencing data, which showed a similar pattern of V and J gene usage for both α and β chains comparing SP-CD4 cells in mice with allogeneic versus autologous thymi in experiment 2 (Supplemental Figure 9).

Supplemental Figure 10A shows plots of VJ usage for CD4⁺ repertoires of 1autoA, 1autoB and 1autoC mice, which received the same thymic tissue and HSCs. These representative plots show no disproportionately favored V-J pairing in the repertoire. We detected a strong correlation between the observed VJ usage and the VJ usage expected from the stochastic combination of V genes with J genes according to the background frequency of each V and J (Supplemental Figure 10B). The observed and expected VJ distributions were compared by Mann-Whitney *U* test, which failed to reject the null hypothesis that VJ pairing is stochastic.

Discussion

We have demonstrated the impact of positive and negative selection in a human thymus on the human TCR repertoire. We observed very high diversity among human thymocytes at all stages of development. This diversity was much greater than that for peripheral blood human T cells that include memory cell populations (18), consistent with studies demonstrating that TCR repertoires of human naive T cells are much more diverse than those of memory T cells (3). The observed lower diversity at the aa versus the nucleotide sequence level, in productive versus nonproductive sequences, and in selected SP cell populations (SP-CD4, SP-CD8, and Tregs) versus nonselected DP thymocytes demonstrates that thymic selection narrows the human TCR repertoire at both the DP and the SP stages. We demonstrated that diversity further decreases in peripheral CD4⁺ and CD8⁺ cells in humanized mice.

The diversity of the SP-CD4 and SP-CD8 TCR repertoires in autologous and allogeneic grafted thymi was greater than that of the original fetal thymus used for the generation of the humanized mice, reflecting the immaturity of the repertoire in the 17-gestational-week fetus used and confirming that the thymocytes that developed in our studies arose de novo from engrafted stem cells as a result of the success of our procedures for purging the graft of preexisting thymocytes prior to implantation. The TCR repertoire of neonatal mice is much narrower than that of adult mice until day 4 or 5 after birth, when the TdT enzyme is acti-

vated (19). Limited studies indicated that CD45⁺ cells enter the human fetal thymus at 8 weeks of gestation and that all steps of TCR development are detectable by week 16 (30), but TdT was reportedly undetectable in fetal human thymus (31), consistent with our observation of very low TdT expression in DP cells of such thymi. In contrast, we detected much higher TdT levels in postnatal human thymi and similar levels in grafted human thymi, thus explaining the greater TCR diversity in grafted human thymi compared with fetal thymus donors. While grafted and human postnatal thymi showed very similar structures and cell populations, we cannot rule out the possibility that undetected structural differences might influence selection.

A high divergence of TCR β repertoires in thymocytes generated in identical thymi from the same HSCs at the same time is consistent with the stochastic nature of TCR repertoire formation and may contribute to the incomplete penetrance of genetically controlled autoimmune diseases in identical twins (32). Increased CDR3 β sharing among productive compared with nonproductive TCR rearrangements and in selected versus unselected thymocyte populations and a high convergence of nucleotide sequences at the aa level demonstrate that thymic selection favors these shared TCR β CDR3s. Although many of these shared CDR3 β s used different V genes but were associated with identical J genes, 20%–25% also had identical V genes, indicating that the same entire β chain was used. Single-cell analysis revealed that TCRs with the same β chain are almost always paired with different α chains.

Our observation of shared sequences between different donors and thymi is consistent with studies in mice showing the existence of MHC-independent public, abundant CDR3 sequences with convergent recombination, also with fewer N insertions than average, among peripheral T cells (6). These authors performed numerical simulations of TCR rearrangements to demonstrate that biases in TCR recombination could only partially explain the observation. To our knowledge, our study is the first to specifically analyze the impact of selection on shared human thymocyte sequences and demonstrates that thymic selection is a significant factor driving their abundance among mature T cells. Previous next-generation sequencing studies of the peripheral blood TCR repertoire of monozygotic twins revealed increased shared CDR3 β s among highly abundant clonotypes and a similar overall overlap between the repertoires of monozygotic twins and unrelated individuals (33), consistent with our observation of similar sharing between mice with the same HSCs but different thymi and between mice with different thymi and HSCs. The increased TCR β sharing observed by Qi et al. in peripheral blood from twins compared with unrelated individuals was more pronounced in memory than in naive T cells (34) and may be explicable by post-thymic selection events rather than thymic selection.

