Published in final edited form as:

Nature. 2019 February ; 566(7744): 398–402. doi:10.1038/s41586-019-0934-8.

High frequency of shared clonotypes in human B cell receptor repertoires

Cinque Soto^{1,2}, Robin G. Bombardi¹, Andre Branchizio¹, Nurgun Kose¹, Pranathi Matta¹, Alexander M. Sevy⁴, Robert S. Sinkovits⁵, Pavlo Gilchuk¹, Jessica A. Finn³, James E. Crowe Jr.^{1,2,3}

(1) The Vanderbilt Vaccine Center, Vanderbilt University Medical Center, Nashville, TN, USA.

(2) Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA.

(3) Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Center, Nashville, TN, USA.

(4) Chemical and Physical Biology Program, Vanderbilt University, Nashville, TN, USA.

(5) San Diego Supercomputer Center, University of California, San Diego, CA, USA.

SUMMARY

The human genome contains approximately 20 thousand protein-coding genes¹, but the size of the collection of adaptive immune system antigen receptors generated by recombination of gene segments with non-templated junctional additions (on B cells) is orders of magnitude larger and unknown. It is not established whether individuals possess unique (private) repertoires or significant components of shared (public) repertoires. Here we sequenced the recombined and expressed B cell receptor gene repertoire in several individuals at unprecedented depth to determine the size of an individual repertoire and the extent of shared repertoire between individuals. The experiments revealed that each individual's circulating repertoire contained between 9 and 17 million B cell clonotypes. The three individuals studied possessed many shared clonotypes, including 1 to 6% B cell heavy chain clonotypes shared between two subjects (0.3% shared by all three) or 20 to 34% of λ or κ light chains shared between two subjects (16 or 22% λ or κ shared by all three). Some of the B cell clonotypes had thousands of clones (somatic variants) within the clonotype lineage. While some of these shared lineages might be driven by exposure to common antigens, prior foreign antigen exposure was not the only force shaping the shared repertoires, as we also identified shared clonotypes present in both human cord blood samples and in all adult repertoires. The unexpectedly high prevalence of shared clonotypes in B cell

Reprints and permissions information is available at www.nature.com/reprints.

Author Information Statement. Address requests for reprints and permissions to James E. Crowe, Jr.; james.crowe@vanderbilt.edu. Author contributions

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

Code availability. The source code (PyIR, Recombinator) and synthetic repertoires (simHIP1–3 and simCORD1–3) are available from <https://github.com/crowelab/PyIR>.

Data Availability Statement. Sequencing data for HIP and CORD data sets have been deposited at the NCBI's Short Read Archive (SRA) under SRP174305. FASTA files for Adaptive Biotechnologies datasets along with V3J and V3DJ clonotypes used for analyses are available from <https://github.com/crowelab/PyIR>.

repertoires, and identification of the sequences of these shared clonotypes, should enable better understanding of the role of B cell immune repertoires in health and disease.

Determination of the complete set of expressed recombined human immune receptor genes is of general interest to understand fundamental aspects of the development and maintenance of the immune system (such as comparing naïve and memory or neonatal and adult repertoires)^{2,3}. We sought to estimate the size and diversity of human B cell receptor (BCR) repertoires of healthy adults or neonates by sequencing samples to extraordinary depth. We designated B cell recombined variable region sequences as members of a single *V3J clonotype* if the sequences were encoded by the same BCR V_H/J_H , V_κ/J_κ or V_λ/J_λ gene segments and possessed identical amino acids in the third complementarity determining region (CDR3). The V3J clonotype provides a minimal representation for a BCR sequence that can be applied across different immune repertoire sequencing methods. We isolated large numbers of peripheral blood mononuclear cells (PBMCs) by leukapheresis from three healthy adults, designated HIP1 (female, age 47 y), HIP2 (male age 22 y) or HIP3 (male age 29 y), obtaining 13, 21, or 30 billion PBMCs, respectively (Extended Data Table 1). To increase sequencing depth, we used diverse methods and primer sets (Extended Data Tables 2, 3, and 4).

The sequencing reactions yielded 1.4 , 1.5 or 1.3×10^9 raw sequencing reads for subjects HIP1, 2 or 3. We processed the sequences to remove low-quality reads (see Supplementary Methods), obtaining about 5.8 , 6.3 , or 5.1×10^8 sequences after quality control filtering for subject HIP1, 2 or 3, respectively. After filtering, sequences were designated *productive reads*. We assigned the inferred germline variable gene segments for BCR sequences and identified junctional residues using the PyIR informatics pipeline based on IgBLAST⁴ and determined unique V3J clonotypes from subject HIP1, 2 or 3.

