HHS Public Access

Author manuscript

Cell Rep. Author manuscript; available in PMC 2020 August 18.

Published in final edited form as:

Cell Rep. 2020 July 14; 32(2): 107882. doi:10.1016/j.celrep.2020.107882.

High Frequency of Shared Clonotypes in Human T Cell Receptor Repertoires

Cinque Soto[1,2], **Robin G. Bombardi**[1], **Morgan Kozhevnikov**[1], **Robert S. Sinkovits**[5], **Elaine C. Chen**[3], **Andre Branchizio**[1], **Nurgun Kose**[1], **Samuel B. Day**[1], **Mark Pilkinton**[4], **Madhusudan Gujral**[5], **Simon Mallal**[3,4], **James E. Crowe Jr.**[1,2,3,6,*]

[1]The Vanderbilt Vaccine Center, Vanderbilt University Medical Center, Nashville, TN 37232, USA

[2]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN 37232, USA

[3]Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN 37212, USA

[4]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37232, USA

[5]San Diego Supercomputer Center, University of California, San Diego, San Diego, CA 92093, USA

[6]Lead Contact

SUMMARY

The collection of T cell receptors (TCRs) generated by somatic recombination is large but unknown. We generate large TCR repertoire datasets as a resource to facilitate detailed studies of the role of TCR clonotypes and repertoires in health and disease. We estimate the size of individual human recombined and expressed TCRs by sequence analysis and determine the extent of sharing between individual repertoires. Our experiments reveal that each blood sample contains between 5 million and 21 million TCR clonotypes. Three individuals share 8% of TCRβ- or 11% of TCRα-chain clonotypes. Sorting by T cell phenotypes in four individuals shows that 5% of naive CD4+ and 3.5% of naive CD8+ subsets share their TCRβ clonotypes, whereas memory CD4+ and CD8+ subsets share 2.3% and 0.4% of their clonotypes, respectively. We identify the sequences of these shared TCR clonotypes that are of interest for studies of human T cell biology.

Graphical Abstract

*Correspondence: james.crowe@vumc.org.

AUTHOR CONTRIBUTIONS

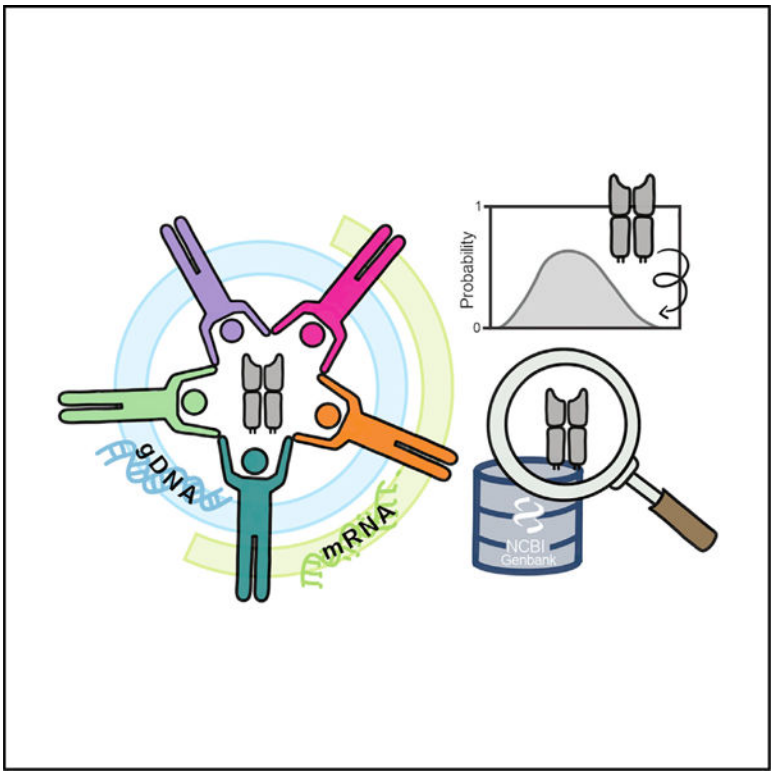
C.S., R.G.B., M.P., S.M., and J.E.C. planned the studies. C.S., R.G.B., M.K., R.S.S., E.C.C., A.B., N.K., S.B.D., M.G., S.M., and M.P. conducted experiments. C.S., R.G.B., R.S.S., E.C.C., S.B.D., M.P., S.M., and J.E.C. interpreted the studies. C.S. and J.E.C. wrote the first draft of the paper. All authors reviewed, edited, and approved the paper. J.E.C. obtained funding.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2020.107882.

DECLARATION OF INTERESTS

J.E.C. has served as a consultant for Takeda Vaccines, Sanofi Pasteur, Pfizer, and Novavax; is on the scientific advisory boards of CompuVax and Meissa Vaccines; and is founder of IDBiologics, Inc.



In Brief

Soto et al. examine the extent to which five healthy adults share their T cell receptor (TCR) repertoire. Using sequencing and bioinformatics, they show a high prevalence of shared clonotypes even considering different T cell phenotypes. Possible functions for some clonotypes are inferred based on homology with TCRs in GenBank.

INTRODUCTION

Healthy immune systems are characterized by diverse T cell receptor (TCR) repertoires. The potential diversity of complete TCR repertoires formed by the process of somatic recombination of variable (V), diversity (D), and joining (J) gene segments (V(D)J recombination) is large. Recent reports of estimates of the size and extent of sharing of B cell receptor (BCR) diversity using next-generation repertoire sequencing showed that there is an un-expectedly high level of sharing in human BCR repertoires (Briney et al., 2019; Soto et al., 2019). A comprehensive estimate of a complete set of recombined human TCR genes has not yet been determined because of the extremely large size. Sharing between TCRβ repertoires has been described previously (Putintseva et al., 2013; Robins et al., 2010; Shugay et al., 2013), but previous efforts to sequence TCRs were not conducted at a scale that enables estimates of the true size of the repertoires or the full extent of sharing. Here, we sought to estimate the size and diversity of human TCR repertoires by sequencing the repertoires of five healthy adults and then determining the number of shared clonotypes present. This dataset is a resource that can facilitate future detailed studies of human TCR repertoires in health and disease.

RESULTS

We used two alternate definitions of clonotypes. We determined the variable (Vα or Vβ) and joining (Jα or Jβ) germline gene and the non-templated regions for each recombined TCR V gene sequence detected. We designated T cell recombined V region sequences as members of a single V3J clonotype if the sequences (1) were encoded by the same TCR Vβ+Jβ or Vα +Jα gene segment combination (ignoring allelic distinctions) and (2) possessed identical amino acid sequences in the complementarity determining region 3 (CDR3). These V3J clonotype identification criteria provide a structured method for grouping TCR sequences and can be applied across immune repertoire sequencing methods, regardless of the amplicon length or the presence of sequence errors in any germline genes. A second, more detailed representation of the TCRβ clonotype includes an accurate diversity (Dβ) germline assignment that we call a V3DJ clonotype. The V3DJ clonotype was used to provide statistical relevance to observed sharing between samples.

We isolated large numbers of peripheral blood mononuclear cells (PBMCs) from five healthy donors enrolled in our Human Immunome Program (HIP) study. We leukapheresed all five donors designated as HIP1 (female, age 47 years), HIP2 (male, age 22 years), HIP3 (male, age 29 years), HIP4 (male, age 32 years), or HIP5 (female, age 30 years) (Table 1) and obtained 0.94 billion, 1.63 billion, 1.71 billion, 3.9 billion, or 8.9 billion PBMCs (Table S1), respectively. We also performed human leukocyte antigen (HLA) typing on all five HIP samples (Table S2). To increase our sequencing depth and limit biases inherent to any single T cell sequencing assay, we used different laboratories and diverse techniques to prepare libraries and to sequence the TCR repertoires of all samples. Although we used total PBMCs as the starting cell population, we analyzed the repertoire of purified memory or naive subsets of CD4+ or CD8+ T cells identified by cell surface marker phenotypes.

We considered two types of sequencing methods for the amplification of TCRs belonging to subjects HIP1, HIP2, and HIP3. The first method is a mRNA-based amplicon sequencing method that provides full-length V(D)J sequences using 5′ multiplex PCR primers that are designed within the leader sequences of each productive V gene and the 3′ primers within the constant regions. The second method is a genomic DNA (gDNA)-based amplicon sequencing method and provides sequencing for the CDR3 junctional region that includes parts of framework 3 and framework 4. The sequencing reactions for the mRNA method yielded 3.2×10^8 raw sequencing reads for subject HIP1, 6.7×10^8 reads for subject HIP2, and 4.6×10^8 reads for subject HIP3. The gDNA method yielded pre-processed data that resulted in 7.5×10^6 sequences for HIP1, 2.3×10^7 sequences for HIP2, and 2.2×10^7 sequences for HIP3. We processed all sequences to remove non-productive reads (see STAR Methods for filtering steps). The total number of unique reads that remained after quality control filtering was 2.7×10^7, 5.4×10^7, or 7.9×10^7 for subject HIP1, HIP2, or HIP3, respectively. All processed reads that remained after quality control filtering were designated productive reads. We assigned the inferred germline V gene segments for TCR sequences and identified junctional residues using the PyIR bioinformatics pipeline based on IgBLAST (Ye et al., 2013).

Estimating the Total Number of TCRβ Clonotypes in the Blood of a Healthy Adult

We used data modeling to determine whether the depth of sequencing was adequate to identify most V3J clonotypes in circulation in subject HIP1, HIP2, or HIP3. The total number of unique V3J clonotypes for each subject represented the species richness value, which is a component of the diversity of each repertoire. We constructed rarefaction curves based on the number of unique V3J clonotypes appearing in the pool of productive reads for subjects HIP1, HIP2, and HIP3 with the program RTK (Saary et al., 2017). The curves did not plateau, suggesting that we did not sequence the TCRβ repertoire of any donor to completion (Figure 1A). HIP1 had the smallest estimated species richness value (n = 5,908,621 unique V3J clonotypes), followed by HIP2 (n = 17,595,796 unique V3J clonotypes) and then HIP3 (n = 21,085,908 unique V3J clonotypes). A lower-bound estimator of species richness often used in ecological surveys to estimate the expected size of a population is the Chao1 estimator (Chao, 1987). Application of the Chao1 estimator to each collection of unique V3J clonotypes from subjects HIP1, HIP2, and HIP3 gave lower-bound species richness estimates of 25.6 million V3J clonotypes for HIP1, 115 million clonotypes for HIP2, and 80.3 million V3J clonotypes for HIP3. These estimates were roughly 4.3 and 3.8 times larger than the experimentally observed number of V3J clonotypes for HIP1 and HIP3. For HIP2, the Chao1 estimate was roughly 6.5 times larger than the experimentally observed number of V3J clonotypes. The large lower-bound estimates result from the occurrence of a large number of clonotypes appearing only once (singleton species) in the gDNA sequencing datasets, as evidenced by pooling the mRNA and gDNA sequence datasets followed by rarefaction analysis (Figure S1A). We also used the program Recon (Kaplinsky and Arnaout, 2016) to estimate the number of V3J clonotypes that we did not observe with experimental sequencing. Estimates from Recon suggested an additional 38.9 million V3J clonotypes should be present for HIP1, an additional 134 million V3J clonotypes should be present for HIP2, and an additional 120 million V3J clonotypes should be present for HIP3 (Figure 1B). The Recon estimates for the missing V3J clonotypes were larger for HIP2 and HIP3 when considering sequencing from the gDNA method. For HIP3, the estimated number of missing V3J clonotypes from the gDNA sequencing method was 3.9 times larger than for the mRNA sequencing method (Figure S1B). For HIP2, the number of missing V3J clonotypes for the gDNA sequencing method was roughly 8 times larger than for the mRNA sequencing method (Figure S1B). To obtain global estimates for the number of V3J clonotypes that we can expect to be unaccounted for, we averaged Chao1 and Recon estimates over these three HIP donors. The average value for the Chao1 estimate suggested a lower-bound value of about 74 million V3J clonotypes. Similarly, the Recon estimate averaged over three donors suggested that we failed to sequence roughly 98 million V3J clonotypes at this depth of sequencing. Summing the averaged Chao1 and Recon values to obtain a rough estimate of clonotype size in the blood suggests that there are at least 171 million TCRβ V3J clonotypes in any individual. However, the differences in the shape of the rarefaction curves between mRNA and gDNA sequencing lead to skewing of these estimates.

Public TCRβ Clonotypes in Healthy Adults

We next determined the degree to which the TCRβ clonotypes were shared among HIP1, HIP2, and HIP3. TCRβ clonotype sharing among individuals has been reported previously

using bulk mRNA sequencing (Putintseva et al., 2013; Shugay et al., 2013) or gDNA sequencing (Robins et al., 2010). Here, we sought to determine whether we would observe a larger degree of sharing using both mRNA and gDNA sequencing methods performed at this depth of sequencing. We first pooled both mRNA and gDNA datasets for subject HIP1, HIP2, or HIP3 to create total individual repertoires, then we estimated the sharing of TCRβ clonotypes among individuals. HIP1 shared 13% of its V3J clonotypes with HIP2 (n = 781,009 common V3J clonotypes) and HIP3 (n = 725,934 common V3J clonotypes). HIP2 shared 12% of its V3J clonotypes with HIP3 (n = 2,063,129 common V3J clonotypes) (Figure 2A). All three HIP subjects shared 8% of their V3J clonotypes (n = 476,920 common V3J clonotypes). The pairwise percentage overlaps among HIP1, HIP2, and HIP3 ranged from 4% to 7% and were 2% overall using mRNA sequencing (Figure S2A). The pairwise percentage overlaps ranged from 12% to 20% and were 11% overall using gDNA sequencing (Figure S2B). Significant overlaps have been described previously with shallower sequencing approaches by Putintseva and co-workers, who observed overlaps of TCRβ CDR3s between two donors as high as 11% (Putintseva et al., 2013; Shugay et al., 2013). Even so, the higher percentages observed between the gDNA sequencing datasets may partly stem from the gDNA method yielding about twice as many V3J clonotypes as the mRNA method for subject HIP2. However, the number of V3J clonotypes obtained by the two methods in the HIP1 and HIP3 sets was comparable (Figures S2A and S2B).