Several possibilities could explain our finding of similar TCR sharing among thymocytes developing in autologous versus allogeneic thymi. First, the allogeneic and autologous thymi may share common peptide motifs that contribute to the selection of shared CDR3 β s in the context of different HLAs, as previously suggested (5). According to a model based on sequencing of adult peripheral blood CD8⁺ T cells, CDR3 β sharing was observed at a 500-fold or greater rate than expected from stochastic rearrangements and was independent of HLA sharing (5). Consistent with

our observation of shared sequences among thymocyte subsets, different peripheral blood T cells subsets revealed significant CDR3 β sharing (35). Although peripheral blood T cell studies may reflect post-thymic, antigen-driven expansion, this had not occurred in the thymic grafts in our studies.

We favor a second possible explanation for interindividual and inter-T cell subset CDR3 β sharing, namely, that the shared CDR3 β s are more highly cross-reactive than are nonshared CDR3 β s and can therefore be positively selected by diverse HLA-peptide complexes. In all 3 experiments, we observed an increased overlap of more abundant CDR3 β s. This may reflect the greater likelihood of detecting more abundant sequences in a given sample size. However, our observation of increased CDR3 β sharing following positive selection of DP thymocytes suggests that they are preferentially selected. We also observed an increase in P6 and P7 hydrophobicity during positive selection, demonstrating the role of HLA-peptide complexes in this process. CDR1 and CDR2 interact prominently with MHC to influence CD4 $^+$ and CD8 $^+$ T cell development, whereas CDR3 β predominantly interacts with peptides bound to MHC (36), which differ for recognition by SP-CD4 and SP-CD8 cells. Thus, the hypothesis that most of the shared CDR3 β s are cross-reactive seems most probable. TCRs have flexible CDR3 loops (37) and can bind to different peptide-MHC complexes through different mechanisms, including binding with an altered angle or register (38). For example, a human, preproinsulin-reactive CD8 $^+$ T cell clone (1E6) is capable of recognizing more than 1 million peptides in the context of a single MHC I molecule (39), including microbial peptides as well as self-peptides (40). Such cross-reactivity is necessary for the limited number of TCRs in the body to recognize more than 10^{15} different peptide-MHC complexes (40). In mice, TCRs that survive thymic selection are enriched for cross-reactive TCRs compared with those that do not. Interestingly, cross-reactive TCRs were frequently reactive to both MHC I and II (41). Furthermore, in mice genetically engineered to express human TCR genes with either a single allele of human MHC II or a single mouse MHC II, a surprisingly high number of CD4 $^+$ TCRs was shared between the 2 groups (42). The finding that single MHCs from different species are able to select shared TCRs provides support for our hypothesis about selection of cross-reactive TCRs.

The shared sequences in our study had a shorter CDR3 β length than did nonshared sequences, consistent with previous studies (5, 33), as a result of lower numbers of N insertions. Since more abundant CDR3 β s were, on average, longer than the less frequent ones, even in the preselection DP-CD69 $^-$ cell population, the increased sharing among shorter sequences does not simply indicate an inherent bias for β -selection of shorter CDR3 β s. We found that the average length of more abundant CDR3 β s decreased during thymic selection, consistent with previous reports in humans and mice (43–46). Shorter CDR3 β s may be positively selected more easily, given the increased low-affinity cross-reactions with diverse MHC-peptide complexes. TCRs in mice deficient for TdT, which mediates N insertion in CDR3s, have a shorter average CDR3 length (8) and show increased inter-animal sharing of TCR sequences with increased cross-reactivity for different peptides compared with WT mice (47). We observed increased usage of neutral and hydrophilic