We used data modeling techniques to determine if the depth of sequencing was adequate to identify a significant proportion of the Ig heavy chain V3J clonotypes in circulation in each subject. We used the program iNEXT⁵ to determine the species richness of V3J clonotypes in the productive read data for each subject. The species richness curves for all three subjects increased asymptotically but never plateaued, suggesting that even at this extreme depth of sequencing we did not identify all the clonotypes in the sample (*left panels* Fig. 1a–c). The number of unique V3J clonotypes approached 80 to 85% of eventual coverage when we collected between 200 to 300 million productive reads. Using the program iNEXT⁵, we also extrapolated the species richness curves out to an additional 100 to 200 million productive reads beyond that obtained with sequencing. The extrapolated data sets yielded an increase in clonotype count of 15 to 25% (*left panels*, Fig. 1–c, see iNEXT extrapolated). We used the program Recon⁶ to estimate the number of missing clonotypes. Estimates from Recon suggested that an additional 38 to 48% of the V3J clonotypes possible at this depth of sequencing were not identified (*right panels*, Fig. 1a–c). The average value for the missing or unobserved clonotypes is about 10.2 million V3J clonotypes, or roughly half of the number of clonotypes we observed from sequencing (Fig. 1d). Lower bound estimates on the size of the repertoires suggest that between 16 to 31 million V3J clonotypes (average 25 million), is expected in circulation (Fig. 1d). To account for the occurrence of somatic

mutations in CDR3s and to group such minor variants into clonotypes, we clustered clonotypes that had 80% sequence identity in the HCDR3 region (Fig. 1e). This procedure suggested that the estimated clonotype number of about 25 million would be reduced by 35 to 46% if clones with small numbers of CDR3 variant residues were grouped. In summary, the experimental results for V3J clonotype number could be adjusted upwards based on extrapolated values (due to incomplete experimental sequencing) but also reduced in number by clustering to accommodate minor somatic mutations in clonotype CDR3s. The data examined in this way suggest that the size of the circulating Ig heavy chain repertoire in individuals is about 11 million, much smaller than originally anticipated⁷. Features of the repertoires were similar between subjects (Extended Data Fig. 1a–1d). Interestingly, the same CDR3 sequence appeared in multiple clonotypes using differing V and J genes. About 12% of all CDR3 amino sequences appeared in multiple Ig V3J clonotypes.

We next sought to determine the extent to which the three experimental repertoires were shared. Subject HIP2 had about 1% of the clonotypes in common with those of subject HIP1 or subject HIP3 (Fig. 2a). Subjects HIP1 and HIP3 shared about 6% of clonotypes. The percentage of shared Ig heavy chain V3J clonotypes between all three subjects HIP1, 2 and 3 (a collection designated: *Shared* HIP1+2+3) was 0.3% ($n = 29,062$ unique V3J clonotypes). We found a similar extent of sharing in our subjects' V3J clonotypes (0.3 to 0.6% shared) with each of three BCR repertoires in an independently derived data set⁸, even though very different methodologies were used for sequencing. The median HCDR3 length of *Shared* HIP1+2+3 ($n = 22,408$ unique CDR3s) was 13 amino acids, which was shorter than the median length of 16 amino acids for *All* HIP1+2+3 ($n = 30,156,947$ unique CDR3s) (Extended Data Fig. 2a).

Previous work^{9,10} showed that V, D and J germline genes pair preferentially. We performed a second analysis of sharing to include only those clonotypes for which a D_H gene assignment could be made in addition to V_H and J_H genes. These “V3DJ clonotypes” were defined similarly to V3J clonotypes but also contained an explicit D_H gene assignment. The percentages of overlapping V3DJ clonotypes were similar to those obtained for V3J clonotypes. Subject HIP2 had about 1% of V3DJ clonotypes in common with subjects HIP1 and 3 (Fig. 2b, *left panel*). Subjects HIP1 and 3 shared about 6% of V3DJ clonotypes. The percentage of shared Ig heavy chain V3DJ clonotypes between all three subjects HIP1, 2 and 3 was 0.2% ($n = 3,464$ unique V3DJ clonotypes). Thus, whether we used V3J or V3DJ clonotype assignments, the percentage of shared clonotypes in the donor repertoires was similar.