We performed a second analysis of sharing that included only V3J clonotypes for which a Dβ gene assignment could be made with confidence. Because our goal was to use accurate Dβ gene assignments, we set the expectation value (E value) threshold in IgBLAST (Ye et al., 2013) for matching to a germline Dβ gene to 0.1. As expected, this threshold eliminated a significant fraction of the TCRβ V3J clonotypes from consideration, because Dβ genes can be assigned unambiguously in only a subset of human TCRβ sequences. V3J clonotypes with an explicit Dβ gene assignment were designated V3DJ clonotypes. The median CDR3 length obtained from the V3DJ clonotypes was 14 amino acids for all three HIP donors (Figure S2C). The percentages of overlapping V3DJ clonotypes was similar to those we obtained with V3J clonotypes (Figure 2B). HIP1 and HIP2 shared 12% of their clonotypes (n = 170,068 common V3DJ clonotypes), HIP1 and HIP3 shared 11% of their clonotypes (n = 156,965 common V3DJ clonotypes), and HIP2 and HIP3 shared 9% of their clonotypes (n = 421,832 common V3DJ clonotypes). All three HIP subjects shared 6% of their clonotypes (n = 89,831 common V3DJ clonotypes). Thus, even with a substantial reduction in clonotype counts when considering V3DJ clonotypes, the degree of TCRβ clonotype sharing remained similar.

To determine whether the degree of clonotype sharing among the three HIP subjects might result from chance or reflected a biological mechanism that causes common selection of certain clonotypes, we constructed null model repertoires using the V(D)J frequencies (VDJ triples) observed in each of the three experimentally determined HIP TCRβ repertoires. The construction of null models requires knowledge of the joint distribution of V(D)J frequencies from the experimentally determined immune repertoire (Arnaout et al., 2011; Boyd et al., 2009). We generated three sets of synthetic TCRβ V3DJ clonotypes using IGoR (Marcou et al., 2018). Each set consisted of three large ensembles (termed simHIP1, simHIP2, and simHIP3), with each containing more than 0.5 billion synthetic V3DJ

clonotypes. We sampled VDJ triples from each set of the three large ensembles of synthetic read collections according to the frequency distribution of VDJ triples from each of the experimentally determined repertoires (Figure S2D). The selection of clonotypes was accomplished by randomly sampling unique amino acid CDR3 sequences from 3 to 19 residues in length (about 99.9% of the experimental V3DJ CDR3 length distribution) from each synthetic VDJ triple until we obtained the same CDR3 length frequency distribution observed in the corresponding experimental VDJ triple set. We subsampled from simHIP1 (n = 1,376,322 unique synthetic V3DJ clonotypes), simHIP2 (n = 4,397,913 unique synthetic V3DJ clonotypes), and simHIP3 (n = 4,933,497 unique synthetic V3DJ clonotypes), each 500 times, and then determined the percentage of overlapping clonotypes from 50,000 permutations of the three subsampled sets. The average percentage overlap in the simulated repertoires ranged from 0.2% to 0.3% between pairs and was 0.01% for the intersection of all pairs (Figure 2C). The overlap count among the three experimental repertoires (89,831 common V3DJ clonotypes) ranked higher than the highest overlap count (184 common V3DJ clonotypes) obtained from any of the 50,000 comparisons of the synthetic repertoires (Figure 2D). When compared against null model repertoires, the high degree of sharing suggests that a strong biological mechanism shapes individual human TCRβ repertoires in a common manner.

We also sequenced the TCRα chains from subjects HIP1, HIP2, and HIP3 using the mRNA method. As expected, the percentage of shared V3J clonotypes for the TCRα datasets was higher, because these chains lack a D gene segment. The median CDR3 length was 12 amino acids across all three HIP subjects (Figure S2E). Subjects HIP1 and HIP2 shared 12% of their clonotypes (n = 196,053 common TCRα V3J clonotypes). HIP3 shared 16% of its clonotypes with HIP1 (n = 261,507 common TCRα V3J clonotypes) and 20% of its clonotypes with HIP2 (n = 410,295 common TCRα V3J clonotypes) (Figure S2F). All three HIP subjects shared 9% of their clonotypes (n = 140,473 common TCRα V3J clonotypes). The high degree of overall sharing between subject HIP2 and subject HIP3 suggests lower diversity in a chains. Similar results for sharing of light-chain immunoglobulins have been observed previously (Hoi and Ippolito, 2013; Soto et al., 2019).

Differences in Immune Repertoires when Using mRNA or gDNA Sequencing

We compared the molecular features of the sequenced TCRβ repertoires based on the different immune repertoire sequencing methods (Table S1). The mRNA-based method (AbHelix) uses a multiplex RT-PCR approach with primers embedded in the leader sequence and constant sequence and provides full-length V(D)J sequencing. The gDNA-based sequencing method (Adaptive ImmunoSEQ) uses a multiplex PCR approach and identifies the sequence for the CDR3 junctional region that includes parts of framework 3 and framework 4. Sequencing with the mRNA method gave median CDR3 lengths of 13 amino acids for HIP1 (n = 3,161,410 unique CDR3s), HIP2 (n = 5,104,666 unique CDR3s), and HIP3 (n = 9,182,164 unique CDR3s) (Figure 3A). Similarly, sequencing with the gDNA method gave median CDR3 lengths of 13 amino acids for HIP1 (n = 2,443,117 unique CDR3s), HIP2 (n = 9,967,538 unique CDR3s), and HIP3 (n = 9,018,345 unique CDR3s) (Figure 3A). Another repertoire feature that we used to characterize the two sequencing methodologies was the frequency of Vβ and Jβ germline gene combinations. The patterns of

Vβ+Jβ germline usage were distinct for each sample in both mRNA and gDNA sequencing methods (Figures S3A and S3B). We also used an alternate approach for assessing differences between Vβ+Jβ germline gene usage based on the Morisita-Horn index of similarity. The Morisita-Horn index of similarity is used in ecology to measure the similarity between two populations and ranges from a value of 0 (dissimilar) to 1 (identical) (Horn, 1966). The Morisita-Horn index of similarity between mRNA and gDNA sequencing methods showed weak (HIP1 = 0.25) to modest (HIP2 = 0.71) levels of similarity when using normalized Vβ+Jβ germline frequencies (Figure 3B). The clonotype abundances between mRNA and gDNA sequencing methods indicated that the mRNA method of sequencing produced a few clonotypes that dominated the individual repertoires; roughly 60 to 80% of the repertoires are accounted for by just 20% of the clonotypes (Figure 3C). Dominance by just a few clonotypes using the mRNA method for sequencing, particularly for the memory component of the TCRβ repertoire, has been described previously by Venturi et al. (2011). For the gDNA method, abundances were not concentrated on just a few clonotypes; instead, they scaled linearly with unique clonotype counts. The repertoire features between mRNA and gDNA sequencing methodologies gave identical median CDR3 lengths but displayed substantial differences in Vβ+Jβ germline gene usage. In addition, the mRNA method resulted in repertoires that were dominated by a subset of V3J clonotypes. This was in contrast to the repertoires from the gDNA method, which showed roughly equal distribution of unique somatic variants to unique V3J clonotypes.

A more precise measurement of similarity between the two sequencing methods requires consideration of the TCRβ sequences. However, the two sequencing methods provided target amplicons of differing lengths. The mRNA method provides sequencing for a portion of the leader region through a portion of the constant region, whereas the gDNA method provides a truncated version of the TCRβ sequence (partial framework 3 through partial framework 4). The V3J clonotype definition that we used provides a minimal unit of information that includes Vβ and Jβ germline information, as well as the CDR3 amino acid sequence. This definition avoids having to deal with differences in amplicon lengths between sequencing methodologies, because both methods provide the sequence of the CDR3 region. The mRNA and gDNA methods yielded between 3 to 13 million unique V3J clonotypes for HIP1, HIP2, or HIP3. HIP1 mRNA shared about 14% of its clonotypes with HIP1 gDNA (n = 398,120 common V3J clonotypes), HIP2 mRNA shared about 13% of its clonotypes with HIP2 gDNA (n = 736,788 common V3J clonotypes), and HIP3 mRNA shared 13% of its clonotypes with HIP3 gDNA (n = 1,440,497 common V3J clonotypes) (Figure 3D). Subsampling down to the size of the smaller of the two populations being compared reduced the overlaps by roughly half, yielding median percentage overlaps of 6.8% ($6.8\% \pm 1.1\% \times 10^{-2}$), 5.6% ($5.6\% \pm 7.5\% \times 10^{-3}$), and 6.6% ($6.6\% \pm 5.5\% \times 10^{-3}$) for subjects HIP1, HIP2, and HIP3, respectively. This modest overlap from deep sequencing from the two methods was expected for several reasons. The principal reason for the modest overlap of the two methods is the incomplete nature of the sequence sets, despite the extreme depth of sequencing; both the rarefaction curves (Figure S1A) and Chao1/Recon estimates suggested incomplete experimental sequencing. To quantify the reproducibility of the mRNA sequencing method, we determined the median percentage of overlapping TCRβ clonotypes that occurred between individual lanes of a HiSeq 2500 flow cell (in which each lane

contained cDNA from independent biological replicates). For the mRNA sequencing method, we observed a median percentage overlap of roughly 59% (Figure S3C). A higher median percentage overlap value of 70% was observed for the TCRα clonotypes (Figure S3D). We also quantified reproducibility of the mRNA sequencing method using the Morisita-Horn index of similarity by comparing the frequency of TCRβ clonotypes between individual lanes of a HiSeq 2500 flow cell. The median value for the Morisita-Horn index of similarity was 0.95 or greater for all pairwise comparisons of all five or six lanes of a HiSeq 2500 flow cell (Figure S3E). Such high concordance between lanes of a flow cell provides one measurement of reproducibility for the mRNA method. Sequencing reproducibility for the gDNA method used here has been described previously (Robins et al., 2012). We also analyzed the degree to which the most abundant clonotypes are shared between mRNA and gDNA methods (Figure S3F). After ranking the most abundant clonotypes from the mRNA and gDNA methods, we observed that 50 clonotypes in the gDNA set for HIP1 are shared and appear in the top 100 ranking. In the mRNA set for HIP1, 72 clonotypes are shared and appear in the top 100 ranking. For HIP2 and HIP3, 85 or more clonotypes are shared and appear in the top 100 ranking for both mRNA and gDNA methods (Figure S3F). These results suggest that the most abundant clonotypes are similar between the two sequencing methods. The less abundant clonotypes, especially the singletons, dominate the repertoire even at this depth of sequencing and ultimately reduce the degree of clonotype sharing between the two methods. Thus, although we expected to see a high degree of sharing when sequencing from the same donor using the same method, the overlap of repertoires obtained from distinctly different methods (mRNA versus gDNA, nucleic acid source, differing amplicon lengths, different amplification primers and conditions, different blood draws, etc.) was expected to be lower and indeed was lower (~13% for TCRβ). Early work by Warren and coworkers showed that even when sequencing mRNA from the same donor, different blood draws can lead to just 13% of nucleotide sequences being shared (Warren et al., 2011). Thus, the added complexity of using distinctly different methods and blood draws led to lower percentage overlaps when compared with using a single method (i.e., mRNA method). The combined large-scale mRNA + gDNA sequence repertoires provided in this resource thus have the theoretical advantages of increased diversity and depth and reduced biases over single-method reference sets.

V3J Clonotype Sharing between TCRβ Cell Subsets

The collection of total PBMCs and sequence analysis of all TCRβ clonotypes in the HIP1, HIP2, and HIP3 samples represented receptors from a mix of diverse T cell types, so we next conducted analyses on subsets of naive or memory CD4+ or CD8+ cells, using sorted cells from subject HIP2, HIP3, HIP4, or HIP5 (Table 1). All sorted cells were sequenced using the gDNA method only. Rarefaction analyses on the sorted datasets displayed evidence of incomplete sequencing (data not shown). The Chao1 and Recon estimates were generally larger for the naive subsets than for the memory subsets (Table S3). If we considered the sum of the Chao1 and Recon estimate to be a naive lower-bound estimate for the number of expected clonotypes at this depth of sequencing, the naive CD4+ set would have the largest number of expected clonotypes based on experimental sequencing, with a median value of about 161 million clonotypes (152.5 million ± 60.7 million). The naive CD8+ set would have a median value of about 84 million clonotypes (88.6 million ± 31.9

million). The memory sets would have smaller median values of 37 million clonotypes (36.3 million ± 17.3 million) for the CD4+ subset and 11 million clonotypes (12.6 million ± 5.3 million) for the CD8+ subset. Pooling the clonotypes from each set and then determining the total number of unique clonotypes suggested that one would need a factor of about 7 to 9 more sequencing depth for the naive sets (CD4+ = 17.2 million unique clonotypes, CD8+ = 11.1 million unique clonotypes) and a factor of about 5 more sequencing depth for the memory sets (CD4+ = 7.4 million unique clonotypes, CD8+ = 1.9 million unique clonotypes) to obtain agreement with expected estimates.

