Determinants governing T cell receptor α/β-chain pairing in repertoire formation of identical twins

Hidetaka Tanno[a,b,1], Timothy M. Gould[c,d,1], Jonathan R. McDaniel[a], Wenqiang Cao[c,d], Yuri Tanno[a], Russell E. Durrett[a], Daechan Park[e], Steven J. Cate[f], William H. Hildebrand[f], Cornelia L. Dekker[g], Lu Tian[h], Cornelia M. Weyand[c,d], George Georgiou[a,b,2,3], and Jörg J. Goronzy[c,d,2,3]

[a]Department of Chemical Engineering, University of Texas at Austin, Austin, TX 78712; [b]Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, TX 78712; [c]Division of Immunology and Rheumatology, Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305; [d]Department of Medicine, Palo Alto Veterans Administration Healthcare System, Palo Alto, CA 94304; [e]Department of Life Sciences, Ajou University, Suwon 16499, South Korea; [f]Department of Microbiology and Immunology, University of Oklahoma Health Sciences Center, Oklahoma City, OK 73104; [g]Department of Pediatrics (Infectious Diseases), Stanford University School of Medicine, Stanford, CA 94305; and [h]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305

Edited by Philippa Marrack, National Jewish Health, Denver, CO, and approved December 4, 2019 (received for review August 30, 2019)

The T cell repertoire in each individual includes T cell receptors (TCRs) of enormous sequence diversity through the pairing of diverse TCR α- and β-chains, each generated by somatic recombination of paralogous gene segments. Whether the TCR repertoire contributes to susceptibility to infectious or autoimmune diseases in concert with disease-associated major histocompatibility complex (MHC) polymorphisms is unknown. Due to a lack in high-throughput technologies to sequence TCR α–β pairs, current studies on whether the TCR repertoire is shaped by host genetics have so far relied only on single-chain analysis. Using a high-throughput single T cell sequencing technology, we obtained the largest paired TCRαβ dataset so far, comprising 965,523 clonotypes from 15 healthy individuals including 6 monozygotic twin pairs. Public TCR α- and, to a lesser extent, TCR β-chain sequences were common in all individuals. In contrast, sharing of entirely identical TCRαβ amino acid sequences was very infrequent in unrelated individuals, but highly increased in twins, in particular in CD4 memory T cells. Based on nucleotide sequence identity, a subset of these shared clonotypes appeared to be the progeny of T cells that had been generated during fetal development and had persisted for more than 50 y. Additional shared TCRαβ in twins were encoded by different nucleotide sequences, implying that genetic determinants impose structural constraints on thymic selection that favor the selection of TCR α–β pairs with entire sequence identities.

T cell repertoire | single-cell sequencing | monozygotic twins | major histocompatibility complex

The T cell repertoire in each individual includes T cell receptors (TCRs) of enormous sequence diversity to be able to recognize a huge variety of peptides displayed by major histocompatibility complex (MHC) molecules. TCRs are heterodimers, composed for most T cells of an α- and a β-chain with each T cell clone expressing a unique combination. Diversity of both TCR chains is generated by somatic recombination of paralogous variable (V), joining (J), and, in the case of the *TRB* gene, diversity (D) gene segments, with additional removal and random addition of nucleotides at the joints (1, 2). In the thymus, TCR β-chains are rearranged first (3); since T cell precursors proliferate until they recombine the *TRA* gene, one single TCR β-chain has been estimated to dimerize with as many as 25 different α-chains to form the peripheral repertoire, emphasizing the contribution of α–β pairing to diversity (4, 5). The resulting theoretical TCR diversity has been proposed to be larger than 10^{15} different TCR α–β dimers (6), and therefore by far exceeds the total number of T cells in an individual, which is $<10^{12}$ (7, 8). However, the realized repertoire is not just the outcome of stochastic selection of all possible TCRs, but instead it is shaped by several mechanisms. The recombination machinery is biased, resulting in the preferred usage of certain V and J elements. How much structural constraints in TCR α–β pairing reduce the number of possible functional TCRs is currently unknown. Pronounced amino acid diversity in each of the 2 TCR chains implicates a very large and complex set of possible protein–protein interactions between them. Putative constraints in TCR α–β dimerization would therefore not be surprising considering the heterogeneity of such protein–protein interactions. Finally, successfully formed TCRs are selected for their binding to HLA peptide complexes. The TCR repertoire is selected in the thymus for TCRs that have a low affinity for self but are functional in MHC peptide recognition (9). Only a small fraction of T cells survives this selection process (10), raising the possibility that structural constraints in TCR α–β pairing limits the functional repertoire.

TCR repertoire analyses so far have relied on determining the number of unique *TRB* sequences in peripheral blood specimens, therefore not including the contribution of TCR α–β pairing. Estimates of true complexity are further problematic, because sizes of individual clonotypes are nonuniformly distributed and measurements in small peripheral blood samples cannot be easily extrapolated to infrequent clonotypes (11, 12). In young

Significance

T cell clones with completely identical T cell receptor (TCR) α–β pairs are frequent in twins, in part due to long-term survival of intrauterine generated T cells, in part due to genetic constraints on the pairing of identical TCR α- and β-chains in thymic selection. In concert with risk-associated major histocompatibility complex (MHC) genes, clonal T cell selection may therefore contribute to disease susceptibility. The reported large dataset on paired human TCRαβ sequences will be an important resource for the scientific community interested in the analysis of the contribution of TCR α–β pairing in human TCR repertoire formation.

Author contributions: C.L.D., C.M.W., G.G., and J.J.G. designed research; H.T., T.M.G., J.R.M., W.C., Y.T., and R.E.D. performed research; D.P., S.J.C., and W.H.H. contributed new reagents/analytic tools; H.T., T.M.G., L.T., G.G., and J.J.G. analyzed data; and H.T., T.M.G., L.T., G.G., and J.J.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Data deposition: Sequences reported in this paper have been deposited in the National Center for Biotechnology Information Sequence Read Archive (accession code PRJNA593622).

[1]H.T. and T.M.G. contributed equally to this work.

[2]G.G. and J.J.G. contributed equally to this work.

[3]To whom correspondence may be addressed. Email: jgoronzy@stanford.edu or gg@che.utexas.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1915008117/-/DCSupplemental.

First published December 26, 2019.

adults, the *TRB* repertoire has been estimated to include close to 10^8 unique nucleotide sequences for naive CD4 and CD8 cells, $1–2 \times 10^6$ for CD4 memory cells, and less than 5×10^5 for CD8 memory cells (13). These data are consistent with the notion that the existing repertoire in each individual is by far smaller than the potential diversity, stressing the importance of selection, which appears to influence susceptibility to autoimmune diseases. Proposed mechanisms include the presence of hydrophobic residues in the peptide-binding complementary determining region 3 (CDR3) of self-reactive TCRs (14), or the selection of TCRs that recognize peptides presented by disease-associated MHC molecules (15).

Studies on whether the human TCR repertoire is shaped by host genetics so far have also only relied on single-chain analysis. Zvyagin et al. (16) compared out-of-frame and in-frame *TRA* and *TRB* sequences of T cells from monozygotic twin pairs and found genetic influence on the rearrangement frequencies of AV, BV, and BJ gene segments independent of the expression of functional chains. A genetic bias was also found for AJ gene segments, but only for expressed *TRA* genes, suggesting a genetic influence on thymic selection. In further support for the latter interpretation, associations between variation in the MHC locus and TCR V gene usage were described by Sharon et al. (17), who applied expression quantitative trait locus mapping to test for transassociations. Obviously, studies on TCR α–β pairs are needed to further characterize the effect of genetic influence because thymic selection depends on the entire TCR dimer.