To assess if the degree of observed sharing between the three HIP subjects might be due to chance or rather reflected a biologic mechanism causing common selection of certain clonotypes, we constructed null model repertoires for the V(D)J assignments (“VDJ triples”) observed in each of the three experimentally determined repertoires. The HCDR3 lengths were longer for the V3DJ clonotypes, with HIP1, 2 and 3 each having a median CDR3 length of 19 amino acids (Extended Data Fig. 2b). Thus, we generated three large ensembles of synthetic reads each containing $> 2 \times 10^9$ simulated (sim) unique clonotypes: simHIP1, simHIP2, and simHIP3. We sampled VDJ triples from each of the synthetic repertoires based on the frequency distribution of the VDJ triples from the experimentally determined

repertoires (Extended Data Fig. 2c). This procedure was accomplished by randomly sampling unique amino acid CDR3 sequences from 3 to 28 residues in length (about 2 SD above the mean CDR3 length for experimentally observed V3DJ clonotypes) from each synthetic VDJ triple until we obtained a similar HCDR3 length frequency distribution as in the experimental repertoire (Extended Data Fig. 2d). We sampled from simHIP1, simHIP2, and simHIP3 and then determined the percentage of overlapping clonotypes. The average percentage overlap in the simulated repertoires ranged from 0.02 to 0.03% between pairs and 0.0004% for the intersection of all pairs (Fig. 2b, *right panel*). The experimental overlap value ($n = 3,641$ common V3DJ clonotypes) ranked highest in the distribution of overlaps obtained from the simulated HIP repertoires (Extended Data Fig. 2e), suggesting the presence of overlapping clonotypes between HIP samples did not occur by chance alone.

Some germline V_H+J_H gene combinations were used more frequently than others in the experimental *Shared* HIP1+2+3 set (Fig. 2c). Clonotype overlap between donors was not expected for CDR3 lengths of 25 amino acids or greater by chance alone. We analyzed all Ig heavy chain CDR3 amino sequences of length 25 or greater in the *Shared* HIP1+2+3 repertoire ($n = 26$ HCDR3s) and found many shared common motifs (Fig. 2d).

We next determined how many unique somatic variants were associated with the V3J clonotypes. Grouping somatic variants associated with each V3J clonotype by requiring the corresponding CDR1 and CDR2 amino acid sequences to be identical, showed thousands of potential lineages (Fig. 2e). We found the maximum number of somatic variants for a single clonotype with identical CDR1 and CDR2 amino sequences for HIP1, 2 or 3 to be 19,209, 22,408 or 26,919 somatic variants, respectively. The number of somatic variants associated with V3J clonotypes was larger, containing a maximum number of variants of 45,873, 34,378 or 85,898 variants in HIP1, 2 or 3, respectively (Extended Data Fig. 2f).

As expected, the percentage of shared V3J clonotypes for the light chain data sets was much higher, since these chains lack a diversity gene segment and have fewer germline gene segments with which to recombine. For the Ig κ chain, subjects HIP1 and HIP2 shared 29% of clonotypes, while HIP3 shared 34% of clonotypes with HIP2 and 25% of clonotypes with HIP1 (Extended Data Fig. 2g). The percentage of unique clonotypes shared between all three subjects in the Ig κ set was 22% ($n = 97,422$ unique V3J clonotypes). For Ig λ , HIP1 and HIP2 shared 23% of clonotypes while HIP3 shared 27% of clonotypes with HIP2 and 20% of clonotypes with HIP1 (Extended Data Fig. 2h). The percentage of unique clonotypes shared between all three subjects in the Ig λ set was 16% ($n = 66,162$ unique V3J clonotypes).