We also analyzed the degree of TCRβ clonotype sharing between naive and memory sets (Figure 4). The median CDR3 length distributions for all subjects was identical with a value of 13 amino acids, ensuring that differences in CDR3 length would not confound analysis involving clonotype sharing (Figure S4A). Pairwise percentage overlaps for the naive sets were higher when compared with the memory sets. The overlaps from the CD4+ naive set were slightly higher than those from the CD8+ naive set. In general, overlaps were all within a narrow range for each of the T cell subsets, with values ranging from 8% to 16% (10.4% ± 0.7%) for the naive sets and 2% to 9% (5.3% ± 0.7%) for the memory cell sets. Because the number of unique TCRβ clonotype counts differed among subjects, we subsampled down to the size of the smaller of the two populations being compared to determine whether the percentage overlaps were largely influenced by differences in the sizes of the two populations. Such a situation might arise when two samples are sequenced to different depths. Subsampling resulted in a decrease in the percentage overlaps in both naive and memory subsets (Table S4). For the naive sets, the percentage overlaps ranged from about 1% to 5% (2.8% ± 0.2%), whereas the overlaps between the memory pools ranged from 1% to 3.5% for CD4+ (2.0% ± 0.3%) and 0.5% to 1.5% for CD8+ (0.9% ± 0.1%). We also determined the percentage overlaps among all four subjects in aggregate. The percentage overlaps for the CD4+ naive cell set belonging to subjects HIP2, HIP3, HIP4, and HIP5 was 5.0% (Figure 5A). The CD4+ memory set had a percentage overlap of about 2.3% (Figure 5B). The CD8+ naive or memory compartments had percentage overlaps of 3.5% or 0.4%, respectively (Figures 5C and 5D). Based on the pairwise subsampling results, we would expect the degree of sharing among all donors to be affected by the sequencing depth.

We also wanted to determine whether our deeper sequencing of the naive CD8+ T cell compartment from the peripheral blood of four healthy adults provided further insights into the degree of clonotype sharing when compared with earlier work using shallower sequencing (Robins et al., 2010). We estimated the number of overlaps in the naive CD8+ compartment by setting a lower-bound value of 10 million on the total number of possible TCRβ clonotypes in any donor. We arrived at this estimate by pooling together all clonotypes we obtained from sequencing of the naive CD8+ compartment from the blood of subjects HIP2, HIP3, HIP4, and HIP5 and determining the total number of unique clonotypes. Our lower-bound estimate is a factor of 10 larger than that used previously (Robins et al., 2010), which makes sense, because our sequencing was deeper. For example, we obtained more than 1 million clonotypes when sequencing the CD8+ cell compartment in each of our subjects, suggesting that a lower-bound value of 1 million would underestimate the size of this compartment. Using a power law fit, we obtained estimates for overlap counts between subject HIP2 and all other subjects that ranged from 395,175 to 562,041

clonotypes (Figure S4B). These overlap count values are higher than those reported previously (Robins et al., 2010), in which estimates of overlap counts ranged from 11,878 to 16,507 TCRβ CDR3s (or percentage overlaps of 1% to 1.7%). Our overlap counts at a cutoff of 1 million (the value suggested by Robins et al., 2010, to be the total number of possible unique TCRβ CDR3s in the CD8+ compartment) range from 73,859 to 105,219. This number is roughly 6 times larger than the values obtained by the early work of Robins et al. (2010), but our sequencing was deeper, making the likelihood of overlaps larger. To summarize, we see percentage overlaps of about 3.5% from the naive CD8+ cell compartment from the blood of four subjects at this depth of sequencing (Figure 5C). We also estimate percentage overlaps from the naive CD8+ compartment to be about 3.95% to 5.62%, assuming the total possible number of clonotypes in any donor is at least 10 million. Altogether, these results imply that any two random individuals can be expected to share a maximum of about 6% of their TCRβ V3J clonotypes from the CD8+ T cell population in the blood.

Although we observed higher TCRβ clonotype sharing among donors than what would be expected to occur by chance alone (Figures 2C–2D and 5), we also sought to determine whether the somatic recombinations encoded in those shared TCRβ clonotypes have higher probabilities of recombining than do non-shared clonotypes. To assess the likelihood of a TCRβ recombination, we computed the generation probability (P_{gen}) using OLGA (Sethna et al., 2019) for all TCRβ clonotypes obtained with the gDNA method for donors HIP2, HIP3, HIP4, and HIP5. OLGA uses a computational model to define the probability of any nucleotide sequence being generated as the sum of the probabilities of the generative events that make up somatic recombination. We found that clonotypes shared by all four donors had higher P_{gen} when compared with clonotypes not shared by all four donors (Figures S5A–S5D). The differences in median values between P_{gen} probabilities for shared and those for not shared was similar across cell sets.

TCRβ Clonotype Sharing with the Emerson Dataset

We also looked at the degree of sharing between our sequenced datasets and those from another large-scale study. In the Emerson et al. (2017) study, 666 healthy bone marrow donors were immunosequenced and generated roughly 90 million unique TCRβ sequences, with an additional 120 subjects that contained roughly 203,000 unique TCRβ sequences used as controls. The dataset from Emerson et al. (2017) has two important features that make it useful for analysis with our dataset. First, this dataset has high statistical power. Second, the dataset uses the Adaptive Biotechnologies gDNA method for sequencing, which avoids some technological biases with the mRNA method employed here. We downloaded and processed the data for all 786 subjects in the same manner as was done for our gDNA sequencing analysis and then determined the percentage overlaps between each of our sorted datasets and the sequencing from the Emerson set. We observed median percentage overlaps between 7% and 17% for CD4+ naive sets (Figure S5E). The CD4+ memory cell sets had median percentage overlaps ranging from about 5% to 10% (Figure S5F). The CD8+ naive and memory cell sets had median percentage overlaps ranging from about 7% to 12% (Figure S5G) and 2% to 5% (Figure S5H), respectively. Subsampling reduced the percentage overlaps to values between 0.5% and 1.5% for both cell types (see smaller plots in each

panel in Figures S5E–S5H). Subsampling consistently resulted in a reduction in the degree of clonotype sharing between our HIP datasets and those from the Emerson dataset. However, the sharing remained around 0.5% to 1% and is consistent with what we observed among our individual HIP datasets (see Table S4). This finding would seem to imply that even at shallower sequencing, we should see percentage overlaps that are larger than what would be expected by chance alone.

Sequence Processing with MiXCR Yields Higher Sharing of V3J Clonotypes

We also used the MiXCR pipeline (Bolotin et al., 2015) to process all TCRβ mRNA sequencing for subjects HIP1, HIP2, and HIP3. We processed the TCRβ mRNA sequencing using MiXCR in two separate modes. One mode performed the assembly of clonotypes based solely on the CDR3 region, whereas the second mode focused the assembly of clonotypes over the entire VDJ region. The percentage of overlapping V3J clonotypes was similar between the CDR3 method of assembly (Figure S5I, left panel) and the VDJ method of assembly (Figure S5I, right panel). Both methods yielded a value of 5% for overall sharing among HIP1, HIP2, and HIP3. The degree of sharing obtained with the MiXCR pipeline was higher when compared with our method that yielded a value of 2% for overall sharing (Figure S2A).

Searching for V3J Clonotypes in GenBank

We next sought to determine whether the V3J clonotypes from our deep sequencing contained biological signatures from past disease exposures. We used data from public repositories containing functionally characterized paired TCR sequences for comparison. Previous studies have shown that mining next-generation sequencing of B cell repertoires can lead to discovery of sequences with a high degree of similarity to functionally known antibodies (Kovaltsuk et al., 2018; Krawczyk et al., 2019). Here, we searched GenBank (Clark et al., 2016) for TCRβ V3J clonotypes that matched to those in our HIP subjects. We used an exact matching criterion that required the V germline gene, J germline gene, and CDR3 amino acid sequences to be identical between clonotypes from GenBank and those sequenced from our HIP subjects. We found 557 TCRβ clonotype matches in GenBank, with 134 matches associated with patented sequences, 49 matches associated with recognition of epitopes from viral pathogens, 8 matches associated with bacterial pathogens, 249 matches associated with autoimmune disorders, 66 matches associated with cancer, and 51 matches associated with other diseases (Figure 6A). We selected full-length sequences from a handful of these matches and generated alignments between the sequences from GenBank and those obtained from our HIP subjects (Figure 6B). The sequences derived from gDNA sequencing were not of full length, and for clarity, we used amino acids from the closest-matching germline gene to fill in the missing framework regions. The alignments show TCRβ chains that bind to epitopes specific to viral antigens, such as herpes simplex virus 2, HIV gag protein, and influenza hemagglutinin (Table S5). Thus, it appears possible to use the GenBank repository to infer possible specificities for several TCRβ clonotypes derived from the HIP subjects.

DISCUSSION

The identification of high frequencies of shared elements in human TCR repertoires is interesting. The application of high-throughput DNA amplicon sequencing of adaptive immune receptor gene repertoires to zebrafish (Weinstein et al., 2009) and shortly thereafter to humans (Arnaout et al., 2011; Boyd et al., 2009; Briney et al., 2012) and other species revealed enormous diversity in these expressed BCR and TCR genes. Studies of TCR V gene lineages increasingly are reported for particular infectious agents, including cytomegalovirus (Alanio et al., 2015; Dash et al., 2017; Emerson et al., 2017; Pogorelyy et al., 2018), Epstein-Barr virus (Dash et al., 2017; Glanville et al., 2017), HIV (Heather et al., 2016), hepatitis C and influenza virus (Dash et al., 2017), yellow fever virus (17D vaccine) (Pogorelyy et al., 2018), and *M. tuberculosis* (Glanville et al., 2017). New computational tools based on structure-based concepts now can group TCRs of common specificity from different donors, and conserved TCR CDR3 motifs can be associated with contact points with the antigenic peptides presented in the context of major histocompatibility complex (MHC) molecules (Dash et al., 2017; Glanville et al., 2017). TCR repertoire response will be of special interest for those who are developing DNA or virus-vectored vaccines for chronic infections that are designed to elicit immunity by induction of cytotoxic T cells. Alterations in TCR repertoire during autoimmune conditions are of intense interest for biomarker and mechanism-of-disease studies, for example, in studies of rheumatoid arthritis (Sakurai et al., 2018) and inflammatory bowel diseases (Werner et al., 2019). For each of these target- or disease-specific TCR studies, large comprehensive datasets of the otherwise healthy TCR repertoire are needed for comparative purposes. Here, we provide such a dataset as a resource for those studies. Deeper understanding of the particular recombined TCRs that are shared frequently in the human population could help us in future studies to understand the variability in immune response of diverse subjects to vaccination, infection, or immunotherapy for cancer. Indeed, using computational V(D)J generative models, we found that shared TCR clonotypes have higher probabilities of being generated through somatic recombination than those appearing in a subset of a population. Targeting highly shared TCR clonotypes by identifying the critical peptides that stimulate those TCRs could be an important approach in future research to develop vaccine formulations that are more broadly effective in diverse populations. Detailed TCR clonotype studies should lead to mechanistic insights into the clonotypes that mediate autoimmune responses or that are reduced or missing in defective immune responses to malignancy.

The datasets we provide here have limitations for some studies of the human TCR immunome. These repertoires were obtained from PBMCs, but it is likely that the TCR repertoire in tissues differs from that in blood. Despite the extreme depth at which we sequenced repertoires here, the number of individuals is small; thus, the diversity of genetics and exposure history in these subjects is limited. To account for this size limitation, we included in extended analyses a published TCR sequencing dataset from the largest cohort of individuals to date. Ignoring the differences in sequencing depth (i.e., sample sizes), we observed a high degree of sharing with the Emerson dataset (as high as 25%). A subset of TCRβ clonotypes that appear in all of our HIP subjects also appear in more than 617 of 778 samples from the Emerson dataset. Chu and coworkers discovered these same clonotypes

and dubbed these persistent TCRβ clonotypes that appeared at multiple time points of a longitudinal sequencing study (Chu et al., 2019). We observed these same persistent TCRβ clonotypes in both gDNA and mRNA sequencing datasets. As more TCRβ immunosequencing becomes available, we expect that even more of these persistent TCRβ clonotypes will become apparent.

To obtain the depth of sequencing required to provide an accurate estimate of overall repertoire size, in the work we present here, it was necessary to use bulk sequencing of unpaired receptor proteins. Despite this limitation, we can infer likely specificity for many interesting TCRs that were observed in this study. We searched through the GenBank repository to find previously reported clonotypes that are identical to those in our experimental repertoires. Many clonotypes for which we found matches have been validated functionally. Matching clonotypes appear in diverse patents and many published academic studies, especially cancer-related studies. Comparing these data with those from studies of single T cell transcriptomics with linked TCRα- and TCRβ-chain repertoires will be of interest for achieving a deeper understanding of human immune responses. Nevertheless, the depth of sequencing we achieved here, along with the observation of the high proportion of shared clonotypes, make these data useful as a comparator set for ongoing and future TCR studies of targeted disease states that are being performed at a shallower sequencing depth.

STAR★METHODS

RESOURCE AVAILABILITY

Lead Contact—Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, James E. Crowe (james.crowe@vumc.org).

Materials Availability—This study did not generate new unique reagents.

Data and Code Availability—Raw sequencing data derived from mRNA sequencing has been deposited in the Sequence Read Archive (SRA). The accession number for the sequencing reported in this paper is SRA: PRJNA511481. Genomic DNA (gDNA) sequencing data that could not be deposited in the SRA has been deposited to Adaptive Biotechnologies: https://doi.org/10.21417/CS2020CR. We also used publicly available TCRβ sequencing from Emerson et al. (2017) that is available from Adaptive Biotechnologies: https://doi.org/10.21417/B7001Z. Synthetic datasets generated using IGoR (Marcou et al., 2018) can be downloaded from https://doi.org/10.6084/m9.figshare.c.5002037.v1. GenBank sequences (release 231) were downloaded from GenBank: ftp://ftp.ncbi.nlm.nih.gov/genbank. PyIR can be found at https://github.com/crowelab/PyIR. All Python scripts used to analyze the sequencing data can be found at https://github.com/crowelab/TCRBmanuscript. Germline gene sequences can be downloaded from IMGT (Lefranc and Lefranc, 2001) and can be found at IMGT: http://www.imgt.org/vquest/refseqh.html#references.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

We leukapheresed five healthy HIV-negative adult subjects with no recently reported acute infections or vaccinations. Adult samples were obtained after informed consent from the Vanderbilt Clinical Trials Center. The study was approved by the Institutional Review Board of Vanderbilt University Medical Center (VUMC). The subjects consisted of two adult females (designated as subject HIP1 or HIP5) and three adult males (designated as subjects HIP2, HIP3 or HIP4) (Table 1). HLA typing (Mack et al., 2013; Robinson et al., 2003) was carried out for all HIP subjects (Table S2). Leukopaks containing large numbers of PBMCs obtained by leukapheresis were collected from all subjects at VUMC. Following leukapheresis, peripheral blood mononuclear cells (PBMCs) were isolated with Ficoll-Histopaque by density gradient centrifugation and cryopreserved in multiple aliquots containing 1×10^7, 2×10^7, 5×10^7, 1×10^8 or 2×10^8 cells in each cryovial in a one mL volume. The cells were cryopreserved in the vapor phase of liquid nitrogen until use.