aa (G and Q, specifically) among shared sequences at P6 and P7 (positions 109 and 110 in Figure 7B) compared with unshared sequences following the transition from DP-CD69 $^-$ to DP-CD69 $^+$. This difference was maintained in the SP cell populations, despite the overall increase in hydrophobicity with selection at these positions in the total repertoire. In view of murine studies indicating that hydrophobicity at P6 and P7 of CDR3 β correlates with reactivity to self-peptide-MHC complexes (28), these data suggest that shared sequences may be preferentially positively selected and survive negative selection because of their low affinity for self-peptide-MHC. Consistently, shared sequences in mice had reduced affinity for self-peptide-MHC (27). Thus, while the low affinity of shared sequences for self-peptide-HLA is sufficient to promote positive selection, the low level of hydrophobicity at the peptide binding site may allow these sequences to preferentially evade negative selection in the thymus. The observed average CDR3 β shortening as negative selection progresses (CD69 $^+$ DP to SP transition) among the most abundant but not among the least abundant sequences is consistent with the avoidance of negative selection by these shorter, more commonly shared CDR3 β s. While one SP thymocyte subset (CD8 $^+$ SP) showed an unexpected increase in P6 and P7 hydrophobicity during the transition from the CD69 $^+$ DP phase, we observed more consistent increases with overall selection between the CD69 $^-$ DP and SP stages for each subset, suggesting that the major increase may reflect positive selection. The CD69 $^+$ DP cell population probably had already undergone considerable negative selection, diluting the increased hydrophobicity associated with positive selection, and negative selection probably continued into the SP phase, making comparisons between these populations too complex to interpret.

Our single-cell TCR-sequencing data showed even greater TCR α sharing than TCR β sharing between animals and also demonstrated that almost all TCRs with shared α or β chains do not form the same $\alpha\beta$ pairs. This result implies that the selective pressure for shared CDR3 β s is applied separately to the β chain. Since we did not perform bulk analysis of CDR3 α at various stages of thymocyte selection, we do not know whether the high level of CDR3 α sharing reflects generational bias or selective pressure. Consistent with our results, greater TCR α than TCR β sharing has been reported among healthy humans, while no difference was found in the diversity of CDR3 α s compared with CDR3 β s (48). Structural analyses have shown that both CDR3 α and CDR3 β interact with peptides and contribute to the recognition of peptide-MHC complexes (49). However, a predominance of charged and polar aa in CDR3 α compared with CDR3 β may affect peptide interactions (50).

The comparison of shared and unshared sequences in our study with known autoreactive and cross-reactive TCRs supports our hypothesis that shared CDR3s are cross-reactive. Cross-reactive CDR3 β s identified through sequencing of healthy donor T cells responding to disparate alloantigens were strikingly enriched for shared compared with unshared sequences identified in thymi in both experiments. Moreover, the frequency of known T1D-associated CDR3 β s was increased during positive selection in the grafted thymi, and these sequences were highly significantly increased among shared CDR3s (Supplemental Table 12). These data suggest that autoreactive T1D-associated sequences are

enriched for cross-reactive TCRs that can be selected by different thymi with different MHCs.

Despite the very divergent CDR3 β TCR repertoires, V and J gene usage was overall very similar among different mice generated with different HSCs and different thymi. There is controversy in the literature about the role of genetic background and thymic selection in determining V and J gene usage (33, 51–54). Our analysis of thymocyte subsets generated from different donors demonstrates only a minor role for thymic selection in determining overall V and J gene usage. Intrinsic genomic factors such as promoter strength or differential structural features are probably the main determinants of V and J gene usage in TCRs. However, the small number of HSC donors in our study did not permit detection of more subtle effects of genetic background and different HLA alleles on V and J gene usage, as has been reported (55, 56). A report of increased similarity in V and J gene usage of peripheral naive T cells between identical twins compared with non-twins but overall similar V and J usage frequencies for any 2 donors, regardless of relatedness (56), is consistent with a subtler effect of HLA alleles in selecting for certain V and J genes. In a mouse study (42), selection of T cells on a single murine or human MHC II molecule resulted in some differences in V and J gene usage. However, the proportion of SP thymocytes in these mice was significantly lower compared with that in normal B6 mice, suggesting inefficient selection on a single MHC allele, which could potentially skew the repertoire toward using certain V and J genes. Our study shows that the thymic MHC does not markedly influence human V and J gene usage when there is a full complement of HLA molecules.