We next sought to determine if human subjects possess common clonotypes prior to environmental exposures by determining the BCR repertoires of three neonates, using umbilical cord white blood cell samples (designated CORD1, 2 or 3). The median Ig HCDR3 lengths for subjects CORD1, 2 or 3 was 14, 15 or 16 amino acids, respectively (Fig. 3a, *left panel*). As expected, the neonatal antibody sequence repertoires lacked somatic mutations when compared to those of adult subjects; 97% of the sequences in each of the cord blood samples had germline divergence values between 0 and 1% (Fig. 3a, *right panel*). There were fewer V_H+J_H combinations in neonatal repertoires, likely reflecting the smaller

blood volume available (Fig. 3b). The percentage of overlapping V3J clonotypes between cord blood samples was smaller than that observed in adult samples. The percentage overlaps ranged from 0.4 to 0.5% for pairwise CORD samples and 0.1% for the intersection of all three samples (Fig. 3c), or for V3DJ clonotypes 0.6 to 0.7% for pairwise CORD samples and 0.1% for the intersection of all three samples (Extended Data Fig. 3a). To test if the amount of sharing between all three CORD subjects was significant, we created three synthetic Ig repertoires based on the V3DJ frequency profile (Extended Data Fig. 3b) and the HCDR3 length distribution of each experimental repertoire (Extended Data Fig. 3c). The experimental overlap value ($n = 45$ common V3DJ clonotypes) ranked highest in the distribution of overlaps obtained from the simulated cord repertoires (Extended Data Fig. 3d).

We next determined the degree of overlap between V3J clonotypes from the adult *Shared* repertoire and the cord blood samples. We identified the presence of 51 shared clonotypes in all six of the subjects (Fig. 3d). HCDR3s with lengths of 10 amino acids or greater lacked mutations in the region encoded by the inferred D gene (Fig. 3e). We also combined BCR sequences from a published report⁸ with the adult sequences described here, which resulted in a total of 5.9×10^7 unique clonotypes from six adult subjects. Determining the percentage overlap of the six adult samples with the three cord bloods identified 130 public BCR clonotypes (Extended Data Fig. 3e). These findings suggest that some shared clonotypes appear in high frequency in all individuals prior to exposure to foreign antigens, and these clonotypes persist in adult repertoires for decades.

The identification of such relatively high frequencies of shared elements in the human BCR repertoires that appear at birth and persist into adulthood was unexpected and interesting. The understanding of which recombined immune receptors 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 or infection. Targeting universally shared clonotypes could be an important approach in future studies for epitope structure-based rational vaccine design¹¹ using “germline targeting”^{12,13}. Monitoring of immune responses to infection or vaccination can be improved with this information, since many adaptive responses have canonical features¹⁴, with some antiviral B cell clonal lineages exhibiting both genetic convergence and divergence¹⁵ to achieve recurring motifs for recognition of viral protein antigens¹⁶. Also, comparisons of healthy shared repertoires shown here with those that appear during disease conditions could lead to development of new biomarker patterns of disease states and mechanistic insights into the clonotypes that mediate undesirable immune responses associated with autoimmune conditions¹⁷ or malignancy. Many questions remain about the complexity of the human immunome¹⁸. First, we only studied circulating blood cells here, but many lymphocyte populations reside in tissues where the repertoire differs from that of blood¹⁹. Also, this study was conducted in a small number of subjects with limited genetic, racial, and geographic diversity, and they were studied only at one time point. Comparing these data from ultra-deep sequencing with that from emerging techniques for single cell lymphocyte transcriptomics and linked heavy and light chain repertoire sequencing^{20,21} also holds promise for deeper understanding of human immune responses.

METHODS

Research subjects

We studied six (three adult and three neonatal) healthy, HIV-negative subjects with no reported acute infections or vaccinations in the months prior to leukapheresis or umbilical cord blood sample collection. The subjects consisted of an adult female (subject HIP1), two adult males (subjects HIP2 and HIP3), and three healthy full-term neonates (research subject demographics shown in Extended Data Table 1). Leukopaks containing large numbers of PBMCs obtained by leukapheresis were collected from subjects HIP1, 2, and 3 at Vanderbilt University Medical Center (VUMC). Cord blood was acquired immediately after term delivery from the placenta and umbilical cord and collected in heparinized tubes (NDRI). Following leukapheresis or cord blood collection, 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. The studies were approved by the Institutional Review Board of Vanderbilt University Medical Center; adult samples were obtained after informed consent was obtained by the Vanderbilt Clinical Trials Center.