In the absence of structural constraints, the magnitude of TCR diversity would imply that sharing of fully identical TCR α–β sequence pairs between individuals should be very rare since the maximum theoretical TCR generation probability is less than 10^{-12} (18). In contrast, studies found unexpectedly high rates of TCR single-chain sequence sharing between individuals. Such public TCR sequences may be of particular interest because they are frequently found in mouse models of autoimmune and infectious diseases (19, 20). Initial human studies defined TCR sequence sharing on the basis of partial sequence identity, typically focusing on antigen-binding CDR3 regions (21, 22). In subsequent work on monozygotic twins, sharing of entire TCR β-chain amino acid sequence was reported to be surprisingly frequent, occurring at a normalized rate of 10^{-7}; moreover, these studies indicated a genetic effect increasing the probability of TCRβ sharing (16, 23, 24). It remains to be determined whether this β-chain sharing reflects clonal identity, i.e., whether the entire TCRαβ dimer is shared.

Prior approaches to study TCR dimers have used single-cell sorting followed by RT-PCR (25), a combinatorial pairing approach [PairSEQ (26)], or the use of a microfluidic device to barcode *TRA* and *TRB* cDNAs from individual cells in emulsion droplets (27). Single-cell RT-PCR in microtiter well plates is low throughput and not suitable for analyzing the repertoire at the requisite depth for delineating rare events such as the prevalence of shared sequences. On the other hand, combinatorial approaches can only detect expanded clones, and therefore it is not suitable for determining the TCRαβ repertoire in T cell populations with very high diversity such as naive T cells. Using a microfluidic single-cell barcoding technique (27), Atwal, Vigneault, and coworkers (28, 29) obtained $\sim 2 \times 10^5$ TCRαβ sequences; their statistical analysis of this dataset led the authors to conclude that TCR α- and β-chain pairing is not stochastic and that α–β pairing is informative of CD4 vs. CD8 T cell lineage development, consistent with the MHC class II vs. class I restriction favoring different V elements.

Here, we adapted a technique developed by one of our laboratories for sequencing the V_H:V_L antibody repertoire (30, 31) to enable the sequencing of the TCRαβ repertoire at very high throughput and with a pairing accuracy of >92%. We used this technique to determine the paired TCR repertoire of 15 healthy individuals including 6 pairs of monozygotic twins and recovered on average tens-of-thousands of unique TCRαβ sequences per sample, to generate by far the largest dataset available, comprising nearly 1 million high-confidence TCRαβ clonotypes. Remarkably, we find only little evidence for α–β pairing restrictions and show that the frequency of α–β pairs is mostly a stochastic product of individual germline encoded V and J gene segment usage. Nonetheless, our data also suggest the possibility of very minor and subtle structural constraints in TCR α/β-chain pairing or a bias in thymic selection favoring combinations of CDR1/2 α and β polymorphic determinants. In our twin studies, we demonstrate a genetic influence on sharing of identical TCR sequences between individuals. This genetic effect was strikingly more pronounced for sharing of TCRαβ sequences compared to that for α or β sequences alone, suggesting presence of identical clonotypes in twin pairs.

Results and Discussion

High-Throughput TCR α–β Sequencing with High Pairing Accuracy. We previously developed a method for encapsulating single B cells in monodisperse water-in-oil droplets formed by using a custom-made flow focusing apparatus. Cells are lysed within droplets and mRNA is captured on oligo-dT beads and then a second emulsion OE-RT-PCR is used to link V_H and V_L cDNAs into a single amplicon that can be sequenced by long-read high-throughput sequencing (30, 31). We modified this experimental pipeline by employing previously reported TCR amplification multiplex primers (28) together with OE-PCR primers required to generate 550-bp TCR α–β amplicons that could be sequenced using Illumina MiSeq 2×300 technology. For data analysis, we developed a bioinformatics pipeline for sequencing quality control, gene annotation, and clustering. To evaluate the pairing accuracy that is achieved using this technology, we first expanded total T cells by stimulating them with anti-CD3/CD28 beads plus IL-2, and then divided them into 2 technical replicates. The 2 replicates yielded a similar number of clonotypes (9,658 and 11,689). A pairing precision of >92% was calculated on the basis that *TRB* sequences (BV-CDR3-BJ) were detected in both replicates, which were linked with the same *TRA* sequence (AV-CDR3-AJ). We then determined the TCRαβ repertoire from naive CD4, memory CD4, and total T cells taken from 15 HLA-typed healthy volunteers (*SI Appendix*, Table S1). We obtained on average ~35,000 distinct TCRαβ clonotypes per sample for a total of 965,523 clonotypes, by sequencing each sample to a depth of about 1 million reads (*SI Appendix*, Table S2).

Lack of Evidence for Major Structural Constraints in TCR α–β Chain Pairing Due to V and J Gene Segment Polymorphisms. We first aligned and aggregated sequences to their corresponding V and J gene segments (e.g., *TRAV* with *TRBV*, *TRAJ* with *TRBJ*), and then determined the total number and respective frequencies of all gene segment pairs. Representative heatmaps show the total number of gene segments within a sample of naive and memory CD4 T cells taken from the same individual (Fig. 1*A*). As previously reported, gene segment usage is highly biased with some gene segments observed very frequently irrespective of the segment(s) to which they pair, e.g., *TRAV*26–1 and *TRBV*20–1 (21, 32, 33). Consequently, certain pairs such as *TRAV*26–1:*TRBV*20–1 (total of 296 in naive and 619 in memory CD4 T cells, respectively, in the sample shown) were highly prevalent in the repertoire. Normalization for absolute usage frequencies eliminated most of the appearance of biased AV–BV pairing (right heatmaps in Fig. 1*A*). Similar results were observed for naive and memory CD4 T cells. Variations in pairing frequencies in the normalized heatmap cell intensities, if present, were mostly with infrequent gene segments and therefore not statistically robust.

To quantitatively explore whether V and J gene segment polymorphisms impose minor structural constraints in TCR α–β