METHOD DETAILS

Cell Sorting—For human subjects HIP2, HIP3, HIP4 and HIP5 subsets of naive and memory CD4+ and CD8+ T cells were obtained by magnetic-activated cell sorting (MACS) prior to gDNA extraction and immunosequencing using the Adaptive Biotechnolgies Immunosequencing TCRβ kit. CD4+ were first bead isolated by negative selection (Miltenyi) then the CD45RA Naive CD4s were isolated by positive selection with the negative fraction including memory cells. CD8 subset include a large fraction of CD8+ effector memory cells that are CD45RA+ (TemRA) which are highly clonal and if included as “naïve” would skew the diversity in the naive fraction. Therefore, after negatively selecting CD8+ T cells, we needed two steps to isolate the naive from the memory cells. We first positively selected the CD45RO+ fraction. The negative fraction included CD8+ TemRA cells and CD8+naive cells. We then used CCR7-PE antibody (Miltenyi) and an anti-PE magnetic bead conjugated antibody (Miltenyi) to positively select the naive cells from the CCR7negative TemRA fraction. The TemRA cells were combined with the CD45RO fraction for CD8 memory TCR sequencing and the CD45RA+CCR7+ was naive. All cell sorted populations were assessed for purity and quantity using analytical flow cytometry. A summary of resulting enriched T cell fractions is provided in Table S1.

Genomic DNA (gDNA) sequencing—For human subject HIP1, approximately 9×10^7 PBMCs were counted by hemacytometer prior to gDNA extraction using the QiaAmp DNA Blood Mini kit (QIAGEN). For human subjects HIP2, HIP3, HIP4 and HIP5, PBMCs were counted by hemacytometer prior to MACs enrichment of naive and memory CD4+ and CD8+ T cells. A summary of resulting enriched T cell fractions is provided in Table S1. Following MACs enrichment, gDNA was isolated from each cell fraction and quantitated using absorbance of UV-visible light on a Nanodrop 2000c (ThermoFisher Scientific). All extractions were performed on separate days and care was taken ensure no cross-contamination between replicate samples or human subject samples. All extracted gDNA from each cell subset per human subject, including gDNA extracted directly from PBMCs, was aliquoted evenly into all wells across 9 to 10 Adaptive ImmunoSEQ TCR α/β deep assay kit 96-well plates according to the manufacturer’s recommendations. Specifically, a total of 200 ng of gDNA corresponding to approximately 30,000 T cell genomes from sorted

T cell subsets (> 90% T cell content after MACs enrichment) entered the PCR1 reaction per replicate in a 96-well replicate plate. In cases where MACs separation was not performed, a total of 300 ng of gDNA corresponding to approximately 30,000 T cell genomes from PBMCs (40%–70% typical T cell content) entered the PCR1 reaction per replicate in a 96-well replicate plate. For subjects HIP2, HIP3, HIP4 and HIP5 three Adaptive ImmunoSEQ TCR α/β deep assay plates were used for the CD4+ naive subset, three Adaptive ImmunoSEQ TCR α/β deep assay plates for the CD4+ memory subset, two Adaptive ImmunoSEQ TCR α/β deep assay plates for the CD8+ naive subset, and one Adaptive ImmunoSEQ TCR α/β deep assay plate for the CD8+ memory subset. For subject HIP1, ten Adaptive ImmunoSEQ TCR α/β deep assay plates were used for bulk PBMCs. TCR α/β CDR3 regions were amplified and sequenced using the deep Adaptive ImmunoSEQ assay in a multiplexed PCR using proprietary primers specific to TCR α/β and J α/β gene segments (Adaptive Biotechnologies). Purified libraries were quantitated using the Qubit 3.0 fluorometer (Thermo Fisher Scientific) prior to size determination using a Bioanalyzer 2100 (Agilent). The final libraries were re-quantified using the KAPA Library Quantification kit (Roche) and sequenced according to the manufacturer's recommendations on an Illumina NextSeq instrument at the VANTAGE core laboratory at Vanderbilt University Medical Center. Each Adaptive Immunoseq TCRβ 96-well kit was dedicated to a single Illumina SR-150 flow cell and sequenced using NextSeq 550.

mRNA sequencing—The mRNA sequencing method, performed by AbHelix LLC (http://www.abhelix.com/, South Plainfield, NJ, USA), was only utilized for human subjects HIP1, HIP2, and HIP3. For each of these human subjects, approximately 9×10^8 PBMCs were counted by hemacytometer and aliquoted into 5 or 6 biological replicates prior to total RNA extraction using the RNAeasy Maxi kit (QIAGEN). All extractions were performed on separate days and care was taken ensure no cross-contamination between replicate samples or human subject samples. Purified total RNA was shipped and processed at AbHelix, LLC. The AbHelix assay is designed to sequence 5 chains targeting B cell receptors (IgG, IgM, IgA, IgK and IgL) and 2 chains targeting T cell receptors TCRα and TCRβ. The total RNA was divided evenly per B or T cell receptor chain type, so only 2/7 of the total RNA provided was utilized for TCRα and TCRβ sequencing. The data from the B cell sequencing of HIP1, HIP2 and HIP3 at AbHelix was used in a separate but similar study (Soto et al., 2019). For the T cells assays used here, total RNA samples were reversed transcribed using the oligo d(T)18 in 3–5 ug per 20 ul reaction (SuperScript IV Reverse Transcriptase, ThermoFisher, CA). Multiple reactions of reverse transcription were combined per biological replicate and purified using magnetic beads. The purified RT products were divided evenly for the first round of PCR amplification specific to human TCRβ and TCRα. The 5′ multiplex PCR primers are designed within the leader sequences of each productive V-gene and the 3′ primers within the constant regions but in close approximation to the J-C junctions. The resulting 1st PCR products were purified with magnetic beads and subject to the second round of PCR amplification to add Illumina index and adaptor sequences. The resulting PCR products were purified with magnetic beads and pooled for sequencing with PE 2×250 on an Illumina HiSeq 2500. Phusion High-Fidelity DNA Polymerase (ThermoFisher, CA) was used in all PCR amplification reactions and care was taken to minimize the number of cycles to achieve adequate amplification. Purified libraries were

quantitated using the Qubit 3.0 fluorometer (Thermo Fisher Scientific) prior to size determination using a Bioanalyzer 2100 (Agilent). Final libraries were quantified using the KAPA Library Quantification kit (Roche) before sequencing on an Illumina HiSeq 2500. A single PE-250 flow cell was dedicated to each biological replicate from each human subject (a total of 5–6 biological replicates and corresponding flow cells per human subject).

Processing of next-generation sequencing (NGS)—The bioinformatics processing of all NGS data was done using our PyIR sequence processing pipeline (https://github.com/crowelab/PyIR) with sample and data management performed using our in-house proprietary laboratory information management system (LIMS). An outline of the process is provided below:

1. *Quality inspection and generation of full length reads.* Paired end (PE) reads generated using the Illumina sequencing platform were assessed using the FASTQC (Andrews, 2012) toolkit (version 0.11.6) to assess of the quality of the run. For those sequencing runs where the average Phred score was 20 or greater by inspection of the FASTQC base sequence quality plot, we proceeded to merge PE reads.

2. *Merging PE reads from mRNA sequencing to generate full-length contigs.* PE reads were merged using the program USEARCH v9.1 (Edgar and Flyvbjerg, 2015) . The overlap region (-*fastq_minovlen*) was set to 15 nucleotides and the maximum number of differences in the overlap region (-*fastq_maxdiffpct*) was set to 10. Data obtained from Adaptive Biotechnologies did not require merging of PE reads.

3. *Gemline gene assignment and definition of CDR3 regions.* We used PyIR, a Python wrapper for IgBLAST v1.9 (Ye et al., 2013), to process the TCR repertoires. PyIR parses IgBLAST output to obtain germline gene assignments, nucleotide sequence for the variable region (V(D)J segment) and CDR3 regions. TCR germline gene sequences were downloaded from IMGT (Lefranc and Lefranc, 2001) (http://www.imgt.org/vquest/refseqh.html#references)

4. *Filtering reads using MongoDB.* Processed reads were imported into our MongoDB database for quality control filtering. Each read was subjected to a series of knowledge-based sieves in the following order: (1) Removal of any read that had an E-value larger than 10^{-6} for TCRβV/TCRβJ germline assignments; (2) Removal of any read that did not have a defined CDR3; (3) Removal of any read containing a stop codon; (4) Removal of any read that was out of frame at the junction region; and (5) Removal of any read that contained an ‘X’ after nucleotide translation to the amino acid sequence. All remaining reads were considered high-quality and labeled as “productive.”

5. *De-replication of productive reads using MongoDB.* To remove redundancy in the data, multiple copies of the same V(D)J segment were removed from the productive pool of sequences. The number of times a V(D)J segment appeared in the dataset was retained and stored in the database. This process of de-replication

effectively reduced the size of the dataset that needed to be integrated into the MongoDB database.

The sequence data obtained from Adaptive Biotechnologies contained only the junctional region of the read with a few nucleotides from the 3′ region of framework 3 and 5′ region of framework 4. This method did not require merging of PE reads and did not provide associated quality scores and thus could not be vetted in the exact same way as the full-length sequencing.

Clonotype definitions—We defined a ***"V3J clonotype"*** by the amino acid sequence of the CDR3 along with the V and J germline gene assignment. If two sequences were encoded by the same inferred V and J genes and had the same CDR3 amino acid sequence, they were considered the same V3J clonotype. For assessing the significance of the amount of clonotype sharing between HIP donors, we used an alternate definition of clonotype that included the diversity (Dβ) germline gene assignment for those sequences where a D gene assignment could be made with reasonably high confidence (see below). When an explicit Dβ germline assignment could be made, we used the combination of the V, D, J gene and an identical CDR3 amino acid sequence to define ***"V3DJ clonotypes"***. In some cases, as indicated in Figure S2D, we grouped sequences with matching V, D, and J gene assignments, regardless of CDR3 sequence, to establish groups termed ***"VDJ triples"***. Finally, in Figures S3A and S3B we show the distribution of V3J clonotypes using heatmaps that only consider the Vβ and Jβ gene assignments (***"VJ heatmap"***).

Defining high-confidence Dβ germline genes—Dβ gene segments are shorter than either Vβ or Jβ germline genes, making high confidence inferred Dβ gene assignments challenging due to exonuclease trimming at the 5′ and 3′ ends of the gene segments. We set the expectation value (E-value) threshold to 10^{-1} for assigning Dβ germline genes to productive reads from the sequenced repertoires. We note that setting the E-value threshold to 10^{-1} resulted in about a 75% reduction in the number of V3J clonotypes when compared to V3DJ clonotypes. However, the remaining population of experimental V3J clonotypes with Dβ gene assignments all had high confidence matches.

CDR3 length and cumulative distributions—The CDR3 length distributions from each subject were determined from the corresponding distributions of unique clonotypes. CDR3 length distributions for mRNA and gDNA datasets belonging to HIP1, HIP2 or HIP3 were represented as Box-and-Whisker plots (Figure 3A). CDR3 length distributions for HIP1, HIP2, HIP3, HIP4 or HIP5 were represented as Box-and-Whisker plots (Figure S4A). Normalized frequency CDR3 length histograms were constructed from V3DJ clonotypes belonging to HIP1, HIP2 or HIP3 (Figure S2C). Normalized CDR3 length histograms were constructed from V3J clonotypes from TCRα sequencing (Figure S2E). Cumulative distributions were constructed using the number of somatic variants associated with a V3J clonotype (Figure 3C) or the number of VDJ triples (Figure S2D).

Percentage overlaps between repertoires—To determine the percentage of clonotypes being shared between subjects, we searched for exact matching clonotypes between subjects. The percentage overlap was defined as the total number of unique

clonotypes shared between donors divided by the size of the smallest population of clonotypes between the donors being compared. A similar definition of overlap has been used by Greiff et al. (2017). All percentage overlaps were rounded to the nearest integer. Percentage overlaps less than 1% were rounded to the nearest decimal place. Percentage overlap calculations were used in figures (Figures 2, 3D, 4, 5, S2A, and S5I).