Interestingly, sequences shared between mice were enriched across all cell populations for the 5' "CASSL" motif in CDR3 β within the V region of CDR3 β . This motif is found in the 3' tail of TRB5-01, which is consistently the most highly expressed V gene across all samples, at roughly 12%. Though it is enriched in shared sequences, the motif has high representation in both shared and unshared sequences sets. Published structural analysis of TCR binding to the MHC-peptide complex for a clone with this motif shows no direct interaction of the motif with peptide-MHC (57). Our observation of enrichment for this motif in shared sequences of preselection CD69⁺ DP thymocytes suggests a more structural role, such as in increased binding to pre-TCR α , rather than a role in self-antigen-driven selection. After successful rearrangement of the TCR β chain, developing thymocytes undergo an estimated 6–7 rounds of cell division before independent rearrangements of TCR α (58, 59), which could enrich such sequences in preselection thymocytes. Increased stability of the TCR tertiary structure or increased binding to TCR α (60) could also explain this enrichment in shared sequences of preselection thymocytes. Our detection of fewer unique CDR3 α sequences than CDR3 β sequences in the single-cell analysis was surprising in light of the ability of early DP thymocytes to expand during β selection and the ability of individual T cell clones to rearrange only 1 β chain and 2 α chains. This result probably reflects reduced efficiency in detecting α chains as opposed to β chains, depending on the sequencing platform used, as less than 100% of cells with β chains sequenced had a sequenced α chain.

Besides V and J gene usage, some reports indicate that the frequency of particular TCR V β -TCR J β recombinations in human

lymphocytes is controlled genetically (61). By comparing the observed VJ frequency distributions in productive and nonproductive repertoires with the VJ frequency distribution expected from a stochastic combination of V genes and J genes according to their background frequency in thymic TCR repertoires, our results show that VJ pairing is only determined by the amount of expression of each gene, with no evidence for preferential V-J recombinations.

In conclusion, our data indicate that human thymus in humanized mice selects a very diverse TCR repertoire. Formation of the human TCR repertoire is largely stochastic and can be almost totally divergent in thymi of animals with identical HSCs, thymus, genetic background, and environment. However, we show that thymic selection increases the overlap between human TCR repertoires and that recognition of self-peptide-HLA plays a role in human thymocyte selection. The overlap of CDR3 β s in disparate thymi and different thymocyte subsets, the shorter CDR3 β lengths of shared sequences, the direct evidence for allo-cross-reactivity and autoreactivity of shared sequences, and the analysis of aa usage in these CDR3 sequences are consistent with the interpretation that shared sequences are preferentially positively selected because of high cross-reactivity and evade negative selection as a result of low affinity for self-peptides. Thus, we have used the humanized mouse model in our studies to obtain insights into the factors determining human T cell repertoire formation.

Methods

Generation of humanized mice. Three, six, and two humanized mice were generated by combined human fetal thymic and hematopoietic cell transplantation in thymectomized immunodeficient mice for experiments 1, 2, and 3, respectively (Figure 1, A–C; see details in the supplemental materials).

FCM of different subsets of grafted thymus and peripheral cells. Mice from experiments 1, 2, and 3 were euthanized 14, 20, and 22 weeks, respectively, after thymus transplantation. Grafted thymi (for mice in all experiments) and spleens and LNs (only for mice in experiment 3) were harvested, and the thymocytes and pooled splenic and LN cells were isolated and subjected to FCM (Supplemental Figure 2; see Supplemental Methods for details on FCM). HLA typing of fetal tissues used to generate humanized mice in all 3 experiments is shown in Supplemental Table 2.

DNA isolation and high-throughput CDR3 β TCR sequencing. Genomic DNA was isolated from sorted cell populations using the QIAGEN DNeasy Blood and Tissue Kit. DNA was frozen at -20°C and shipped on dry ice to Adaptive Biotechnologies for high-throughput TCR β CDR3 sequencing. The TCR-sequencing data were retrieved using ImmunoSEQ software (Adaptive Biotechnologies). The Adaptive Biotechnologies raw sequencing data on all samples are available at <https://github.com/aleksobrad/humanized-mouse-data> (GitHub commit ID: 96d3c83).

Single-cell TCR sequencing. Single-cell TCR sequencing was performed using the 10 \times Genomics platform as detailed in the Supplemental Method.

Computational and statistical analysis. Computational and statistical analysis methods are described in Supplemental Methods.