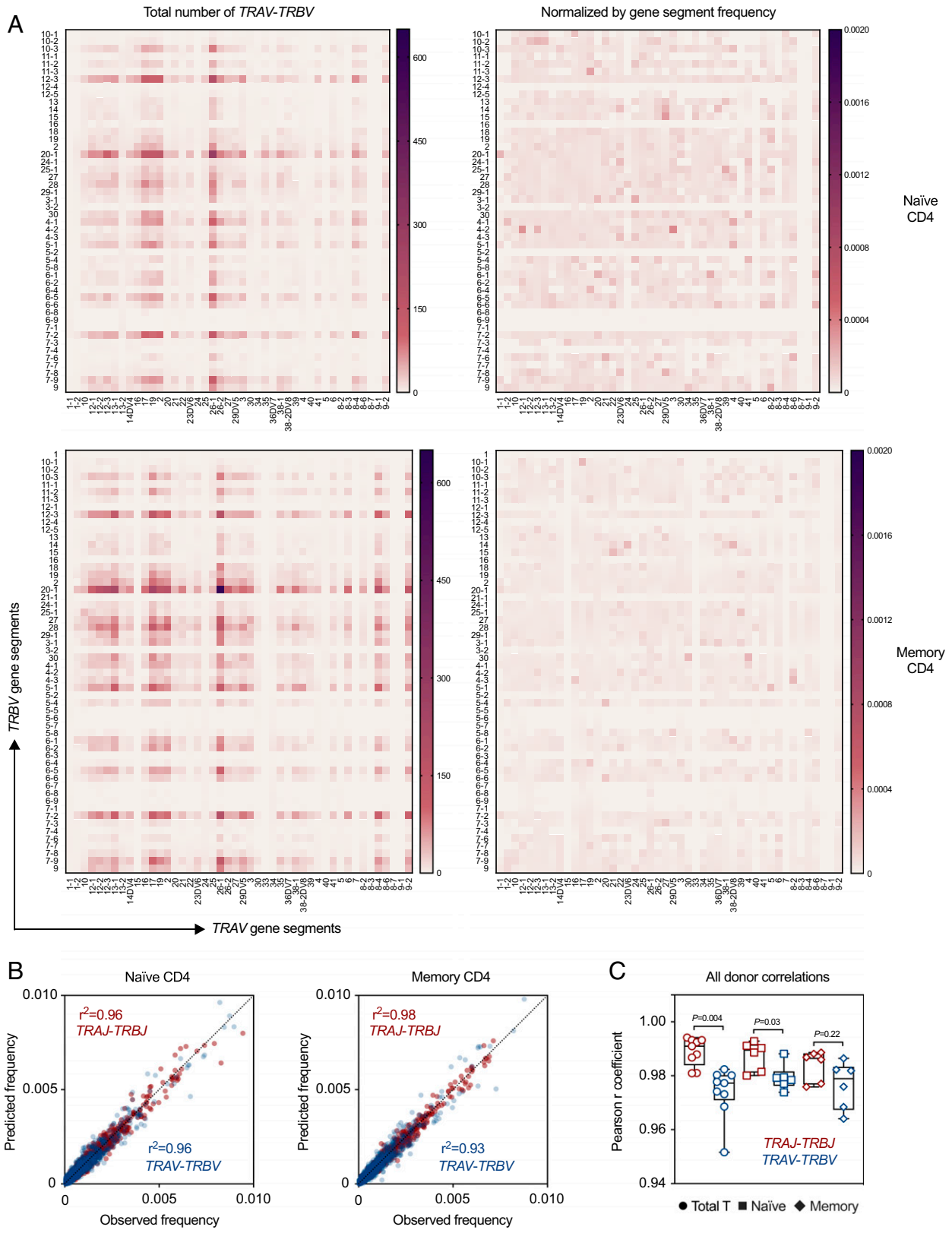

Fig. 1. Pairing of V and J elements of TCR α- and β-chains is largely stochastic. (*A*) Heatmaps show the number of *TRAV–TRBV* sequence combinations detected in naive (*Top*) and memory CD4 (*Bottom*) T cells from one representative individual before (*Left*) and after (*Right*) normalization for gene segment frequencies. (*B*) Observed frequencies of TCR α–β pairs were correlated to those predicted based on the product of the respective gene segment frequencies. Scatter plots show V (blue) and J (red) segment frequencies for naive (*Left*) and memory (*Right*) CD4 subsets from one representative individual. Correlation coefficients are from fitting to an identity function without linear regression. (*C*) Summary box plots of Pearson correlation coefficients for all individuals as shown in Fig. 1*B*. Different cell populations are indicated as circles (total T), squares (naive CD4), or diamonds (memory CD4), with variable (blue) and joining (red) regions shown separately. Two-sided *P* values were determined by one-sample Wilcoxon test.

pairing, we correlated frequencies of observed pairing to those predicted based on gene segment usage. Correlation plots of V and J regions within naive and memory CD4 T cells from one representative donor are shown in Fig. 1*B*; Pearson correlation coefficients across all donors and cell populations examined are summarized in Fig. 1*C*. Coefficients were generally higher than 0.94, suggesting the lack of a major structural constraint in pairing. However, when we compared the observed r^2 to distributions of r^2 values generated from a model in which we assume independent pairing, the observed r^2 for V region pairs were below the 5% percentile expected for independent pairing for all naive and memory samples analyzed, indicating the presence of a slight bias (*SI Appendix*, Fig. S1*B*). Interestingly, a single combination, BV20-1 with AV26-1, accounted for most of this bias in all individuals, both for naive and for memory CD4 T cells. Correlations of observed J region pairing relative to those predicted based on gene segment usages were slightly stronger than those for V segments, indicated in Fig. 1*C* (total T, $P = 0.004$; naive CD4, $P = 0.03$; memory CD4, $P = 0.22$). Moreover, observed correlation coefficients were as expected (i.e., largely within the standard distribution range of simulated correlation coefficients) for independent J pairing, indicating a lack in structural constraints (*SI Appendix*, Fig. S1*B*). We also examined the correlation between V regions on one chain and J regions on the other (*SI Appendix*, Fig. S1*E*); for all T cell populations, we found that V–J correlation coefficients were greater than 0.95, indicating that such associations are also highly random. Grigaityte et al. (29) recently postulated a bias in AV and BV gene segment pairing based on mutual information analysis of 2×10^5 paired sequences from 5 donors. Their mutual information estimates were larger than zero (with zero indicating independent pairing), but the deviations were small and may have been overestimated in 2 individuals due to the inclusion of sequences from clonally expanded cells. Of note, within each population, we only included unique sequences in our analysis. We performed a mutual information analysis on our data and included a 95% confidence interval using a bootstrapped mutual information estimator (*SI Appendix*, Fig. S1*D* and Table S3). Consistently, the mutual information values were increased for AV–BV, but not for AJ–BJ combinations, indicating a bias in V gene segment pairing. In addition to a structural constraint in pairing, this bias in V gene segment pairing could also be caused by thymic selection. Structural studies of human and mouse TCRs have mapped the variable regions CDR1 and CDR2 encoded by the AV and BV germline gene segments to the region contacting the MHC (34–36). CDR1α and CDR1β frequently also contribute to binding of amino- and carboxyl-terminal peptide segments, respectively. The hypervariable CDR3α and CDR3β loops of the TCR, to which the J region contributes, are placed over the center of the bound peptide. This central position of the J segments in