Generating synthetic repertoires using IGoR—Three independent sets of simulations (one per HIP subjects HIP1, HIP2 or HIP3) were carried out using IGoR (Marcou et al., 2018) to generate billions of synthetic V3DJ clonotypes. Only those VDJ triples that appeared in each experimentally derived repertoire were simulated. In total, we generated 1.2×10^9, 0.72×10^9 and 1.1×10^9 unique synthetic V3DJ clonotypes for simHIP1, simHIP2 and simHIP3 respectively. Clonotypes from each of these distributions were subsampled based on the and CDR3 length distributions (Figure S2C) and VDJ triple frequencies (Figure S2D) observed in the experimentally determined HIP repertoires. A total of 500 synthetic repertoires were generated for each HIP subject through subsampling of the larger synthetic repertoires. There were some TCRβ variable (Vβ) germline genes that we did not consider for overlap comparisons since IGoR could not generate a sufficiently large enough set. The following set of 425 VDJ triples appearing in HIP1 were not considered: TRBV5–6_TRBD2_TRBJ2–6,TRBV5–6_TRBD1_TRBJ2–6,TRBV5–7_TRBD1_TRBJ1–6,TRBV5–8_TRBD2_TRBJ2–6,TRBV5–8_TRBD1_TRBJ2–6,TRBV11–2_TRBD2_TRBJ1–2,TRBV13_TRBD1_TRBJ2–4,TRBV11–1_TRBD2_TRBJ2–2,TRBV11–2_TRBD2_TRBJ1–3,TRBV4–3_TRBD2_TRBJ2–4,TRBV28_TRBD1_TRBJ2–4,TRBV3–1_TRBD1_TRBJ2–6,TRBV10–2_TRBD1_TRBJ1–1,TRBV11–2_TRBD1_TRBJ1–4,TRBV6–6_TRBD2_TRBJ1–3,TRBV11–1_TRBD1_TRBJ1–6,TRBV6–6_TRBD2_TRBJ2–4,TRBV11–2_TRBD1_TRBJ2–5,TRBV6–1_TRBD2_TRBJ1–4,TRBV11–1_TRBD2_ TRBJ2–7,TRBV11–1_TRBD1_TRBJ2–7,TRBV7–8_TRBD2_TRBJ2–4,TRBV5–8_TRBD2_TRBJ2–2,TRBV11–3_TRBD1_TRBJ2–4,TRBV4–1_TRBD1_TRBJ1–3

The following set of 42 VDJ triples appearing in HIP2 were not considered: TRBV11–2_TRBD1_TRBJ1–4,TRBV5–8_TRBD1_TRBJ2–6,TRBV15_TRBD1_TRBJ2–5,TRBV5–6_TRBD2_TRBJ2–6,TRBV3–1_TRBD2_TRBJ1–6,TRBV5–7_TRBD1_TRBJ1–5,TRBV5–7_TRBD1_TRBJ1–6,TRBV4–1_TRBD1_TRBJ1–3,TRBV6–9_TRBD2_TRBJ2–3,TRBV3–1_TRBD1_TRBJ2–6,TRBV11–3_ TRBD2_TRBJ2–6,TRBV11–2_TRBD2_TRBJ1–2,TRBV5–8_TRBD2_TRBJ2–6,TRBV7–8_TRBD1_TRBJ1–3,TRBV24–1_TRBD2_TRBJ2–4,TRBV6–6_TRBD2_TRBJ2–4,TRBV11–2_TRBD1_TRBJ2–5,TRBV11–1_TRBD1_TRBJ1–5,TRBV11–1_TRBD1_TRBJ1–6,TRBV13_ TRBD1_TRBJ1–2,TRBV2_TRBD2_TRBJ2–4,TRBV6–5_TRBD1_TRBJ2–4,TRBV28_TRBD1_TRBJ2–4,TRBV12–2_TRBD1_TRBJ2–5,TRBV10–2_TRBD1_TRBJ1–1,TRBV18_TRBD2_TRBJ1–3,TRBV6–5_TRBD2_TRBJ2–1,TRBV28_TRBD2_TRBJ2–2,TRBV6–6_TRBD1_TRBJ1–3,TRBV11–1_TRBD1_TRBJ2–7,TRBV29–1_TRBD2_TRBJ2–6,TRBV5–6_TRBD1_TRBJ2–6,TRBV13_TRBD1_TRBJ2–4,TRBV7–8_TRBD2_TRBJ2–4,TRBV4–3_TRBD2_TRBJ2–4,TRBV11–1_TRBD2_TRBJ2–1,TRBV6–1_TRBD2_TRBJ1–

3,TRBV11–1_TRBD2_ TRBJ2–2,TRBV6–1_TRBD2_TRBJ1–4,TRBV11–3_TRBD1_TRBJ2–4,TRBV5–8_TRBD1_TRBJ2–1,TRBV11–1_TRBD2_TRBJ2–7

The following set of 42 VDJ triples appearing in HIP3 were not considered: TRBV5–6_TRBD1_TRBJ2–6,TRBV6–8_TRBD2_TRBJ1–5,TRBV5–7_TRBD1_TRBJ1–6,TRBV6–9_TRBD2_TRBJ2–3,TRBV13_TRBD1_TRBJ1–2,TRBV5–8_TRBD2_TRBJ2–6,TRBV5–8_TRBD1_ TRBJ2–6,TRBV28_TRBD2_TRBJ2–2,TRBV3–1_TRBD2_TRBJ1–6,TRBV11–2_TRBD2_TRBJ1–2,TRBV13_TRBD1_TRBJ2–4,TRBV11–1_TRBD2_TRBJ2–2,TRBV11–2_TRBD2_TRBJ1–3,TRBV4–3_TRBD2_TRBJ2–4,TRBV28_TRBD1_TRBJ2–4,TRBV12–2_TRBD1_TRBJ1–4,TRBV10–2_TRBD1_TRBJ1–1,TRBV11–1_TRBD1_TRBJ1–5,TRBV11–2_TRBD1_TRBJ1–4,TRBV6–6_TRBD1_TRBJ1–3,TRBV6–5_ TRBD1_TRBJ2–4,TRBV11–1_TRBD1_TRBJ1–6,TRBV6–6_TRBD2_TRBJ2–4,TRBV6–1_TRBD2_TRBJ1–3,TRBV7–8_TRBD1_TRBJ1–3,TRBV11–2_TRBD1_TRBJ2–5,TRBV6–1_TRBD2_TRBJ1–4,TRBV11–1_TRBD1_TRBJ2–7,TRBV7–8_TRBD2_TRBJ2–4,TRBV5–8_ TRBD1_TRBJ2–1,TRBV5–7_TRBD1_TRBJ1–3,TRBV2_TRBD2_TRBJ2–4,TRBV5–8_TRBD2_TRBJ2–2,TRBV11–3_TRBD1_TRBJ2–4,TRBV4–1_TRBD1_TRBJ1–3,TRBV6–2_TRBD1_TRBJ2–2

Processing mRNA sequencing data with MiXCR—TCRβ sequencing data from the mRNA sequencing method for HIP1, HIP2 and HIP3 were merged in a paired wise fashion. Duplicate reads were collectively removed using ‘clumpify’ from ‘BBMap’ package. For data analysis, MiXCR (version 2.1.10) (Bolotin et al., 2015) was used to align reads to the reference IMGT library, assemble them by turning off clustering (to remain consistent with PyIR), and eventually make clonotype calls. All the MiXCR runs were made on a compute cluster employing threading option to make use of 16 cores to satisfy the memory requirement for our large datasets. By default, MiXCR assembles the reads based on CDR3 region, however, we employed an alternate option as well by assembling the reads for entire VDJ region. As a result, we had two call sets for each HIP dataset. It is pertinent to mention that when using VDJ region to assemble the reads, MiXCR puts base quality score requirements over a wider region compared to when the CDR3 region is selected for assembly. Hence many more reads are dropped from VDJ assembly and it leads to smaller call set compared to when assembly is performed over CDR3 region. Finally, the non-functional clonotypes were removed by VDJTools (Shugay et al., 2015) and unique clonotypes were sorted to obtain a final call set. V3J clonotype percentage overlaps obtained by processing with MiXCR appear in Figure S5I.

TCRβ V3J clonotypes in GenBank—The entire GenBank (Clark et al., 2016) database (release 231) was downloaded from GenBank: ftp://ftp.ncbi.nih.gov/genbank/ and processed using PyIR. Only those sequences from GenBank with Vβ and Jβ matches and that passed our quality filtering were considered. We used all V3J clonotypes from HIP1, HIP2, HIP3, HIP4 or HIP5 to search through the processed set of GenBank V3J clonotypes to find exact matches (Figure 6; Table S5).

QUANTIFICATION AND STATISTICAL ANALYSIS

Constructing species richness curves with RTK—Species richness is a widely used concept in ecological sciences where the goal is to determine the diversity of species in a certain area or community. The most straightforward method for doing this analysis involves subsampling a population and then determining the number of unique species present in that subsampled population. We used the program RTK (Saary et al., 2017) to determine species richness by subsampling populations of clonotypes based on their total frequency of occurrence in high-quality reads. In order to use RTK, we first generated a list depth sizes based on the number of high-quality reads such that there would be 100 total depth sizes. Each depth size then was used to rarefy the total frequency of occurrence of each clonotype. The Chao1 estimator (Chao, 1987) was used to compute a lower-bound estimate of the total number of V3J clonotypes we should expect from the current sequencing. Calculation of the Chao1 estimator requires counts for the frequency of singleton (f_1) and doubleton V3J clonotypes (f_2) and the total number of observed clonotypes in the sample ($S_{observed}$):

$$S_{Chao1} = S_{observed} + \frac{{f_1}^2}{2f_2}$$

The variance of the Chao1 estimator was also determined using the formula from Chao (1987):

$$\sigma^2_{Chao1} = f_2 \times \left(0.5 \times \left(\frac{f_1}{f_2}\right)^2 + \left(\frac{f_1}{f_2}\right)^3 + 0.25 \times \left(\frac{f_1}{f_2}\right)^4\right)$$

Classical approximate 95% confidence intervals based on the Chao1 estimator (S_{Chao1}) were computed based on the assumption of normality [$S_{Chao1}-1.96\sigma_{Chao1}$, $S_{chao1}+1.96\sigma_{Chao1}$]. RTK also provides other alpha diversity measures for a sample population, however for this study we considered only the species richness measure and the Chao1 estimator. The command line used with RTK was*: rtk memory* -i < *total frequency of occurrence of clonotypes file* > -d< *depth size* > -w1 < *create on table* > -t 10< *threads to use* >. Since one table was created for each depth size, we concatenated all of the tables and then sorted by depth size. The species richness values and Chao1 estimates were based on the average of 10 independent runs. We denoted the largest depth value as the endpoint estimate on both the species richness curves and the Chao1 estimate curves. RTK was used to subsample V3J clonotypes from TCRβ chains based on their frequency of occurrence in productive reads and appear in Figures 1A and S1A. Chao1 estimates were computed using the program RTK and appear in Figure S1B and Table S3. All species richness curves were generated using ORIGIN(Pro) (version 2018b).

Determining missing clonotype counts with Recon—We used the program Recon (Kaplinsky and Arnaout, 2016) (version 2.1) to determine the number of clonotypes that were unaccounted for in the TCRβ repertoires belonging to HIP1, HIP2, HIP3, HIP4 or HIP5. Recon was run in two steps. The first step generated a set of best fit parameters. The second step provided an estimate for the number of “missing species” or clonotypes. In the

first step, the *–run_recon* and*–clone_distribution_in_file* options were used, with the second option requiring a tab delimited file containing the clonotype group size bin and the frequency of clonotypes occurring in the pool of productive reads for a particular group size bin. The best-fit parameters obtained in the first step along with default parameter settings that included the *error_bar_params.txt* file supplied with the program were then used to calculate an estimate of the number of V3J clonotypes unaccounted for in our sequencing.

Power law model fits using naive CD8+ data—We used a non-linear regression model similar to that employed by Robins et al. (2010) to estimate the total number of V3J TCRβ clonotypes shared between the naive T cell CD8+ compartments between two subjects. We fit the observed overlap data with the same two parameter model used by Robins et al. (2010):$Y= aX^b$ where a and b are the parameters to be estimated, X is the input variable and Y is the number of overlapping V3J TCRβ clonotypes between two repertoires. In order to generate input data for X, we first sorted in descending order the number of times the V3J TCRβ clonotype appeared in the pool of deduplicated and productive sequences. Then assuming a lower-bound of 10 million possible V3J clonotypes in the CD8+ naive compartment for each donor, we created a scale from 0 to 10 that could be incremented in units 0.01. To determine the number of sequences per donor that would be represented with this scale, we used the following formula:

$$N = 10^{(\log_{10} i + 13)/2}$$

Where i represents the scaled values along the x axis in Figure S4B starting from a value of 0.01 and N represents the top number of V3J clonotypes to be considered for overlap analysis from the two sorted datasets. For example, a value i = 1.0 would denote the overlaps from the top 3,162,278 V3J clonotypes from each of two subjects. The value of 13 is required so that we arrive at a value of 10 on the x axis when considering the overlaps at a value of N = 10,000,000 V3J clonotypes. For the case where N = 10,000,000, i = 10.0 and we would arrive at:

$$N_1 \times N_2 \times 10^{-13} = 10^{(\log_{10} 10 + 13)/2} \times 10^{(\log_{10} 10 + 13)/2} \times 10^{-13} = 10^{14} \times 10^{-13} = 10$$

on the x axis scale denoting the estimated number of overlaps from the entire naive CD8+ compartments of two subjects. See the x axis legend in Figure S4B for estimated values. All fits were done in Python (Van Rossum and Drake, 1995) (version 2.7.12) using the Numpy module (version 1.13.3). All fits had r^2 values better than 0.99. The estimated overlaps were based on the fits and appear in Figure S4B.

Standardization of sample sizes—In order to standardize the sample size between two TCRβ V3J clonotype populations, we adopted the method from Venturi et al. (2008). Briefly, in our rendition of their method we first pooled both populations of clonotypes together into a linear arrangement and then randomly shuffled the linear arrangement 100 times. This preconditioning of the linear arrangement was meant provide a starting randomized linear arrangement. The shuffled linear arrangement was then shuffled (or

permuted) 1,000 times each time selecting a subset of clonotypes of size M from one end of the linear arrangement and a second subset of clonotypes of size M from the opposite end of the linear arrangement. Each time, the percentage overlap was then computed between the two subsets. M represents the size of the smaller of the populations being compared (i.e., the number of unique V3J clonotypes). Subsampling and standardization were used in Table S4 and Figure S5.

Heatmap construction of Vβ+Jβ usage—The frequency of Vβ+Jβ usage for each subject was determined using unique V3J clonotypes. All raw counts were transformed into Z scores using the total frequency of unique Vβ+Jβ combinations for that donor. Any row of the heatmap with a sum of less than 10 counts and any pseudo-genes were excluded from consideration. Heatmaps were created in Python (Van Rossum and Drake, 1995) (version 2.7.12) using the Seaborn plotting module (Waskom et al., 2017) (version 0.8.1) and appear in Figures S3A and S3B.