Cross-reactive TCR list. PBMCs from a human subject were stained with CFSE and cocultured separately for 6 days with 2 fully HLA-mismatched irradiated PBMCs as stimulator cells. Dividing CD4⁺ and CD8⁺ T cells (CFSE^{lo}) and unstimulated CD4⁺ and CD8⁺ T cells

from the same donor were subjected to FACS, and after DNA isolation their TCR β was sequenced as described above. Alloreactive sequences were defined as CDR3 β aa sequences that were expanded in frequency by at least 2-fold from unstimulated to stimulated samples, above a minimum frequency in the stimulated samples of 1×10^{-5} templates, as described previously (18, 62). Alloreactivity was determined separately for CD4 $^{+}$ and CD8 $^{+}$ samples, and alloreactive CDR3 β s responding to both stimulators were considered cross-reactive, since they responded to different stimulators with no MHC sharing. We compared the repertoires of T cell populations in grafted thymi of humanized mice with the list of cross-reactive TCRs to investigate the selection of cross-reactive TCRs in human thymus, calculating the OR of allo-cross-reactivity in shared versus unshared clones as well as the OR of shared clone status in allo-cross-reactive versus allo-non-cross-reactive sequences, with significance assessed by Fisher's exact test. Lists of the CDR3 sequences determined to be cross-reactive and allo-non-cross-reactive are available as .csv files at <https://github.com/aleksobrad/humanized-mouse-data> (GitHub commit ID: 96d3c83), along with the raw adaptive sequencing data from healthy control mixed lymphocyte reactions (MLRs) of the same responders against 2 fully HLA-mismatched stimulators.

T1D-reactive TCR list. Our T1D-reactive TCR data set contains 2208 unique CDR3 β aa sequences associated with T1D derived from peripheral blood, pancreas, LNs, and spleen of T1D donors from the nPOD program (20). These sequences were derived from several assays including sequencing of T cells following FACS proliferation of dye-labeled responding T cells harvested in response to culture with autoantigens (21), direct MHC tetramer isolation of autoreactive T cells (21–24), and, in certain situations, an assay following the isolation and examination of peptide reactivities from islet-infiltrating T cells (25). T1D reactivity for these sequences was defined as reactivity to islet antigens such as GAD65 and insulin as described previously (26). A single, prominent T1D-reactive nucleotide sequence (encoding CASSFWGSDTGELFF TCRBV11-02 TCRBJ02-02) present in bulk-sequencing data but missing from single-cell data was removed from experiment 2 because of suspected contamination from a vector with the same nucleotide sequence that was present in the laboratory. We compared the repertoires of T cell populations in grafted thymi in humanized mice with this list of T1D-reactive sequences to investigate the selection of T1D-reactive TCRs in human thymus, calculating the OR of T1D reactivity among shared versus unshared sequences, with significance assessed by Fisher's exact test. The .csv files with the CDR3 sequences of T1D-reactive sequences are available at <https://github.com/aleksobrad/humanized-mouse-data> (GitHub branch name: Humanized-Mouse-Data/t1d_sequences.csv).

Statistics. For comparisons between different cell populations analyzing clonality, JSD and fraction of shared sequences, paired *t* tests with Bonferroni's multiple testing correction were performed. For studies that involved calculation of the OR, we performed a Fisher's exact test. The Spearman's correlation coefficient *R* value and *P* value from the nonparametric Spearman's correlation test were reported

where fold changes in the relative aa frequencies are plotted against aa hydrophobicity based on Gibbs free energy. An unpaired *t* test with Bonferroni's correction for multiple testing was performed to compare the V β and J β gene usages between different groups for each gene. For comparison of the observed and expected frequencies of VJ gene pairs, Mann-Whitney *U* tests were performed, with the null hypothesis that there is no difference between the distribution of observed frequencies and the distribution of frequencies expected from random VJ combination. In all statistical tests, corrected *P* values of less than 0.05 were considered significant.

Study approval. Protocols involving the use of human tissues and animals were approved by the IRB and IACUC of Columbia University, and all of the experiments were performed in accordance with the protocols.

Author contributions

Conceptualization: MKM and MS; methodology: MKM, AO, and MS; software and analysis: AO, KM, HRS, and AM; validation: RW, YS, TMB, and MS; experiments: MKM, AO, MH, AM, GN, ND, HL, SH, and SDW; writing of the original draft of the manuscript: MKM, AO, AM, and MS; review and editing of the manuscript: MKM, AO, MH, RW, YS, HRS, TMB, and MS; funding acquisition: MS; study supervision: MS.