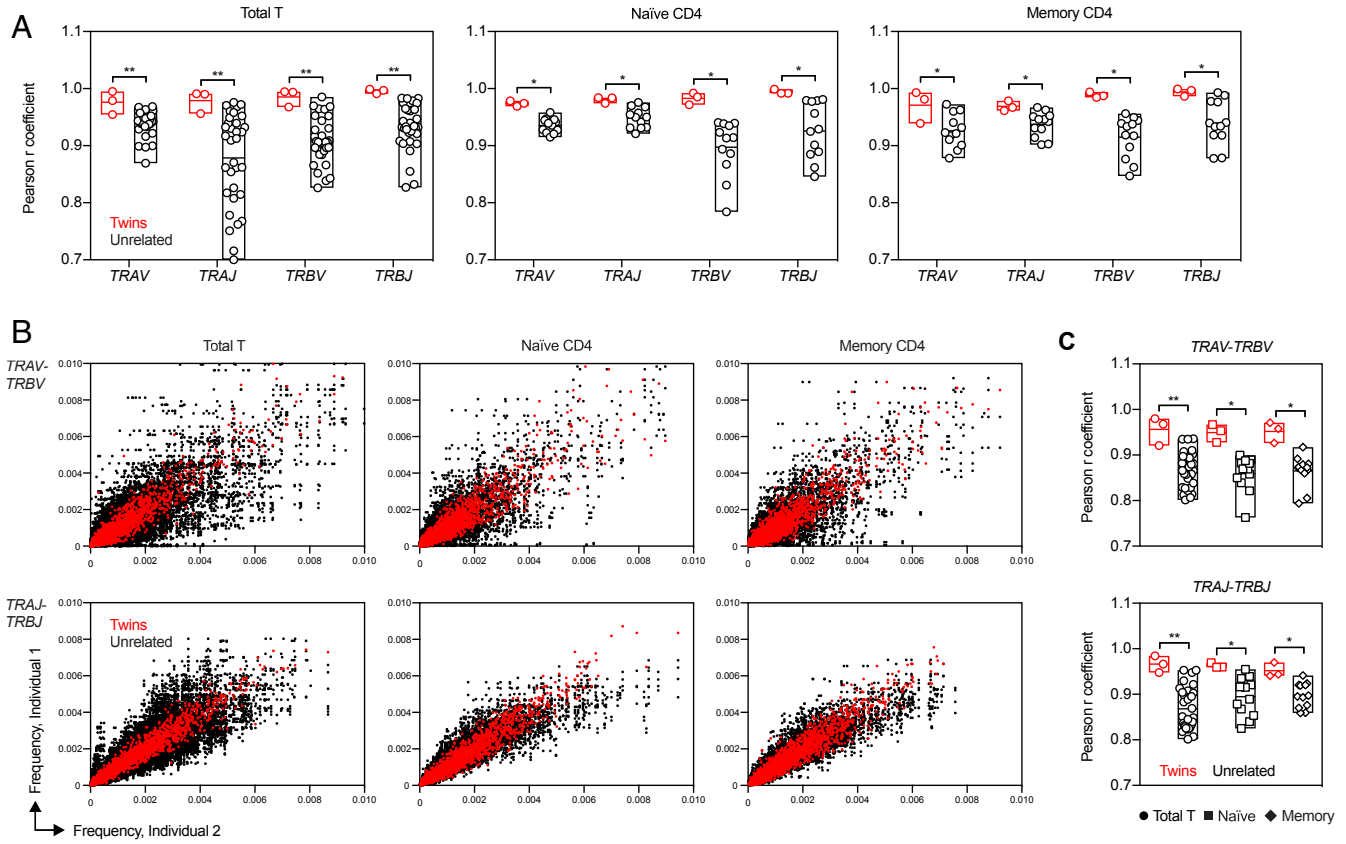

Fig. 2. Genetic influence on frequencies of TCRαβ pairs is due to biased gene segment usage. (*A*) Gene segment usage of *TRAV*, *TRAJ*, *TRBV*, and *TRBJ* elements were compared in twin pairs (red) and unrelated individuals (black). Pearson correlation coefficients for all possible comparisons are shown as box plots for total T (*Left*), naive CD4 (*Middle*), and memory CD4 (*Right*). One-sided *P* values were determined by permutation test (*$P = 0.067$; **$P = 0.002$). (*B*) Frequencies of *TRAV–TRBV* (*Top*) and *TRAJ–TRBJ* (*Bottom*) gene segment combinations were compared between twins and unrelated individuals. Scatter plots show correlations in total T (*Left*), naive CD4 (*Middle*), and memory CD4 (*Right*) cells. For each cell population, all possible combinations between 2 individuals are shown comparing twin pairs (red dots) or unrelated pairs (black dots). (*C*) Summary box plots show Pearson *r* coefficients calculated for each possible pair of individuals comparing twin pairs (red) vs. unrelated pairs (black). Different cell populations are indicated as circles (total T), squares (naive CD4), or diamonds (memory CD4); *Top* shows correlation coefficients for variable regions comparisons, and *Bottom* for joining regions. One-sided *P* values were determined by permutation test (*$P = 0.067$; **$P = 0.002$).

the TCR was not associated with any structural constraints imposed by J segment polymorphisms.

Genetic Influence on the TCR α/β-Chain Repertoire Is Driven by Biased Gene Segment Usage. To explore the genetic influence on the TCRαβ repertoire, we included 6 monozygous twin pairs in our study population. We sequenced total T cells from 3 twin pairs and purified naive and memory CD4 T cells from 3 additional pairs. Previous studies have shown a genetic influence on the usage of V, and potentially to a lesser extent J, gene segments (16, 37). We correlated individual gene segment frequencies for all possible interindividual combinations and cell populations and found significantly higher correlation coefficients among twins compared to unrelated individuals for all TCR gene segments (Fig. 2*A*). This genetic influence was seen in naive as well as in memory cells. If recombination is driving the bias, then we should not see a genetic impact on TCR α–β pairing beyond that for gene segment usage. This was indeed the case. We correlated frequencies of gene segment pairs for both variable and joining regions (i.e., *TRAV*:*TRBV*, *TRAJ*:*TRBJ*) for all interindividual combinations (Fig. 2*B*). While correlation coefficients of gene segment pair frequencies were greater in twins for each cell population (Fig. 2*C*), such differences were of about the same magnitude as they were for the case of individual gene segment usages (Fig. 2*B*), thus suggesting that there was no major additional genetic impact on pairing. (Note that significance levels were determined by permutation test and *P* values are bottomed due to the relatively small twin sample size, e.g., the case of $n = 6$ individuals comprising 3 twin pairs results in 30 possible permutations and thus the smallest one-sided *P* value is 1/15 or 0.067.)

Increased Sharing of Identical TCR α–β Pairs in Twins. To further examine the genetic influence on the TCR repertoire formation, we analyzed the TCR chain sequences that are shared among donors at the amino acid level. The reidentification of clonotypes in a second sample depends on the overall TCR diversity and is therefore limited by the number of cells sequenced and the frequency of a given sequence within a sample (38). To determine

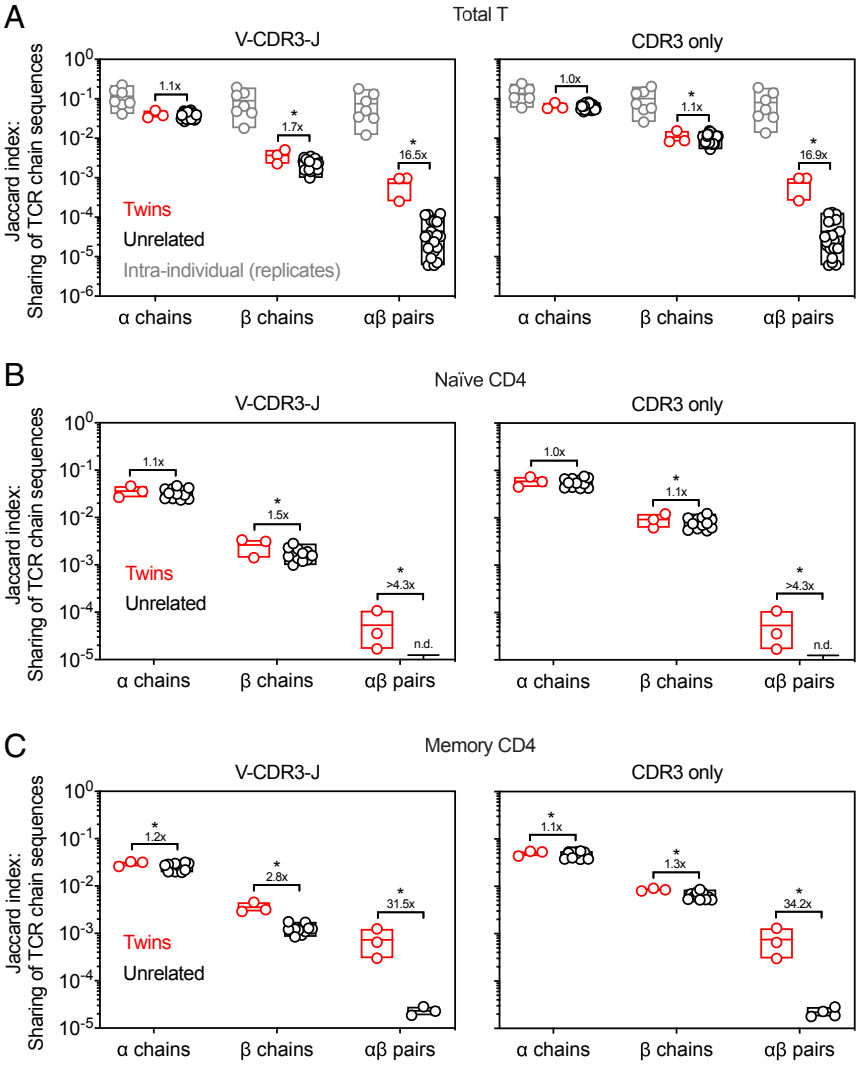

Fig. 3. Increased sharing of identical TCRαβ pairs in twins. Jaccard indices were computed for sharing of identical TCR α- and TCR β-chain sequences between twin pairs (red) or unrelated individuals (black) as well as for TCRαβ pairs. As lower boundary of detecting identical sequences, sharing in replicate samples (gray) is shown. Sharing across entire variable and joining regions (V-CDR3-J, *Left*) and sharing within isolated CDR3 regions (*Right*) are shown. Box plots show results from total T (*A*), naive CD4 (*B*), and memory CD4 (*C*) cells; if no identical sequences were found (TCRαβ pairs in *B*), the detection limit is shown. Fold differences between twins and unrelated individuals average sharing of identical sequences are indicated. One-sided *P* values were determined by permutation test (*$P = 0.067$).

whether our sample size of 30,000 to 40,000 sequences out of about 1 million cells is sufficient to detect identical sequences and to develop an upper boundary of detection, we first performed experiments on replicate total T cell samples taken from the same individual (Fig. 3*A*). The rate of reidentifying sequences in replicate samples was similar for TCR α or β single chains and for TCRαβ sequence pairs and was on average ∼10 to 20%. As previously described, the TCR repertoire is extremely diverse and reidentification is therefore biased toward clonally expanded T cells. While quantitative predictions on the sensitivity of TCRαβ repertoire analyses are not possible in the absence of information on the clonal size distributions, rates of sequence sharing observed between samples from different individuals must be clear underestimates.