Nonparametric statistics using R—The Kruskal-Wallis was used to determine if the median CDR3 lengths between HIP1-HIP5 were statistically different and was done using the R statistical package (R Core Team, 2014) (version 3.2.3) (Figure 3A). The Mann-Whitney test was used to determine if the median P_{gen} values generated using OLGA (Sethna et al., 2019) were statistically different between the *Shared* and *Not Shared* sets using the R statistical package (R Core Team, 2014) (version 3.2.3) (Figures S5A–S5F).

Morisita-Horn index of similarity—We computed the Morisita-Horn (Horn, 1966) index of similarity between Vβ+Jβ frequencies derived from mRNA and gDNA sequencing methods for HIP1, HIP2 or HIP3 (Figure 3B). The formula used for computing the Morisita-Horn index is:

$$C_{MH} \frac{2\sum X_{ij}X_{ik}}{\left[\left(\sum X_{ij}^2/N_j^2\right)+\left(\sum X_{ik}^2/N_k^2\right)\right]N_jN_k}$$

Where X_{ij} and X_{ik} are the number of individuals of species in sample j and sample k respectively. N_j and N_k are the total number of individuals in sample j or sample k respectively. We also computed the Morisita-Horn index between lanes of the same flow cell using V3J clonotype abundances for HIP1, HIP2 or HIP3 (Figure S3E).

Percentage overlaps for synthetic repertoires—We started with 500 random subsamples from each of the large synthetic V3DJ clonotype distributions. The number of unique synthetic V3DJ clonotypes was kept constant between subsamples and was similar to the number of experimentally observed V3DJ clonotypes (compare Figures 2C and 2B). There was a total of 125,000,000 three-way comparisons between the three subsampled datasets. For the sake of computational feasibility, we randomly selected 50,000 triplet combinations where each element of the triplet represents one of the three subsampled datasets. For each randomly selected triplet combination, we determined the pairwise percentage overlaps (simHIP1 with simHIP2, simHIP1 with simHIP3 or simHIP2 with simHIP3) and the total number of shared synthetic V3DJ clonotypes between all three sets

(simHIP1 with simHIP2 with simHIP3). To obtain precision measures, we computed the mean and SEM of the percentage overlaps for 50,000 comparisons (Figure 2D). We restricted our overlap analysis involving synthetic V3DJ clonotypes to CDR3 lengths from 3 through 19 amino acids.

P value estimation for repertoire overlaps—We used a method similar to that of Arnaout et al. (2011) to estimate the significance of the overlaps between synthetic repertoires simHIP1, simHIP2 and simHIP3. Five hundred synthetic repertoires were subsampled (with replacement) from each of the synthetic repertoires. We then used a total of 50,000 overlap comparisons between synthetic repertoires to determine the average percentage overlap (and SEM). To estimate the P value associated with the overlaps between synthetic repertoires, we ranked the overlap count between synthetic repertoires with the corresponding overlap counts between the experimentally determined repertoires (Figure 2D).

Generation probabilities using OLGA—We downloaded the OLGA package (version 1.1.0) (Sethna et al., 2019) from https://github.com/statbiophys/OLGA. To compute the generation probabilities (P_{gen}) for *Shared* and *Not Shared* V3J clonotypes (Figures S5A–S5H), we ran the *compute_pgen.py* script using the default setting where we explicitly defined the location of the V and J TCRB germline gene locations on the command line using the *–v_in* and *–j_in* command line switches.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Merissa Mayo and Ardina Pruijssers for regulatory and human subjects support. We thank Gopal Sapparapu and Olivia Koues for technical help and advice on experiments. We thank scientists at the VANTAGE core of Vanderbilt University Medical Center, Adaptive Biotechnologies, and Douglas Zhang and team at AbHelix (now Novogene), where DNA sequence analysis experiments were performed. We thank Tong Jin and Jimmy Quach of the San Diego Supercomputer Center and University of California, San Diego, for assistance with generating some synthetic repertoires. We thank Karen Trochez and Jill Janssen of the Clinical Trials Center at Vanderbilt University Medical Center and staff and physicians of the Vanderbilt University Medical Center leukapheresis clinic for assistance with large-scale human cell collections. We thank Richard Scheuermann (JCVI), Wayne Koff, Ted Schenkelberg, and the Advisory Board of the Human Vaccines Project for helpful discussions. We thank Katie Boland, David Hamm, and staff members from Adaptive Biotechnologies for helpful discussions involving their TCRβ assay. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, Nashville, Tennessee. This work also used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF grant ACI-1548562, and specifically the Comet supercomputer at the San Diego Supercomputer Center, which is supported by NSF grant ACI-1341698. This work was supported by a grant from the Human Vaccines Project and institutional funding from Vanderbilt University Medical Center. The authors acknowledge support from TN-CFAR grant P30 AI110527.

REFERENCES

Alanio C, Nicoli F, Sultanik P, Flecken T, Perot B, Duffy D, Bianchi E, Lim A, Clave E, van Buuren MM, et al. (2015). Bystander hyperactivation of preimmune CD8+ T cells in chronic HCV patients. eLife 4, e07916. [PubMed: 26568315]

Andrews S (2012). FastQC: A quality control tool for high throughput sequence data https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, Nusbaum C, Rajewsky K, and Koralov SB (2011). High-resolution description of antibody heavy-chain repertoires in humans. PLoS ONE 6, e22365. [PubMed: 21829618]

Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, and Chudakov DM (2015). MiXCR: software for comprehensive adaptive immunity profiling. Nat. Methods 12, 380–381. [PubMed: 25924071]

Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, Jones CD, Simen BB, Hanczaruk B, Nguyen KD, et al. (2009). Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci. Transl. Med 1, 12ra23.

Briney BS, Willis JR, McKinney BA, and Crowe JE Jr. (2012). High-throughput antibody sequencing reveals genetic evidence of global regulation of the naı¨ve and memory repertoires that extends across individuals. Genes Immun 13, 469–473. [PubMed: 22622198]

Briney B, Inderbitzin A, Joyce C, and Burton DR (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature 566, 393–397. [PubMed: 30664748]

Chao A (1987). Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43, 783–791. [PubMed: 3427163]

Chu ND, Bi HS, Emerson RO, Sherwood AM, Birnbaum ME, Robins HS, and Alm EJ (2019). Longitudinal immunosequencing in healthy people reveals persistent T cell receptors rich in highly public receptors. BMC Immunol 20, 19. [PubMed: 31226930]

Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, and Sayers EW (2016). GenBank. Nucleic Acids Res 44 (D1), D67–D72. [PubMed: 26590407]

Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, Crawford JC, Clemens EB, Nguyen THO, Kedzierska K, et al. (2017). Quantifiable predictive features define epitope-specific T cell receptor repertoires. Nature 547, 89–93. [PubMed: 28636592]

Edgar RC, and Flyvbjerg H (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics 31, 3476–3482. [PubMed: 26139637]

Emerson RO, DeWitt WS, Vignali M, Gravley J, Hu JK, Osborne EJ, Desmarais C, Klinger M, Carlson CS, Hansen JA, et al. (2017). Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. Nat. Genet 49, 659–665. [PubMed: 28369038]

Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, Ji X, Han A, Krams SM, Pettus C, et al. (2017). Identifying specificity groups in the T cell receptor repertoire. Nature 547, 94–98. [PubMed: 28636589]

Greiff V, Weber CR, Palme J, Bodenhofer U, Miho E, Menzel U, and Reddy ST (2017). Learning the High-Dimensional Immunogenomic Features That Predict Public and Private Antibody Repertoires. J. Immunol 199, 2985–2997. [PubMed: 28924003]

Heather JM, Best K, Oakes T, Gray ER, Roe JK, Thomas N, Friedman N, Noursadeghi M, and Chain B (2016). Dynamic Perturbations of the T-Cell Receptor Repertoire in Chronic HIV Infection and following Antiretroviral Therapy. Front. Immunol 6, 644. [PubMed: 26793190]

Hoi KH, and Ippolito GC (2013). Intrinsic bias and public rearrangements in the human immunoglobulin Vλ light chain repertoire. Genes Immun 14, 271–276. [PubMed: 23535864]

Horn HS (1966). Measurement of “Overlap” in comparative ecological studies. Am. Nat 100, 419–424.

Kaplinsky J, and Arnaout R (2016). Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. Nat. Commun 7, 11881. [PubMed: 27302887]

Kovaltsuk A, Leem J, Kelm S, Snowden J, Deane CM, and Krawczyk K (2018). Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. J. Immunol 201, 2502–2509. [PubMed: 30217829]

Krawczyk K, Raybould MIJ, Kovaltsuk A, and Deane CM (2019). Looking for therapeutic antibodies in next-generation sequencing repositories. MAbs 11, 1197–1205. [PubMed: 31216939]

Lefranc M-P, and Lefranc G (2001). The T cell receptor FactsBook (Academic Press).

Mack SJ, Cano P, Hollenbach JA, He J, Hurley CK, Middleton D, Moraes ME, Pereira SE, Kempenich JH, and Reed EF (2013). Common and well-documented HLA alleles: 2012 update to the CWD catalogue. Tissue Antigens 81, 194–203. [PubMed: 23510415]

Marcou Q, Mora T, and Walczak AM (2018). High-throughput immune repertoire analysis with IGoR. Nat. Commun 9, 561. [PubMed: 29422654]

Pogorelyy MV, Minervina AA, Touzel MP, Sycheva AL, Komech EA, Kovalenko EI, Karganova GG, Egorov ES, Komkov AY, Chudakov DM, et al. (2018). Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. Proc. Natl. Acad. Sci. USA 115, 12704–12709. [PubMed: 30459272]

Putintseva EV, Britanova OV, Staroverov DB, Merzlyak EM, Turchaninova MA, Shugay M, Bolotin DA, Pogorelyy MV, Mamedov IZ, Bobrynina V, et al. (2013). Mother and child T cell receptor repertoires: deep profiling study. Front. Immunol. 4, 463. [PubMed: 24400004]

R Core Team (2014). R: A language and environment for statistical computing http://www.R-project.org/.

Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, Carlson CS, and Warren EH (2010). Overlap and effective size of the human CD8+ T cell receptor repertoire. Sci. Transl. Med 2, 47ra64.

Robins H, Desmarais C, Matthis J, Livingston R, Andriesen J, Reijonen H, Carlson C, Nepom G, Yee C, and Cerosaletti K (2012). Ultra-sensitive detection of rare T cell clones. J. Immunol. Methods 375, 14–19. [PubMed: 21945395]

Robinson J, Waller MJ, Parham P, de Groot N, Bontrop R, Kennedy LJ, Stoehr P, and Marsh SG (2003). IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. Nucleic Acids Res 31, 311–314. [PubMed: 12520010]

Saary P, Forslund K, Bork P, and Hildebrand F (2017). RTK: efficient rarefaction analysis of large datasets. Bioinformatics 33, 2594–2595. [PubMed: 28398468]

Sakurai K, Ishigaki K, Shoda H, Nagafuchi Y, Tsuchida Y, Sumitomo S, Kanda H, Suzuki A, Kochi Y, Yamamoto K, and Fujio K (2018). HLA-DRB1 Shared Epitope Alleles and Disease Activity Are Correlated with Reduced T Cell Receptor Repertoire Diversity in CD4+ T Cells in Rheumatoid Arthritis. J. Rheumatol 45, 905–914. [PubMed: 29657145]

Sethna Z, Elhanati Y, Callan CG, Walczak AM, and Mora T (2019). OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. Bioinformatics 35, 2974–2981. [PubMed: 30657870]

Shugay M, Bolotin DA, Putintseva EV, Pogorelyy MV, Mamedov IZ, and Chudakov DM (2013). Huge Overlap of Individual TCR Beta Repertoires. Front. Immunol 4, 466. [PubMed: 24400005]

Shugay M, Bagaev DV, Turchaninova MA, Bolotin DA, Britanova OV, Putintseva EV, Pogorelyy MV, Nazarov VI, Zvyagin IV, Kirgizova VI, et al. (2015). VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. PLoS Comput. Biol 11, e1004503. [PubMed: 26606115]

Soto C, Bombardi RG, Branchizio A, Kose N, Matta P, Sevy AM, Sinkovits RS, Gilchuk P, Finn JA, and Crowe JE Jr. (2019). High frequency of shared clonotypes in human B cell receptor repertoires. Nature 566, 398–402. [PubMed: 30760926]

Van Rossum G, and Drake FL Jr. (1995). Python reference manual (Centrum voor Wiskunde en Informatica)

Venturi V, Kedzierska K, Tanaka MM, Turner SJ, Doherty PC, and Davenport MP (2008). Method for assessing the similarity between subsets of the T cell receptor repertoire. J. Immunol. Methods 329, 67–80. [PubMed: 18001765]

Venturi V, Quigley MF, Greenaway HY, Ng PC, Ende ZS, McIntosh T, Asher TE, Almeida JR, Levy S, Price DA, et al. (2011). A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. J. Immunol 186, 4285–4294. [PubMed: 21383244]

Warren RL, Freeman JD, Zeng T, Choe G, Munro S, Moore R, Webb JR, and Holt RA (2011). Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome Res 21, 790–797. [PubMed: 21349924]

Waskom M, Botvinnik O, O'Kane D, Hobson P, Lukauskas S, Gemperline D, Augspurger T, Halchenko Y, Cole J, Warmenhoven J, et al. (2017). Seaborn: statistical data visualization, Published online September 3, 2017 10.5281/zenodo.883859.