Acknowledgments

This research was supported by the following NIH grants: P01 AI04589716 and R01DK103585 (to MS) and P01 AI42288 and DK106191 (to TMB). Funding was also provided by the NIH Human Islet Research Network (HIRN) Opportunity Pool Fund (RRID:SCR_014393; <https://hirnetwork.org>; U01 DK104162, to MS and TMB). This research was performed using resources and/or funding provided by the NIH HIRN, which is supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) (RRID: SCR_014393; <https://hirnetwork.org>; CMAI UC4 DK104207, to MS, and DK104194, to TMB). Research was performed at the Columbia Center for Translational Immunology (CCTI) Flow Cytometry Core facility, which is supported in part by the Office of the Director of the NIH (S10OD020056, S10RR027050, P30CA013696, 5P30DK063608, and R01DK106436). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. MKM was supported by a Friedman Award from the University of British Columbia (Canada) and an American Diabetes Association (ADA) Postdoctoral Fellowship. We thank Arup Chakraborty (Massachusetts Institute of Technology) and Peter Sims (Columbia University) for helpful comments on the manuscript and Nicole Casio (Columbia University) for assistance with the submission.

Address correspondence to: Megan Sykes, Columbia Center for Translational Immunology, 650 West 168th Street, Black Building, 1512, (Mailbox 127), New York, New York 10032, USA. Phone: 212.304.5696; Email: megan.sykes@columbia.edu.

1. Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P. A direct estimate of the human alpha beta T cell receptor diversity. *Science*. 1999;286(5441):958–961.
2. Robins HS, et al. Comprehensive assessment of

- T-cell receptor beta-chain diversity in alpha beta T cells. *Blood*. 2009;114(19):4099–4107.
3. Qi Q, et al. Diversity and clonal selection in the human T-cell repertoire. *Proc Natl Acad Sci U S A*. 2014;111(36):13139–13144.

4. Vanhanen R, et al. T cell receptor diversity in the human thymus. *Mol Immunol*. 2016;76:116–122.
5. Robins HS, et al. Overlap and effective size of the human CD8 $^{+}$ T cell receptor repertoire. *Sci Transl Med*. 2010;2(47):47ra64.