Next, we analyzed sharing of TCR chain sequences comparing twins and unrelated individuals. Results are shown as Jaccard indices (Fig. 3) as well as normalized sharing rates (*SI Appendix*, Fig. S2*A*). In addition to *P* values that plateaued again at 0.067, we provide the fold differences in twins vs. unrelated sharing rates as a metric to highlight the extent of the observed difference. Sharing of identical TCR α-chain sequences in total T cells was high among all individuals irrespective of twin status with ∼10% of the α-chain repertoire being shared (Fig. 3*A*). Sharing of identical β-chain sequences was roughly an order of magnitude less frequent than that of α-chains and slightly higher in twin pairs, consistent with the previously reported normalized frequency of 10^{-7} derived from isolated *TRB* sequencing (24). Sharing of identical TCRαβ amino acid sequences was extremely low in unrelated individuals, but markedly higher in twin pairs. Read counts shown for shared TCRαβ sequences were higher (*SI Appendix*, Fig. S2*B*), indicating that reidentification was biased toward larger clonal sizes and the observed frequency is a minimal estimate. Interestingly, we observed almost identical sharing frequencies for full TCRαβ sequences (V-CDR3-J) as for CDR3 regions alone, i.e., sharing of isolated identical CDR3 combinations was infrequent.

Previous results on *TRB* sequencing of antigen-specific T cells have reported sharing frequencies as high as 10^{-5} for isolated VZV-specific CD4 and yellow fever-specific CD8 memory T cells (23, 24). Since total T cells include diverse naive and more oligoclonal memory T cells, we analyzed TCR α–β pairing in purified naive and memory CD4 T cell subsets (Fig. 3 *B* and *C*). Identical TCRαβ clones were detected in naive T cells from twins but were not detected in unrelated individuals. Identical CD4 memory T cell clonotypes were detectable at a low level in unrelated individuals and were more than 30-fold enriched in twin pairs.

Fewer N-Region Addition and Higher MHC Similarity Drive Sharing of Identical TCR α–β Sequences. Public TCR sequences have been mostly attributed to recombinatorial bias in the generation of single TCR α- or β-chains, also termed convergent recombination (39); however, the processes underlying shared, identical TCRαβ pairs as found here have not been studied. To identify mechanisms driving TCR αβ-chain sequence sharing, we first examined whether shared clonotypes represent public sequences. Identical TCRα

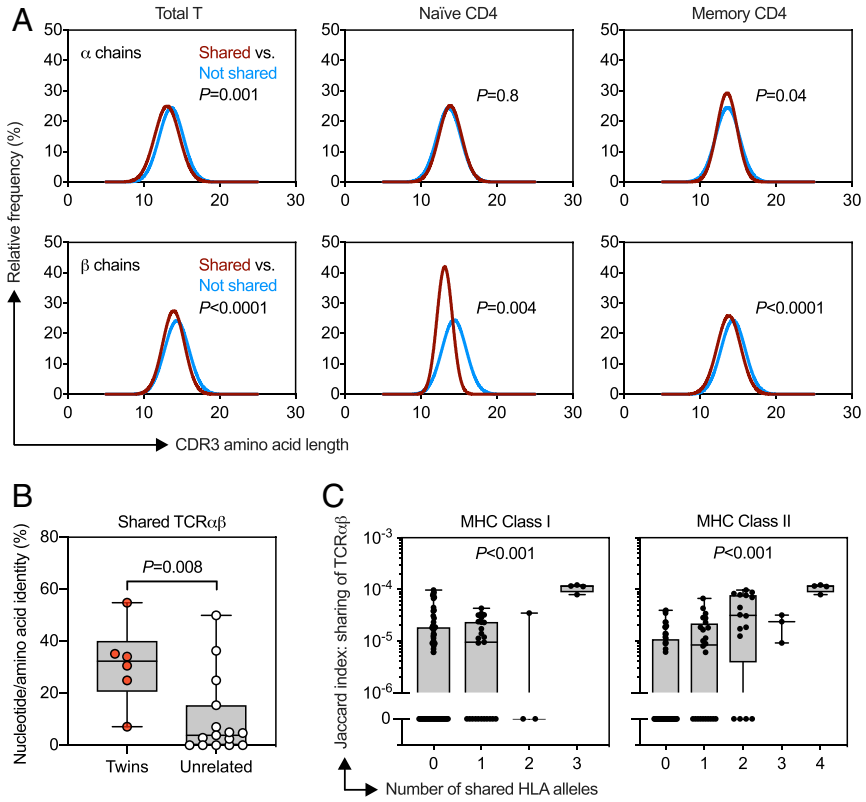

Fig. 4. Mechanisms driving sharing of identical TCRαβ sequences. (*A*) TCR α- and β-chain amino acid CDR3 lengths were analyzed independently comparing TCRαβ sequences shared vs. those not shared between twin–twin pairs. Data are shown as Gaussian distributed histograms. *Top* shows the distribution of α-chain CDR3 lengths in total T (*Left*), naive CD4 (*Middle*), and memory CD4 (*Right*) cells; *Bottom* shows those for β-chains. Two-sided *P* values were determined by unpaired *t* test. (*B*) Box plots show percentage of identical nucleotide TCRαβ sequences in the shared TCRαβ amino acid clonotypes among twins and unrelated individuals across all cell types; *P* value was determined by Mann–Whitney *U* test. (*C*) Jaccard indices were computed for the sharing of identical sequences in combined total T and memory CD4 samples between any 2 unrelated individuals. Results are plotted vs. the number of shared MHC class I (*Left*) and class II (*Right*) alleles; *P* values were determined by trend test.

sequences have been identified for natural killer T (NKT) cells and for mucosa-associated invariant T (MAIT) cells (40, 41). Only one clonotype with paired identical α- and β-chains identified in our analysis of twins displayed a sequence characteristic of NKT cells. TCRs expressing AV–AJ combinations pertinent for MAIT cells made up about 10 to 15% of all shared clonotypes in twins, but no shared TCRαβ sequences characteristic of MAIT cells were observed among unrelated individuals, suggesting that MAIT cells make only a minor contribution to the shared TCRαβ repertoire.

We examined in more detail the CDR3 amino acid sequences of the shared TCR αβ-chains in twin pairs. The CDR3α sequences were about of the same average length, while CDR3β regions of shared TCRαβ clones were nearly 1 amino acid shorter compared to that of nonshared (Fig. 4*A*). These observations were consistent for all 3 comparisons, i.e., for naive CD4, memory CD4, and total T cells. This finding may reflect that shorter sequences have fewer nucleotide additions at nongermline encoded positions, thus resulting in a higher probability of sequence overlap. We noted that the majority of amino acid CDR3β sequences of the shared TCRαβ (i.e., identical amino acid sequences detected in 2 or more unrelated donors) were different at the nucleotide sequence level; a representative example is shown in *SI Appendix*, Fig. S2*C*. Shared CDR3β sequences that had different nucleotide sequence differed by 2 to 2.5 nucleotide substitutions on average in all individuals and T cell subsets. However, identical CDR3α–CDR3β nucleotide sequences were found at a significantly higher frequency in twins compared to unrelated individuals ($P = 0.008$), with about one-third of shared sequences in twins identical at the nucleotide level (Fig. 4*B*); in contrast, complete nucleotide sequence identity was absent or very infrequent (median <10%) among unrelated individuals. The increased frequency of nucleotide sequence identity indicates that a large fraction of shared clones derived from the same progenitor T cells generated during fetal development (42). Twins in this study were between 50 and 63 y of age, implying that these clones persisted and expanded over decades. We noted that increased frequencies of shared amino acid sequences in twins compared to unrelated donors was still evident even when the TCRs with identical nucleotide sequences were excluded from the analysis (*SI Appendix*, Fig. S2*D*), suggesting that additional mechanisms, most likely similarities in thymic selection among twins, are involved in generating an increase in the shared repertoire.