Weinstein JA, Jiang N, White RA 3rd, Fisher DS, and Quake SR (2009). High-throughput sequencing of the zebrafish antibody repertoire. Science 324, 807–810. [PubMed: 19423829]

Werner L, Nunberg MY, Rechavi E, Lev A, Braun T, Haberman Y, Lahad A, Shteyer E, Schvimer M, Somech R, et al. (2019). Altered T cell receptor beta repertoire patterns in pediatric ulcerative colitis. Clin. Exp. Immunol 196, 1–11. [PubMed: 30556140]

Ye J, Ma N, Madden TL, and Ostell JM (2013). IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res 41, W34–W40. [PubMed: 23671333]

Highlights

- TCR clonotype sharing between individuals is higher than expected by chance
- High-frequency TCR clonotypes are captured using gDNA and mRNA sequencing methods
- Shared TCR clonotypes have a higher probability of being generated
- Functional annotation for TCR clonotypes can be obtained using GenBank

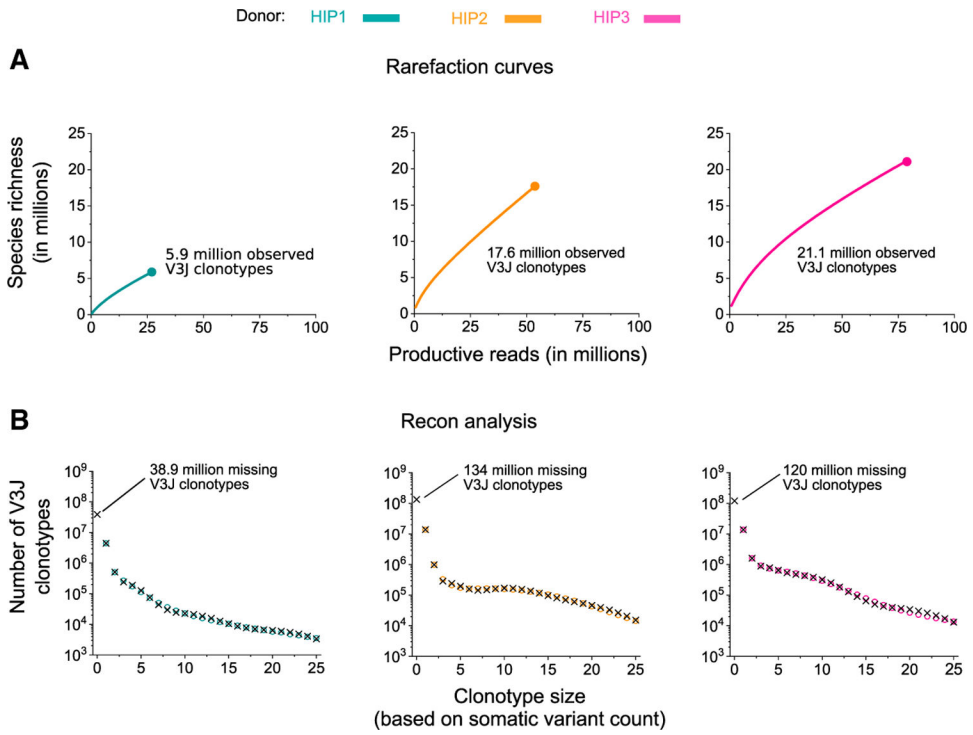


Figure 1. TCRβ V3J Clonotype Diversity for Subjects HIP1, HIP2, and HIP3

(A) Rarefaction curves for species richness of V3J clonotypes generated using the program RTK (Saary et al., 2017). HIP1 had an endpoint species richness value of 5.9 million V3J clonotypes (left panel), HIP2 had a value of 17.6 million V3J clonotypes (middle panel), and HIP3 had a value of 21.1 million V3J clonotypes (right panel). The endpoint estimates for species richness appear on each plot with a filled-in circle.

(B) Recon (Kaplinsky and Arnaout, 2016) estimates suggested about 38.9 million V3J clonotypes were not observed at this depth of sequencing for HIP1 (left panel), 134 million for HIP2 (middle panel), or 120 million for HIP3 (right panel). The observed values for clonotypes binned by their repeat frequency (clonotype group size) is represented by an open circle, and theoretical fits obtained using Recon are represented by an X. For clarity, only the first 25 clonotype group sizes are shown on the plot.

See also Figure S1.

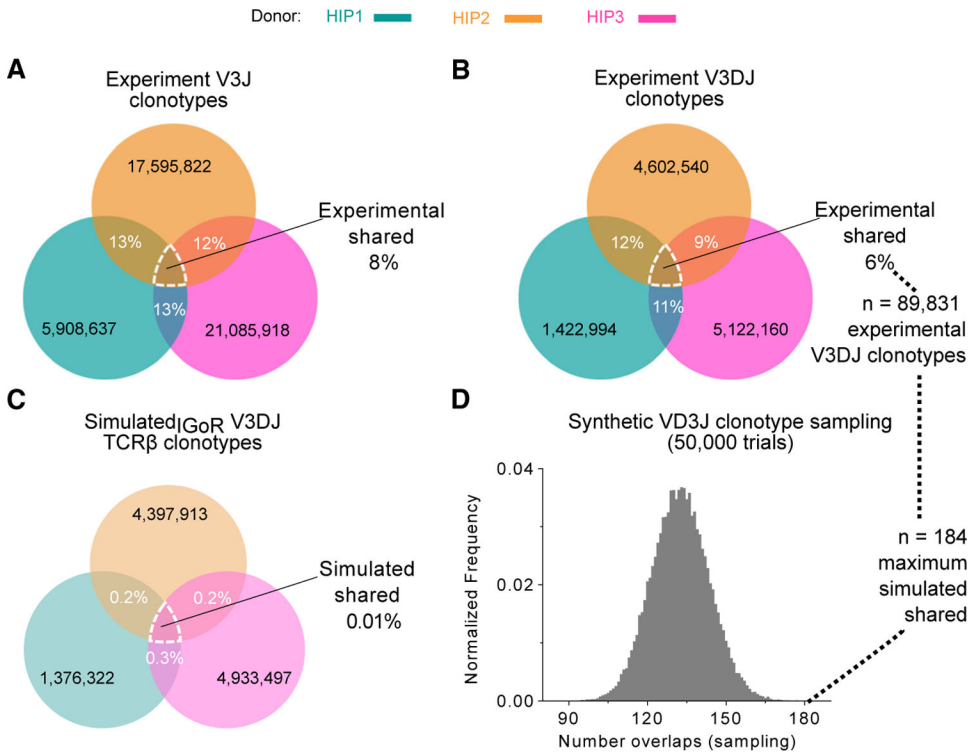


Figure 2. Shared V3J Clonotypes for TCRβ Chains Belonging to Subjects HIP1, HIP2, and HIP3

(A) Shared V3J clonotypes for experimentally determined TCRβ chains.

(B) Shared V3DJ clonotypes for experimentally determined TCRβ chains.

(C) Shared V3DJ clonotypes for synthetic repertoires generated using the program IGoR (Marcou et al., 2018). The pairwise percentage overlaps were based on the average computed for 50,000 comparisons. The average and standard error of the mean (SEM) for the pairwise percentage overlaps were 0.2% (2.0×10^{-5}) between simHIP1 and simHIP2, 0.3% (2.0×10^{-5}) between simHIP1 and simHIP3, and 0.2% (1.0×10^{-5}) between simHIP2 and simHIP3. The average (SEM) for the percentage overlaps among all three synthetic repertoires was 0.01% (4.0×10^{-6}). Reducing the CDR3 lengths of the V3DJ clonotypes in (B) so that the maximum length is 19 amino acids did not change the value of the percentage overlaps.

(D) Histogram of overlap counts among all three synthetic repertoires, simHIP1, simHIP2, and simHIP3, for 50,000 comparisons. Ranking the overlap count among the three experimentally determined TCRβ repertoires (n = 89,831 shared V3DJ clonotypes) against those obtained from the three synthetic repertoires gave an estimated p = 0.002.

See also Figure S2.

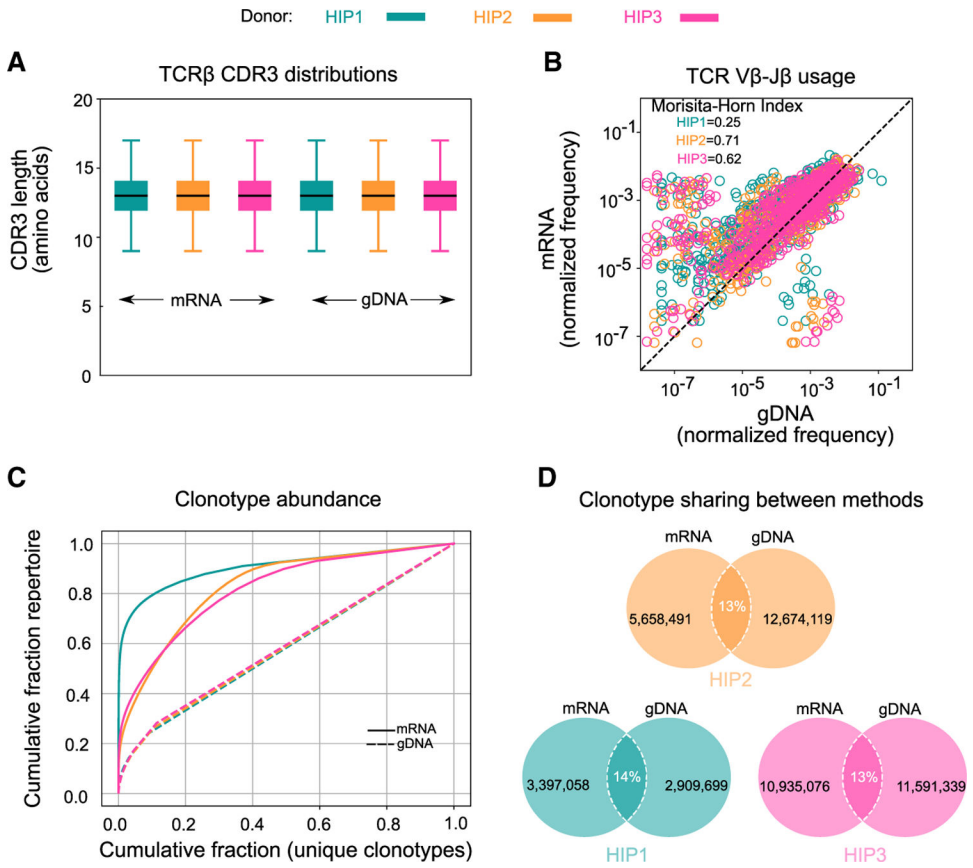


Figure 3. TCRβ Repertoire Statistics for Subjects HIP1, HIP2, and HIP3 Using the mRNA or gDNA Sequencing Methods

(A) Boxplot showing CDR3 length distribution of repertoires obtained from the mRNA or gDNA sequencing methods. Sequencing from the mRNA method resulted in a median CDR3 length of 13 amino acids for HIP1 (n = 3,161,410 unique CDR3s), HIP2 (n = 5,104,666 unique CDR3s), and HIP3 (n = 9,182,164 unique CDR3s). Sequencing from the gDNA method resulted in median CDR3 lengths of 13 amino acids for HIP1 (n = 2,443,117 unique CDR3s), HIP2 (n = 9,967,538 unique CDR3s), and HIP3 (n = 9,018,345 unique CDR3s). Each box represents the interquartile range (IQR) from 25% to 75%. The median is represented by a line within each box, and the whiskers are each within 1.5 of the IQR. There was no statistical difference in the median values according to a Kruskal-Wallis test ($p < 2.2E{-}16$).

(B) Morisita-Horn indices for subject HIP1, HIP2, or HIP3 between mRNA and gDNA sequencing methods.

(C) Distributions of clonotype abundances for the mRNA and gDNA methods represented as the cumulative fraction of the repertoire versus the cumulative fraction of unique TCRβ V3J clonotypes. The clonotypes were ordered from largest to smallest based on the total number of unique somatic variants associated with each clonotype.

(D) TCRβ V3J clonotype overlap percentages for subject HIP1, HIP2, or HIP3 between mRNA and gDNA sequencing methods. Subsampling yielded median percentage overlaps of $6.8\% \pm 1.1\% \times 10^{-2}$, $5.6\% \pm 7.5\% \times 10^{-3}$ or $6.6\% \pm 5.5\% \times 10^{-3}$ for subjects HIP1, HIP2 or HIP3, respectively.

See also Figure S3.

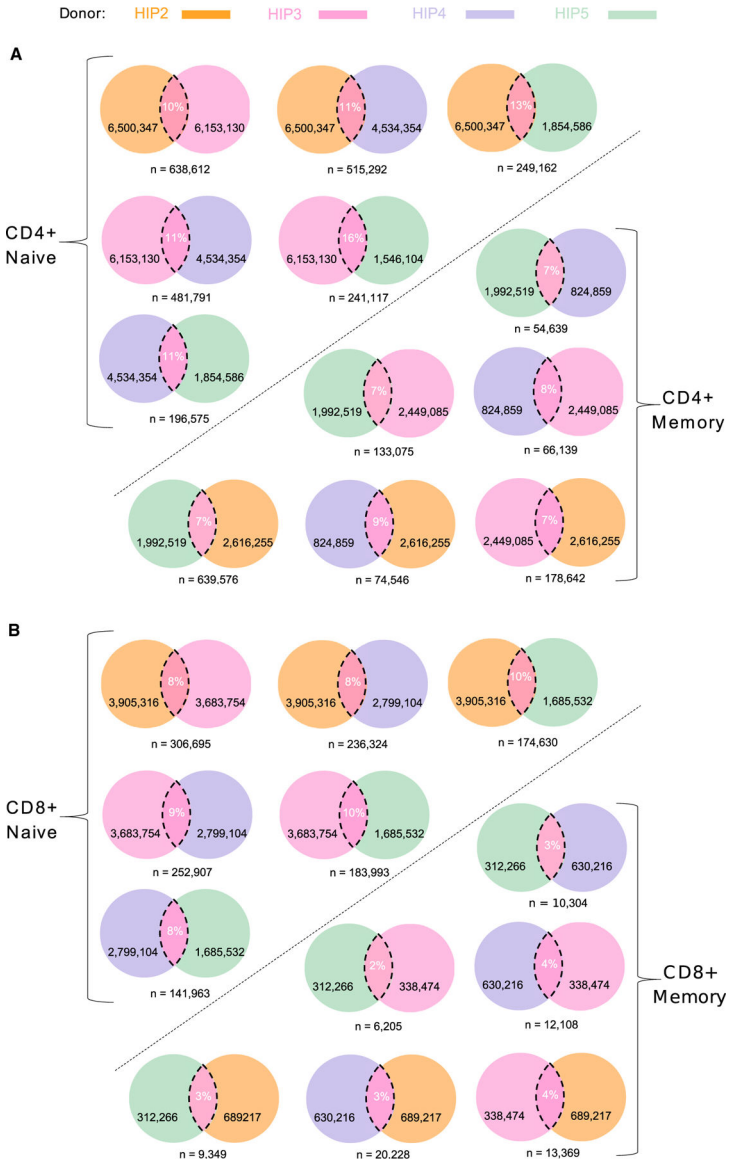


Figure 4. Pairwise Percentage Overlaps of T Cell Subsets Based on gDNA Sequencing for Subject HIP2, HIP3, HIP4, or HIP5

(A) Pairwise percentage clonotype overlaps for naive or memory CD4+ T cell subsets.