Molecular techniques for RNA/DNA extraction, RT-PCR or 5' RACE amplification, and next generation sequencing procedures

Multiple techniques and sequencing laboratories were used for these procedures to increase our sampling depth (see Supplementary Methods for details). Briefly, total RNA or genomic DNA was extracted from unsorted PBMCs, and antibody heavy and light chain recombined genes were amplified by RT-PCR or PCR using multiple commercial vendor kits, commercial services, or previously published methods^{23–25} followed by DNA sequencing on the Illumina MiSeq and HiSeq 2500 platform (Extended Data Tables 2, 3, and 4). In one case subsets of pan B cells were used as input material for library preparations. Each profiling protocol varied in terms of reverse transcription and amplification strategy (multiplex PCR or 5' RACE), primer sets (V and J gene primers, leader and constant primers or 5' RACE template switching oligo and constant primers), and incorporation of unique molecular identifiers (UMI) for sequence error correction. The molecular amplification fingerprinting²⁶ (MAF) method incorporated UMIs. Protocols that did not incorporate UMIs are the AbHelix service, Adaptive immunoSEQ B cell service, and BIOMED-2 method.

Processing of raw reads

We processed the raw reads using our in-house pipeline and briefly summarize the five steps below (Extended Data Fig. 4, see Supplementary Methods for details): 1) Check quality control (QC) of the sequencing using the FASTQC toolkit²⁷; 2) Generation of full-length contigs from Illumina paired end (PE) reads using the software package USEARCHv9.1²⁸; 3) Removal of the BIOMED-2 primers using the software package FLEXBARv3.0²⁹ (primer sequences in Extended Data Table 3, and schematic of placement in Extended Data Fig. 5); 4) Assign germ lines, determine CDR3 regions and filter out poor quality reads using our PyIR tool (a Python wrapper for IgBLAST v1.6⁴, available from <https://github.com/crowelab/PyIR>); and 5) Deduplication of all redundant reads in the data set was based on the

nucleotide sequence in the framework 1–4 region. It should be noted that the final filter in step 4 of our pipeline uses the Phred score of each base in the CDR3 to determine the plausibility of the read. Any read with a Phred score in the CDR3 region below 30 was discarded. Using such a filter enforced a very high level of stringency, but we considered this desirable in order to normalize QC across divergent laboratories and methods. The filter focused on the CDR3 region, since these residues formed the basis for defining clonotypes. For those methods that provided processed FASTA data (like Adaptive Biotechnologies), we reprocessed the data using PyIR with only minimal filters. To facilitate downstream repertoire analysis, all productive reads (see Supplementary Methods for details) were uploaded to our custom SEEQ database.

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. In some cases, as indicated in Figure 1e, clustering was used to group together V3J clonotypes sharing identical V and J germline gene assignments with CDR3 amino acid sequences sharing 80% or greater sequence identity. For assessing the significance of the amount of clonotype sharing between donors, we used an alternate definition of clonotype that included the D_H germline gene assignment for those sequences where a D gene assignment could be made with high confidence (see below). When an explicit D_H germline assignment could be made, we used the combination of the V, D, J gene and an identical CDR3 aa sequence to define “**V3DJ clonotypes**”. We also grouped together and determined the number of unique and productive reads associated with each V3J clonotype for HIP1, HIP2 or HIP3 (see Extended Data Fig. 2f). We segregated these groupings further by determining those unique and productive reads for nucleotide sequences that contained identical CDR1 and CDR2 amino acid sequences (see Fig. 2e). In some cases, as indicated in Extended Data Figures 2c and 3b, we grouped sequences with matching V, D, and J gene assignments, regardless of CDR3 sequence, to establish groups termed “**VDJ triples**”. Finally, in Figures 2c and 3b and Extended Data Figure 1d, we show the distribution of V3J clonotypes using heatmaps that only consider the V_H + J_H gene assignments (“**VJ heatmap**”).

Defining high-confidence D_H germline gene assignments

D_H gene segments are shorter than either V_H or J_H germline genes making their assignments in sequencing challenging due to high levels of somatic mutation⁹. We set the E-value threshold to 10⁻⁶ for assigning D_H germline genes to productive reads from the sequenced repertoires (identical thresholds were used for V_H and J_H). We note that setting the E-value threshold to 10⁻⁶ resulted in a 75–80% loss in V3J clonotypes. However, the remaining population of experimental V3J clonotypes with D_H gene assignments all had high confidence matches and contained longer HCDR3s.

Construction of clonotype repertoires

Clonotypes obtained from each subject across all sequencing methods were combined into separate pools and dereplicated for each subject HIP1, HIP2, or HIP3. We also pooled clonotypes from HIP1, HIP2, and HIP3 into collections designated *All* HIP1+2+3 and *Shared* HIP1+2+3 (containing common clonotypes). Pooling allowed us to achieve a superior depth of sequencing.