Prior studies have shown that individuals with overlapping TCR sequence motifs share common MHC haplotypes (43). We therefore examined whether frequencies of TCR sharing between individuals correlated with sharing of HLA alleles. We combined all total T and memory CD4 TCRαβ sequences and then determined Jaccard ratios for any samples from unrelated individuals sharing identical sequences. Jaccard values were then plotted vs. the number of HLA alleles shared between individuals. Sharing of identical TCRαβ sequences was significantly increased in unrelated individuals with higher HLA similarity as determined by separate trend analyses (Fig. 4*C*). This HLA-mediated effect was evident for MHC class I and II molecules. Our combined sequence pool derived from total T cells samples, thus including both CD4 and CD8 T cells, which may explain the correlation with both HLA regions. However, we also noted that the unrelated individuals sharing 3 MHC class I alleles also shared 4 MHC class II alleles. While this strong linkage disequilibrium does not allow us to distinguish between MHC class I and II association, nonetheless MHC polymorphisms appear to account for most if not all of the genetic influence on repertoire sharing.

In summary, by analyzing TCR α/β-chain pairing in close to 1 million clonotypes, we observed that the frequencies of *TRAV*:*TRBV* and *TRAJ*:*TRBJ* combinations were mostly determined by the respective expression frequencies of germline V and J gene segments; pairing of *TRAV*:*TRBV* elements was not completely stochastic, implying that structural constraints exist; however, they were very subtle. Concordant with previously published observations, we observed a genetic influence on a biased recombination of TCR α- as well as β-chains. The by-far-largest genetic effect was seen for the sharing of identical TCRαβ sequences that was in part explained by the survival of T cells generated in the twins during fetal development. This genetic influence was not very apparent at the level of single public TCR α- or β-chain sequences, but most evident in identical pairing, implying a strong selective force in thymic selection, presumably imposed by identical MHC peptide complexes. Our data therefore suggest that the TCR repertoires in nontwin pairs are more dissimilar than previously suggested by single-chain studies. All twin pairs were older than 50 y, and clonotype sharing was even more detectable in memory than naive CD4 T cells, emphasizing the importance of the genetic over environmental influences.

Materials and Methods

Sample Collection and HLA Genotyping. All research involving human subjects has been approved by institutional review boards (Stanford University, The University of Texas at Austin). Whole blood was collected from healthy volunteers after informed consent had been obtained (Stanford University, Stanford, CA, and Gulf Coast Regional Blood Center, Houston, TX). Monozygotic twins were recruited from the SRI twin registry (44). GoldenGate genotyping (Illumina) was performed to determine zygosity by IGenix (Bainbridge Island). Peripheral blood mononuclear cells (PBMCs) were isolated by density centrifugation using Ficoll media at a density of 1.077 g/mL (#17-829E, Lonza; or #07851, Stemcell). PBMC were collected, resuspended at 10^6 cells/mL in complete RPMI-1640 medium containing 10% DMSO, and cryopreserved for up to 10 wk. A total of 2×10^5 PBMCs was used for subsequent HLA genotyping. In-house sequence-based typing of the HLA loci was performed in the Clinical Laboratory Improvement Amendments/American Society for Histocompatibility and Immunogenetics-accredited laboratory of University of Oklahoma Health Science Center as previously described (45).

Cell Purification and Flow Cytometry. On the day prior to experiments, total T cells were isolated using Pan T cell isolation kit (#130-096-535, Miltenyi Biotec). Naive CD4 and memory CD4 T cell populations were purified using magnetic-bead–based negative EasySep selection reagents (#19555, #17952; Stemcell). In separate initial experiments, we confirmed greater than 95% purity as measured by identification of $CD4^+CD45RA^+CCR7^+$ (naive) and $CD4^+CD45RA^-$ (memory) populations via flow cytometry analysis (Biolegend; clones OKT4, HI100, G043H7). Cells were stained with antibodies for 30 min at 4 °C prior to analysis. Purified T cell subsets were allowed to recover at 37 °C with 5% CO_2 in 1 mL of complete media for 2 to 3 h. After an approximate 24-h overnight transit in 1- to 2-mL vials and 30 U/mL rhIL-2, cells were harvested for microfluidic TCR library preparation. To avoid cross-contamination between samples, we restricted sample preparation to one individual per day.

High-Throughput Single-Cell Paired TCRαβ Sequencing. T cell samples were stimulated with 100 ng/mL phorbol 12-myristate 13-acetate (PMA) (# P8139; Sigma) and 100 ng/mL ionomycin (#I9657; Sigma) for 4 h to enrich TCRαβ gene transcripts. Paired TCRαβ sequencing was performed using the flow-focusing technology described earlier for the sequencing of the V_H:V_L repertoire from B cells, which we adapted here for TCR sequencing (30, 31). Primers used for *TRA* and *TRB* amplification are listed in *SI Appendix*, Table S4. Briefly, single T cells were sequestered and lysed within monodisperse water-in-oil emulsions and mRNA from the lysed encapsulated cells was captured by oligo-dT beads. The emulsion was broken, and the oligo-dT beads were used for emulsion overlap extension RT-PCR using the primers in *SI Appendix*, Table S4. This second emulsion was broken with diethyl ether; cDNA was isolated and amplified by PCR in a total volume of 250 μL using DreamTaq Hot Start DNA Polymerase (#EP1702; Thermo Fisher Scientific), with the primers shown in *SI Appendix*, Table S4 under the following conditions: 94 °C for 3 min, followed by 25 to 30 cycles of PCR amplification (94 °C for 30 s, 62 °C for 30 s, 72 °C for 1 min) and a final extension of 72 °C for 7 min. The resulting ∼550-bp *TRA*-*TRB* amplicon was gel purified (#C1003-50, #D4001-1-100, #D4003-2-48; Zymo Research) and sequenced on Illumina MiSeq 2 × 300. The MiSeq sequences were quality filtered using Trimmomatic (46) by trimming sequences following a 5-bp stretch with an average Phred score of <20; V, D, and J genes were assigned using the MiXCR software (47). TCRαβ pairs with greater than 1 read were clustered at

95% CDR3β nucleotide identity. In the cluster, a given CDR3β was paired with multiple different CDR3α. To discount cross-contamination by different cells, we selected the highest-frequency CDR3α in each cluster as the correct paired chain for the particular CDR3β. For the purpose of this study, we ignored the possibility that dual TCRα genes can be rearranged in a single cell and our technology should therefore produce 2 TCRαβ fusion amplicons from some T cells. Recovery of unique TCRαβ sequences from each sample independent of read counts is reported in *SI Appendix*, Table S2.

To validate the pairing precision of this technology, total T cells were isolated from the PBMCs and stimulated with CD3/CD28 Dynabeads (#11161D; Thermo Fisher Scientific) and 30 units/mL IL-2 (PeproTech) for 1 wk. The medium was exchanged every 3 d, and fresh beads and IL-2 were added. The expanded T cells were divided into 2 technical replicates with each containing ∼1.25 million T cells, and TCRαβ sequencing was performed for each replicate. The pairing precision was calculated with the following formula as described before (30, 31):

$$P=\sqrt{\frac{TP_{1\text{ and }2}}{TP_{1\text{ and }2}+FP_{1\text{ or }2}}}.$$

$TP_{1\text{ and }2}$ is the number of TCRβ amino acid sequences paired with identical TCRα amino acid sequences in both replicates. $FP_{1\text{ or }2}$ is the number of TCRβ amino acid sequences paired with different TCRα amino acid sequences across the replicates. P is the TCRαβ pairing precision, which was >0.92%.