(B) Pairwise percentage clonotype overlaps for naive or memory CD8+ T cell subsets.

See also Tables S3 and S4 and Figure S4.

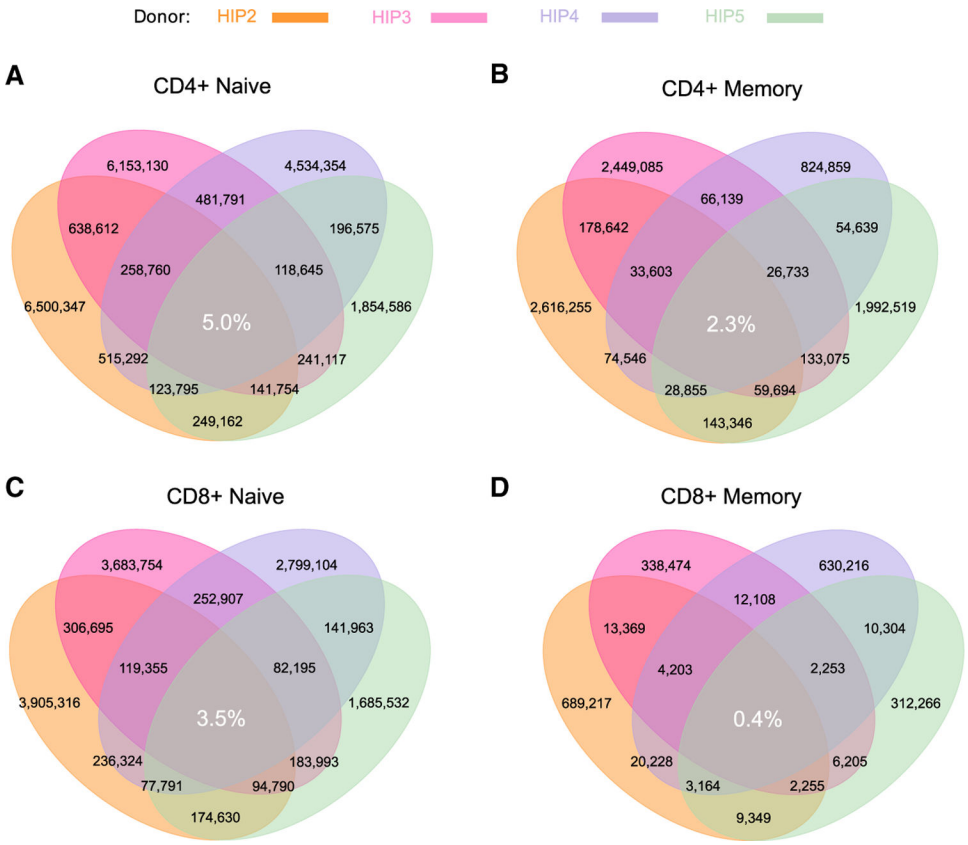


Figure 5. TCRβ Clonotype Sharing for T Cell Subsets

The total number of shared clonotypes in each T cell subset belonging to subjects HIP2, HIP3, HIP4, and HIP5 was determined. The percentage overlaps for all five subjects were also determined in each T cell subset and are represented as white text in each figure.

(A) CD4+ naive cell subset.

(B) CD4+ memory cell subset.

(C) CD8+ naive subset.

(D) CD8+ memory cell subset.

See also Figure S5.

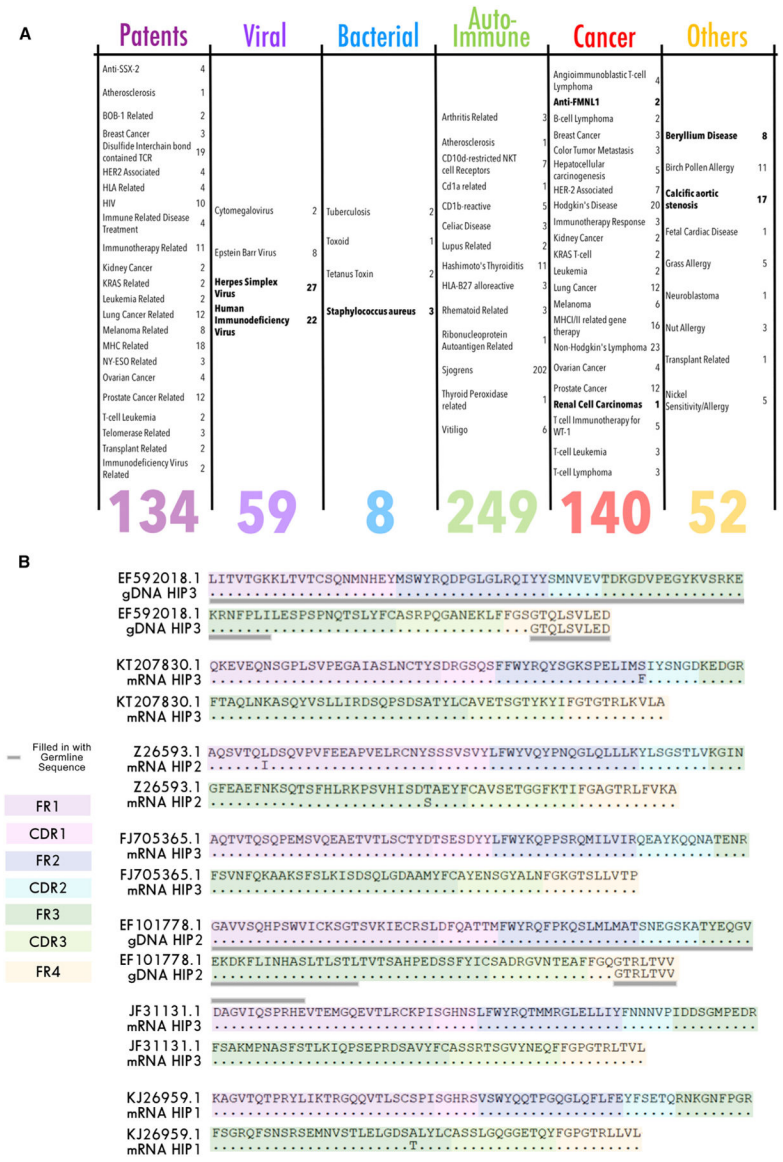


Figure 6. TCRβ Clonotypes from HIP Subjects Appearing in GenBank

TCRβ V3J clonotypes from HIP subjects were used to search against the entire GenBank database for possible matches.

(A) Exact matches were grouped either as patented sequences or into one of five categories that focused on the target: viral target, bacterial target, autoimmune target, cancer target, and other.

(B) Representative amino acid sequence alignments between the V region from GenBank and the V region from the HIP subject. In cases in which the sequence was missing in the framework region or regions, the closest-matching germline sequence was used to fill in the missing region. The filled-in portion of the sequence is highlighted by the thick gray line of the alignment.

See also Table S5.

Table 1.

Research Subject Demographics

Subject[a]	Subject No. from Clinical Site	Gender	Age (years)
HIP1	VVC[b] 1051	F	47
HIP2	VVC 657	M	22
HIP3	VVC 1056	M	29
HIP4	VVC 1124	M	32
HIP5	VVC 1386	F	30

See also Tables S2 and S3.

[a] All samples were collected by leukapheresis from healthy Caucasian subjects living in Nashville, Tennessee.

[b] VVC, Vanderbilt Vaccine Center.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
CD4+ T Cell Isolation Kit, Human	Miltenyi Biotec	Cat# 130–096-533
CD8+ T Cell Isolation Kit, Human	Miltenyi Biotec	Cat# 130–096-495
CD4+ Central Memory T Cell Isolation Kit, Human	Miltenyi Biotec	Cat# 130–094-302
CD45RO MicroBeads, Human	Miltenyi Biotec	Cat# 130–046-001
LS Columns	Miltenyi Biotec	Cat# 130–042-401
Biological Samples		
Human adult leukapheresis sample, HIP1	Vanderbilt Vaccine Center, https://www.vumc.org/vvc/contact-us	VVC 1051
Human adult leukapheresis sample, HIP2	Vanderbilt Vaccine Center, https://www.vumc.org/vvc/contact-us	VVC 657
Human adult leukapheresis sample, HIP3	Vanderbilt Vaccine Center, https://www.vumc.org/vvc/contact-us	VVC 1056
Human adult leukapheresis sample, HIP4	Vanderbilt Vaccine Center, https://www.vumc.org/vvc/contact-us	VVC 1124
Human adult leukapheresis sample, HIP5	Vanderbilt Vaccine Center, https://www.vumc.org/vvc/contact-us	VVC 1386
Critical Commercial Assays		
Immunoseq TCRB Kit	Adaptive Biotechnologies	Cat# hsTCRB
TCRB Library and Sequencing Service	AbHelix LLC	Cat# AH91004
QIAGEN QIAamp DNA Blood Maxi Kit	QIAGEN	Cat# 51192
QIAGEN RNeasy Maxi Kit	QIAGEN	Cat# 75162
HiSeq Rapid SBS Kit v2 (500 cycles)	Illumina	Cat# FC-402–4023
HiSeq PE Rapid Cluster Kit v2	Illumina	Cat# PE-402–4002
HiSeq Rapid Duo cBot Sample Loading Kit	Illumina	Cat# CT-403–2001
NextSeq 500/550 Mid Output Kit v2.5 (150 cycles)	Illumina	Cat# 20024904
Qubit dsDNA HS Assay Kit	Thermofisher Scientific	Cat# Q32851
High Sensitivity DNA Kit	Agilent	Cat# 5067–4626
KAPA Library Quantification Kit	Roche	Cat# KK4835
SuperScript IV First-Strand Synthesis System	ThermoFisher Scientific	Cat# 18091050
Phusion High-Fidelity DNA Polymerase	ThermoFisher Scientific	Cat# F530S
Deposited Data		
gDNA sequencing (Adaptive Biotechnologies)	This paper	https://doi.org/10.21417/CS2020CR
mRNA sequencing (AbHelix LLC)	This paper	SRA: PRJNA511481
Recombinant DNA		

REAGENT or RESOURCE	SOURCE	IDENTIFIER
HLA typing HIP1	Institute for Immunology Murdoch University Western Australia	VVC 1051
HLA typing HIP2	Institute for Immunology Murdoch University Western Australia	VVC 657
HLA typing HIP3	Institute for Immunology Murdoch University Western Australia	VVC 1056
HLA typing HIP4	Institute for Immunology Murdoch University Western Australia	VVC 1124
HLA typing HIP5	Institute for Immunology Murdoch University Western Australia	VVC 1386
Software and Algorithms		
PyIR v1.0	This paper and Soto et al., 2019	https://github.com/crowelab/PyIR
USEARCH v9.1	Edgar and Flyvbjerg, 2015	https://www.drive5.com/usearch/manual9.2/
FASTQC v.0.11.6	Andrews, 2012	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
IgBLAST v.1.9	Ye et al., 2013	https://ncbi.github.io/igblast/
Recon v.2.1	Kaplinsky and Arnaout, 2016	https://github.com/ArnaoutLab/Recon
IGOR v1.1.0	Marcou et al., 2018	https://qmarcou.github.io/IGoR/
OLGA v1.1.0	Sethna et al., 2019	https://github.com/statbiophys/OLGA
MiXCR v.2.1.10	Bolotin et al., 2015	https://mixcr.readthedocs.io/en/master/
RTK v.0.93.1	Saary et al., 2017	https://github.com/hildebra/Rarefaction
VDJTools v1.2.1	Shugay et al., 2015	https://github.com/mikessh/vdjtools
Python v.2.7.12	Python software foundation	http://www.python.org
Numpy v.1.13.3 (distributed with Python)	Python software foundation	http://www.python.org
Seaborn Python plotting module v.0.8.1	Waskom et al., 2017	https://doi.org/10.5281/zenodo.883859
MongoDB v.3.4.0	MongoDB Inc. New York, NY, USA	https://www.mongodb.com
R statistical package v3.2.3	R Core Team	https://www.r-project.org/
ORIGIN(Pro) 2018b	OriginLab Corporation, Northampton, MA, USA.	https://www.originlab.com/
Other		
Processed sequence data, analyses, Python scripts and other resources related to the sequencing of human subjects HIP1, HIP2, HIP3, HIP4 and HIP5.	This paper	https://github.com/crowelab/TCRBmanuscript
gDNA (Adaptive Biotechnologies)	Emerson et al., 2017	https://doi.org/10.21417/B7001Z
Synthetic repertoire sets created with IGoR	This paper	https://doi.org/10.6084/m9.figshare.c.5002037.v1
GenBank (release 231)	Clark et al., 2016	ftp://ftp.ncbi.nlm.nih.gov/genbank
IMGT Reference directory in FASTA format	Lefranc and Lefranc, 2001	http://www.imgt.org/vquest/refseqh.html#VQUEST