6. Madi A, 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):1603–1612.
7. Clambey ET, Davenport B, Kappler JW, Marrack P, Homann D. Molecules in medicine mini review: the $\alpha\beta$ T cell receptor. *J Mol Med.* 2014;92(7):735–741.
8. Cabaniols JP, Fazilleau N, Casrouge A, Kourilsky P, Kanellopoulos JM. Most alpha/beta T cell receptor diversity is due to terminal deoxynucleotidyl transferase. *J Exp Med.* 2001;194(9):1385–1390.
9. Klein L, Kyewski B, Allen PM, Hogquist KA. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nat Rev Immunol.* 2014;14(6):377–391.
10. Wieggers GJ, Kaufmann M, Tischner D, Villunger A. Shaping the T-cell repertoire: a matter of life and death. *Immunol Cell Biol.* 2011;89(1):33–39.
11. Allen S, Turner SJ, Bourges D, Gleeson PA, van Driel IR. Shaping the T-cell repertoire in the periphery. *Immunol Cell Biol.* 2011;89(1):60–69.
12. Pham HP, et al. Half of the T-cell repertoire binomial diversity is genetically determined in humans and humanized mice. *Eur J Immunol.* 2012;42(3):760–770.
13. Marodon G, et al. High diversity of the immune repertoire in humanized NOD.SCID.gamma c-/- mice. *Eur J Immunol.* 2009;39(8):2136–2145.
14. Vandekerckhove BA, Bacalla R, Jones D, Kono DH, Theofilopoulos AN, Roncarolo MG. Thymic selection of the human T cell receptor V beta repertoire in SCID-hu mice. *J Exp Med.* 1992;176(6):1619–1624.
15. Shimizu I, Fudaba Y, Shimizu A, Yang YG, Sykes M. Comparison of human T cell repertoire generated in xenogeneic porcine and human thymus grafts. *Transplantation.* 2008;86(4):601–610.
16. Kalscheuer H, et al. A model for personalized in vivo analysis of human immune responsiveness. *Sci Transl Med.* 2012;4(125):125ra30.
17. Mitchell JL, Seng A, Yankee TM, Ikaros, Helios, and Aiolos protein levels increase in human thymocytes after β selection. *Immunol Res.* 2016;64(2):565–575.
18. Morris H, et al. Tracking donor-reactive T cells: Evidence for clonal deletion in tolerant kidney transplant patients. *Sci Transl Med.* 2015;7(272):272ra10.
19. Bogue M, Candéias S, Benoist C, Mathis D. A special repertoire of alpha:beta T cells in neonatal mice. *EMBO J.* 1991;10(12):3647–3654.
20. Jacobsen LM, Posgai A, Seay HR, Haller MJ, Brusko TM. T cell receptor profiling in type 1 diabetes. *Curr Diab Rep.* 2017;17(11):118.
21. Eugster A, et al. High diversity in the TCR repertoire of GAD65 autoantigen-specific human CD4⁺ T cells. *J Immunol.* 2015;194(6):2531–2538.
22. Yang J, et al. Antigen-specific T cell analysis reveals that active immune responses to β cell antigens are focused on a unique set of epitopes. *J Immunol.* 2017;199(1):91–96.
23. Michels AW, et al. Islet-derived CD4 T cells targeting proinsulin in human autoimmune diabetes. *Diabetes.* 2017;66(3):722–734.
24. Bulek AM, et al. Structural basis for the killing of human beta cells by CD8(+) T cells in type 1 diabetes. *Nat Immunol.* 2012;13(3):283–289.
25. Kent SC, et al. Expanded T cells from pancreatic lymph nodes of type 1 diabetic subjects recognize an insulin epitope. *Nature.* 2005;435(7039):224–228.
26. Seay HR, et al. Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes. *JCI Insight.* 2016;1(20):e88242.
27. Greiff V, et al. Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody Repertoires. *J Immunol.* 2017;199(8):2985–2997.
28. Stadinski BD, et al. Hydrophobic CDR3 residues promote the development of self-reactive T cells. *Nat Immunol.* 2016;17(8):946–955.
29. White SH, Wimley WC. Membrane protein folding and stability: physical principles. *Annu Rev Biophys Biomol Struct.* 1999;28:319–365.
30. Haynes BF, Heinly CS. Early human T cell development: analysis of the human thymus at the time of initial entry of hematopoietic stem cells into the fetal thymic microenvironment. *J Exp Med.* 1995;181(4):1445–1458.
31. Deibel MR, Riley LK, Coleman MS, Cibull ML, Fuller SA, Todd E. Expression of terminal deoxynucleotidyl transferase in human thymus during ontogeny and development. *J Immunol.* 1983;131(1):195–200.
32. Redondo MJ, et al. Heterogeneity of type 1 diabetes: analysis of monozygotic twins in Great Britain and the United States. *Diabetologia.* 2001;44(3):354–362.
33. Zvyagin IV, et al. Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing. *Proc Natl Acad Sci U S A.* 2014;111(16):5980–5985.
34. Qi Q, et al. Diversification of the antigen-specific T cell receptor repertoire after varicella zoster vaccination. *Sci Transl Med.* 2016;8(332):332ra46.
35. Wang C, et al. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc Natl Acad Sci U S A.* 2010;107(4):1518–1523.
36. Petrova G, Ferrante A, Gorski J. Cross-reactivity of T cells and its role in the immune system. *Crit Rev Immunol.* 2012;32(4):349–372.