Statistical Analyses. We adopted the following method to analyze the degree to which the frequencies of *TRA*:*TRB* gene segment combinations exhibited a bias that could not be attributed to the observed *TRA* or *TRB* gene segment frequencies. In essence, we calculated the Pearson correlation coefficients comparing the observed to the expected (based on gene segment frequencies) *TRA:TRB* combinations. We then compared observed correlations to distributions of correlation coefficients generated from modeling independent pairing of the α- and β-chains as follows. The data consisted of numbers of sequenced reads from different TCRα gene segments (denoted by $n_{\alpha_1}, n_{\alpha_2}, \cdots, n_{\alpha_I}$), numbers of reads from different TCRβ gene segments (denoted by $m_{\beta_1}, \cdots, m_{\beta_J}$), as well as numbers of reads from different TCRαβ gene segments pairs (denoted by $o_{\alpha_1\beta_1}, o_{\alpha_1\beta_2}, \cdots, o_{\alpha_I\beta_J}$). The observed frequency of TCRαβ gene segments pairs (i,j) was calculated as $r_{ij}=o_{\alpha_i\beta_j}/O$, where $O=\sum_{i=1}^{I}\sum_{j=1}^{J}o_{\alpha_i\beta_j}$. The expected frequency of the same gene pair was calculated as p_iq_j, where $p_i=n_{\alpha_i}/N$, $q_j=m_{\beta_j}/M$, $N=\sum_{i=1}^{I}n_{\alpha_i}$, and $M=\sum_{j=1}^{J}m_{\beta_j}$. The correlations coefficients were between 2 vectors $\{p_iq_j, i=1,\cdots,I, j=1,\cdots,J\}$ and $\{r_{ij}, i=1,\cdots,I, j=1,\cdots,J\}$. The comparison of correlation coefficients can be conducted with 1-sample or 2-sample permutation test as appropriate. The observed R-square was calculated as follows:

$$R^2=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(r_{ij}-p_iq_j\right)^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(r_{ij}-(IJ)^{-1}\right)^2}.$$

To generate the distribution of R-square under the independent pairing assumption, we

1. simulated $(o^*_{\alpha_1\beta_1}, o^*_{\alpha_1\beta_2}, \cdots, o^*_{\alpha_I\beta_J}) \sim MNOM(O, (p_1q_1, p_1q_2, \cdots, p_Iq_J))$;
2. calculated $(n^*_{\alpha_1}, n^*_{\alpha_2}, \cdots, n^*_{\alpha_I})$ as $n^*_{\alpha_i}=\sum_{j=1}^{J}o^*_{\alpha_i\beta_j}$, $i=1,\cdots,I$;
3. calculated $(m^*_{\beta_1}, m^*_{\beta_2}, \cdots, m^*_{\beta_J})$ as $m^*_{\beta_j}=\sum_{i=1}^{I}o^*_{\alpha_i\beta_j}$, $j=1,\cdots,J$;
4. calculated $p^*_i=n^*_{\alpha_i}/N$, $q^*_j=m^*_{\beta_j}/M$, and $r^*_{ij}=o^*_{\alpha_i\beta_j}/O$, $i=1,\cdots,I, j=1,\cdots,J$;
5. calculated

$$R^{*2}=\frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(r^*_{ij}-p^*_iq^*_j\right)^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(r^*_{ij}-(IJ)^{-1}\right)^2};$$

6. repeated 1 to 5 multiple times to generate many R^{*2} values to approximate the distribution of R-square under the independent pairing assumption;
7. compared the observed R-square with the expected distribution of R-square under the independent pairing assumption.

This analysis was conducted separately for each individual, each type of cells (memory and naive), and V and J gene segments (e.g., *TRAV* with *TRBV*, *TRAJ* with *TRBJ*) (*SI Appendix*, Fig. S1). Specifically, the paired permutation test was used to compare the correlation coefficient related to *TRAV* and *TRBV* pairing with the correlation coefficient related to *TRAJ* and *TRBJ* pairing (Fig. 1*C*).

As a supplementary sensitivity analysis, we have also repeated the test based on the mutual information, which is also a dependence metric for the distribution of *TRA:TRB* combinations. Specifically, the mutual information can be estimated as follows:

$$MI=\sum_{i=1}^{I}\sum_{j=1}^{J}r_{ij}\log\frac{r_{ij}}{p_iq_j},$$

where $0\times\log 0$ is defined as zero. $MI=0$ represents completely random combinations between *TRA* and *TRB*. In the hypothesis test, we have compared the observed MI with the distribution of MI^*s from the simulated data under the independent pairing assumption as the analysis for R^2. The results were similar to those based on R^2 and reported in *SI Appendix*, Fig. S1.

In addition, we also constructed a debiased point estimator for the mutual information and the associated 95% confidence interval. The naive estimator of the mutual information above can be biased when the numbers of reads of some combinations are very small or zero. To remove this bias, we used the bootstrap method. Specifically, we bootstrapped the original data and constructed the bootstrapped mutual information estimator as following:

1) simulated $(o^*_{\alpha_1\beta_1}, o^*_{\alpha_1\beta_2}, \cdots, o^*_{\alpha_I\beta_J})$; where $o^*_{\alpha_i\beta_j}\sim\text{Poisson}(o_{\alpha_i\beta_j})$, $i=1,\cdots,I$, $j=1,\cdots J$;
2) calculated $(n^*_{\alpha_1}, n^*_{\alpha_2}, \cdots, n^*_{\alpha_I})$ as $n^*_{\alpha_i}=\sum_{j=1}^{J}o^*_{\alpha_i\beta_j}$, $i=1,\cdots,I$;
3) calculated $(m^*_{\beta_1}, m^*_{\beta_2}, \cdots, m^*_{\beta_J})$ as $m^*_{\beta_j}=\sum_{i=1}^{I}o^*_{\alpha_i\beta_j}$, $j=1,\cdots,J$;
4) calculated $p^*_i=n^*_{\alpha_i}/N$, $q^*_j=m^*_{\beta_j}/M$, and $r^*_{ij}=o^*_{\alpha_i\beta_j}/O$, $i=1,\cdots,I, j=1,\cdots,J$;
5) obtain the bootstrapped mutual information estimator:

$$MI^*=\sum_{i=1}^{I}\sum_{j=1}^{J}r^*_{ij}\log\frac{r^*_{ij}}{p^*_iq^*_j}.$$

The bias of the naive mutual information estimator can be approximated as follows:

$$\text{average}(MI^*)-MI,$$

and the new debiased mutual information estimator is as follows:

$$MI-(\text{average}(MI^*)-MI)=2MI-\text{average}(MI^*),$$

where $\text{average}(MI^*)$ is the empirical average of bootstrapped mutual information estimators from multiple, e.g., 500, bootstrapped datasets. The 95% confidence interval of the debiased estimator is constructed by double bootstrap. If the 95% confidence interval excludes zero, then *TRA* and *TRB* pairing are not completely independent with statistical confidence. If the 95% confidence interval includes zero and is adequately narrow, *TRA* and *TRB* pairings are approximately independent. The results are reported in *SI Appendix*, Table S3.

Comparison of the interperson correlation coefficients of gene segment pair frequencies between twins and unrelated individuals (Fig. 2) was done by permutation test as follows. We selected the average twin–twin correlation coefficients as the test statistics. To generate the null distribution under the hypothesis that the correlations between twins are identical to those between unrelated individuals, we randomly permuted the twin labels and recalculated the test statistics accordingly. Since there are 15 different ways to form 3 pairs from 6 individuals, the test statistics can only take 15 distinct values under permutation and the smallest one-sided P value is 1/15 = 0.067. The same permutation test is used to compare the Jaccard indices between twin pairs and between unrelated individuals (Fig. 3).

The Jaccard index, also known as the intersection over the union, was calculated as follows:

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|}.$$

Normalized sharing frequency was calculated as follows:

$$N(A,B)=\frac{|A\cap B|}{|A|\times|B|},$$

where $|A\cap B|$ is the number of identical sequences shared among any 2

samples (A and B), and $|A|$ and $|B|$ are the total number of unique sequences present in samples A and B, respectively.

ACKNOWLEDGMENTS. We thank the Genome Sequencing and Analysis Facility at the University of Texas at Austin for performing Illumina sequencing. This work was supported by NIH Grants U19 AI057266 (to G.G. and J.J.G.) and R01 AI129191 (to J.J.G.) and US Defense Threat Reduction Agency Grant HDTRA1-12-C-0105 (to G.G.). H.T. was supported by University of Texas Health Innovation for Cancer Prevention Research Training Program Postdoctoral Fellowship (Cancer Prevention and Research Institute of Texas Grant RP160015), Japan Society for the Promotion of Science Postdoctoral Fellowships for Research Abroad, and Uehara Memorial Foundation Research Fellowship. We thank the study participants and the staff of the Stanford–Lucile Packard Vaccine Program for conducting the twin study; the Stanford Clinical and Translational Research Unit provided resources through NIH–National Center for Advancing Translational Sciences–Clinical and Translational Science Award Grant UL1 TR001085. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the Cancer Prevention and Research Institute of Texas.

1. B. Toyonaga, Y. Yoshikai, V. Vadasz, B. Chin, T. W. Mak, Organization and sequences of the diversity, joining, and constant region genes of the human T-cell receptor beta chain. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 8624–8628 (1985).
2. Y. Yoshikai *et al*., Organization and sequences of the variable, joining and constant region genes of the human T-cell receptor alpha-chain. *Nature* **316**, 837–840 (1985).
3. D. M. Pardoll, B. J. Fowlkes, R. I. Lechler, R. N. Germain, R. H. Schwartz, Early genetic events in T cell development analyzed by in situ hybridization. *J. Exp. Med.* **165**, 1624–1638 (1987).
4. T. P. Arstila *et al*., A direct estimate of the human alphabeta T cell receptor diversity. *Science* **286**, 958–961 (1999).
5. T. P. Arstila *et al*., Diversity of human alpha beta T cell receptors. *Science* **288**, 1135 (2000).
6. M. M. Davis, P. J. Bjorkman, T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).
7. J. J. Goronzy, Q. Qi, R. A. Olshen, C. M. Weyand, High-throughput sequencing insights into T-cell receptor repertoire diversity in aging. *Genome Med.* **7**, 117 (2015).
8. E. Bianconi *et al*., An estimation of the number of cells in the human body. *Ann. Hum. Biol.* **40**, 463–471 (2013).
9. L. Klein, B. Kyewski, P. M. Allen, K. A. Hogquist, Positive and negative selection of the T cell repertoire: What thymocytes see (and don't see). *Nat. Rev. Immunol.* **14**, 377–391 (2014).
10. M. Egerton, R. Scollay, K. Shortman, Kinetics of mature T-cell development in the thymus. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 2579–2582 (1990).
11. D. J. Laydon, C. R. Bangham, B. Asquith, Estimating T-cell repertoire diversity: Limitations of classical estimators and a new approach. *Philos. Trans. R Soc. Lond. B Biol. Sci.* **370**, 20140291 (2015).
12. E. S. Lee, P. G. Thomas, J. E. Mold, A. J. Yates, Identifying T cell receptors from high-throughput sequencing: Dealing with promiscuity in TCRα and TCRβ pairing. *PLoS Comput. Biol.* **13**, e1005313 (2017).
13. Q. Qi *et al*., Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 13139–13144 (2014).
14. B. D. Stadinski *et al*., Hydrophobic CDR3 residues promote the development of self-reactive T cells. *Nat. Immunol.* **17**, 946–955 (2016).
15. K. W. Wucherpfennig, D. Sethi, T cell receptor recognition of self and foreign antigens in the induction of autoimmunity. *Semin. Immunol.* **23**, 84–91 (2011).
16. I. V. Zvyagin *et al*., Distinctive properties of identical twins' TCR repertoires revealed by high-throughput sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 5980–5985 (2014).
17. E. Sharon *et al*., Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nat. Genet.* **48**, 995–1002 (2016).
18. T. Dupic, Q. Marcou, A. M. Walczak, T. Mora, Genesis of the αβ T-cell receptor. *PLoS Comput. Biol.* **15**, e1006874 (2019).
19. J. S. Menezes *et al*., A public T cell clonotype within a heterogeneous autoreactive repertoire is dominant in driving EAE. *J. Clin. Invest.* **117**, 2176–2185 (2007).
20. Y. Zhao *et al*., Preferential use of public TCR during autoimmune encephalomyelitis. *J. Immunol.* **196**, 4905–4914 (2016).
21. H. S. Robins *et al*., Overlap and effective size of the human $CD8^+$ T cell receptor repertoire. *Sci. Transl. Med.* **2**, 47ra64 (2010).
22. V. Venturi *et al*., A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *J. Immunol.* **186**, 4285–4294 (2011).
23. M. V. Pogorelyy *et al*., Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 12704–12709 (2018).
24. Q. Qi *et al*., Diversification of the antigen-specific T cell receptor repertoire after varicella zoster vaccination. *Sci. Transl. Med.* **8**, 332ra46 (2016).
25. A. Han, J. Glanville, L. Hansmann, M. M. Davis, Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat. Biotechnol.* **32**, 684–692 (2014).
26. B. Howie *et al*., High-throughput pairing of T cell receptor α and β sequences. *Sci. Transl. Med.* **7**, 301ra131 (2015).
27. A. W. Briggs *et al*., Tumor-infiltrating immune repertoires captured by single-cell barcoding in emulsion. bioRxiv:10.1101/134841 (5 May 2017).
28. J. A. Carter *et al*., T-cell receptor alphabeta chain pairing is associated with $CD4^+$ and $CD8^+$ lineage specification. bioRxiv:10.1101/293852 (3 April 2018).
29. K. Grigaityte *et al*., Single-cell sequencing reveals alphabeta chain pairing shapes the T cell repertoire. bioRxiv:10.1101/213462 (2 November 2017).
30. B. J. DeKosky *et al*., In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* **21**, 86–91 (2015).
31. J. R. McDaniel, B. J. DeKosky, H. Tanno, A. D. Ellington, G. Georgiou, Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. *Nat. Protoc.* **11**, 429–442 (2016).
32. J. D. Freeman, R. L. Warren, J. R. Webb, B. H. Nelson, R. A. Holt, Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res.* **19**, 1817–1824 (2009).
33. E. Quirós Roldan *et al*., Different TCRBV genes generate biased patterns of V-D-J diversity in human T cells. *Immunogenetics* **41**, 91–100 (1995).
34. D. N. Garboczi *et al*., Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. *Nature* **384**, 134–141 (1996).
35. K. C. Garcia *et al*., An alphabeta T cell receptor structure at 2.5 Å and its orientation in the TCR-MHC complex. *Science* **274**, 209–219 (1996).
36. M. G. Rudolph, R. L. Stanfield, I. A. Wilson, How TCRs bind MHCs, peptides, and coreceptors. *Annu. Rev. Immunol.* **24**, 419–466 (2006).
37. F. Rubelt *et al*., Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. *Nat. Commun.* **7**, 11112 (2016).
38. Y. Elhanati, Z. Sethna, C. G. Callan, Jr, T. Mora, A. M. Walczak, Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunol. Rev.* **284**, 167–179 (2018).
39. V. Venturi, D. A. Price, D. C. Douek, M. P. Davenport, The molecular basis for public T-cell responses? *Nat. Rev. Immunol.* **8**, 231–238 (2008).
40. H. Y. Greenaway *et al*., NKT and MAIT invariant TCRα sequences can be produced efficiently by VJ gene recombination. *Immunobiology* **218**, 213–224 (2013).
41. V. Venturi, B. D. Rudd, M. P. Davenport, Specificity, promiscuity, and precursor frequency in immunoreceptors. *Curr. Opin. Immunol.* **25**, 639–645 (2013).
42. M. V. Pogorelyy *et al*., Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. *PLoS Comput. Biol.* **13**, e1005572 (2017).
43. J. Glanville *et al*., Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).
44. R. E. Krasnow, L. M. Jack, C. N. Lessov-Schlaggar, A. W. Bergen, G. E. Swan, The twin research registry at SRI international. *Twin Res. Hum. Genet.* **16**, 463–470 (2013).
45. M. C. Lanteri *et al*., Association between HLA class I and class II alleles and the outcome of west nile virus infection: An exploratory study. *PLoS One* **6**, e22948 (2011).
46. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
47. D. A. Bolotin *et al*., MiXCR: Software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).