37. Reiser JB, et al. CDR3 loop flexibility contributes to the degeneracy of TCR recognition. *Nat Immunol.* 2003;4(3):241–247.
38. Sewell AK. Why must T cells be cross-reactive? *Nat Rev Immunol.* 2012;12(9):669–677.
39. Wooldridge L, et al. A single autoimmune T cell receptor recognizes more than a million different peptides. *J Biol Chem.* 2012;287(2):1168–1177.
40. Cole DK, et al. Hotspot autoimmune T cell receptor binding underlies pathogen and insulin peptide cross-reactivity. *J Clin Invest.* 2016;126(6):2191–2204.
41. McDonald BD, Bunker JJ, Erickson SA, Oh-Hora M, Bendelac A. Crossreactive $\alpha\beta$ T cell receptors are the predominant targets of thymocyte negative selection. *Immunity.* 2015;43(5):859–869.
42. Chen X, Poncette L, Blankenstein T. Human TCR-MHC coevolution after divergence from mice includes increased nontemplate-encoded CDR3 diversity. *J Exp Med.* 2017;214(11):3417–3433.
43. Nishio J, Suzuki M, Nanki T, Miyasaka N, Kohsaka H. Development of TCRB CDR3 length repertoire of human T lymphocytes. *Int Immunol.* 2004;16(3):423–431.
44. Yassai M, Gorski J. Thymocyte maturation: selection for in-frame TCR alpha-chain rearrangement is followed by selection for shorter TCR beta-chain complementarity-determining region 3. *J Immunol.* 2000;165(7):3706–3712.
45. Matsutani T, et al. Shortening of complementarity determining region 3 of the T cell receptor α chain during thymocyte development. *Mol Immunol.* 2011;48(4):623–629.
46. Matsutani T, et al. Comparison of CDR3 length among thymocyte subpopulations: impacts of MHC and BV segment on the CDR3 shortening. *Mol Immunol.* 2007;44(9):2378–2387.
47. Gavin MA, Bevan MJ. Increased peptide promiscuity provides a rationale for the lack of N regions in the neonatal T cell repertoire. *Immunity.* 1995;3(6):793–800.
48. Kitaura K, Shini T, Matsutani T, Suzuki R. A new high-throughput sequencing method for determining diversity and similarity of T cell receptor (TCR) α and β repertoires and identifying potential new invariant TCR α chains. *BMC Immunol.* 2016;17(1):38.
49. Adams JJ, et al. T cell receptor signaling is limited by docking geometry to peptide-major histocompatibility complex. *Immunity.* 2011;35(5):681–693.
50. Moss PA, Bell JL. Sequence analysis of the human alpha beta T-cell receptor CDR3 region. *Immunogenetics.* 1995;42(1):10–18.
51. Hawes GE, Struyk L, van den Elsen PJ. Differential usage of T cell receptor V gene segments in CD4⁺ and CD8⁺ subsets of T lymphocytes in monozygotic twins. *J Immunol.* 1993;150(5):2033–2045.
52. Malhotra U, Spielman R, Concannon P. Variability in T cell receptor V beta gene usage in human peripheral blood lymphocytes. Studies of identical twins, siblings, and insulin-dependent diabetes mellitus patients. *J Immunol.* 1992;149(5):1802–1808.
53. Kohsaka H, Taniguchi A, Chen PP, Ollier WE, Carson DA. The expressed T cell receptor V gene repertoire of rheumatoid arthritis monozygotic twins: rapid analysis by anchored polymerase chain reaction and enzyme-linked immunosorbent assay. *Eur J Immunol.* 1993;23(8):1895–1901.
54. Reinhardt C, Melms A. Skewed TCRV beta repertoire in human thymus persists after thymic emigration: influence of genomic imposition, thymic maturation and environmental challenge on human TCRV beta usage in vivo. *Immunobiology.* 1998;199(1):74–86.
55. Sharon E, Sibener LV, Battle A, Fraser HB, Garcia KC, Pritchard JK. Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nat Genet.* 2016;48(9):995–1002.
56. Rubelt F, et al. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naive and antigen-experienced cells. *Nat Commun.* 2016;7:11112.
57. Kjer-Nielsen L, et al. A structural basis for the selection of dominant T cell receptors in antiviral immunity. *Immunity.* 2003;18(1):53–64.
58. Hamrouni A, Aublin A, Guillaume P, Maryanski JL. T cell receptor gene rearrangement lineage analysis reveals clues for the origin of highly restricted antigen-specific repertoires. *J Exp Med.*

- 2003;197(5):601-614.
59. Dudley EC, Petrie HT, Shah LM, Owen MJ, Hayday AC. T cell receptor beta chain gene rearrangement and selection during thymocyte development in adult mice. *Immunity*. 1994;1(2):83-93.
60. Bartok I, Holland SJ, Kessels HW, Silk JD, Alkhinji M, Dyson J. T cell receptor CDR3 loops influence alphabeta pairing. *Mol Immunol*. 2010;47(7-8):1613-1618.
61. Nanki T, Kohsaka H, Mizushima N, Ollier WE, Carson DA, Miyasaka N. Genetic control of T cell receptor BJ gene expression in peripheral lymphocytes of normal and rheumatoid arthritis monozygotic twins. *J Clin Invest*. 1996;98(7):1594-1601.
62. DeWolf S, et al. Quantifying size and diversity of the human T cell alloresponse. *JCI Insight*. 2018;3(15):e.121256.