Rarefaction analysis and constructing species richness curves using VJ3 clonotypes

We used the program iNEXT⁵ to subsample populations of VJ3 clonotypes from Ig heavy chains belonging to subjects HIP1, HIP2 or HIP3 based on their frequency of occurrence in productive reads. The iNEXT⁵ program was also used to extrapolate beyond the number of experimentally observed productive reads to 500 million total productive reads in order to obtain estimates for additional VJ3 clonotype counts we could expect with additional sequencing. Chao1 estimates also were computed using the program iNEXT⁵ (see Supplementary Methods for details on this estimate). The program Recon⁶ was used to estimate of the number of missing VJ3 clonotypes in the Ig heavy chain data sets belonging to subjects HIP1, HIP2 or HIP3. The command line arguments used for Recon can be found in Supplementary Methods

Determination of CDR3 length distributions and germline divergence distributions

The CDR3 distributions from each subject were determined from the corresponding distributions of unique clonotypes. All normalized CDR3 length histograms were constructed from *unique* CDR3 amino acid sequences. Germline divergence was defined as 100 percent minus the percent identity that an Ig nucleotide sequence had with its closest matching germline Variable (V) gene sequence. Germline divergence values were converted to integers before constructing normalized histograms.

Determination of the extent of overlapping clonotypes between experimental data sets

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. The search for shared clonotypes included comparisons of clonotypes from adult subjects (HIP1, 2 and 3), three adult subjects in a previously described BCR database⁸, and cord blood samples (CORD1, 2 and 3). All percentage overlaps were rounded to the nearest integer. Percentage overlaps less than 1% were rounded to the nearest decimal place.

Generating synthetic repertoires and determining the extent of overlapping V3DJ clonotypes between synthetic data sets

We used our tool Recombinator to generate synthetic V3DJ clonotypes based on the VDJ triple frequency and CDR3 length distribution of the experimentally derived repertoires (see Supplementary Methods for details). We generated three large synthetic repertoires corresponding to the HIP data sets (denoted as simHIP1, simHIP2 or simHIP3). In total, we ended up with 2.37×10^9 , 2.42×10^9 and 2.49×10^9 unique synthetic V3DJ clonotypes for

simHIP1, simHIP2 and simHIP3 respectively. Five hundred synthetic repertoires were subsampled (with replacement) from each of these larger sets. A total of 1,000 overlap comparisons was used to obtain an estimate of the P value by ranking the overlap count between the experimentally determined repertoires against the corresponding overlap counts from the synthetic repertoires. We generated 100 synthetic repertoires for each CORD sample (simCORD1, simCORD2 and simCORD3) since the VDJ triple frequencies were smaller than those from the HIP sets. The P value was estimated in the same way using 1,000 comparisons.

Clustering somatic variants to handle length variations

To remove any methodological biases in the length of the nucleotide sequences that occurred from using different sequencing strategies, VSEARCH³⁰ was used to cluster the somatic variants associated with each unique V3J clonotype. The sequence identity threshold used for clustering was set to 100%. The goal here was to determine the number of possible unique somatic variants and not to correct or “average” out the error associated with sequencing.

Collapsing heavy chain V3J clonotypes using complete-linkage clustering

We clustered heavy chain clonotypes belonging to subjects from HIP1, HIP2 or HIP3 using complete-linkage clustering at a sequence identity threshold of 80% (converted from a Hamming distance). V3J clonotypes with the same CDR3 length and V and J germline gene assignments first were grouped together and then clustered separately. All clustering was carried out using the Scipy package (version 1.0) in Python (versions 3.6.1 and 3.6.4).

Figure and plot generation

All plots and normalized frequency histograms were generated using OriginPro 2018. Heat maps were generated using the Seaborn plotting module (version 0.8.1) in Python (version 2.7.12). Web logos were created using WebLogo (version 2.8.2)²². The Mann-Whitney U test and Pearson's correlation coefficient (r) were both computed using the R statistical package (version 3.2.3).

Extended Data

Extended Data Figure 1. Repertoire properties for Ig V3J clonotype data belonging to HIP1-3. (a) Normalized frequency histogram of HCDR3 sequence lengths belonging to Ig heavy chain V3J clonotypes for HIP1 (left plot, $n = 8,623,076$ unique CDR3s with a median CDR3 length 16 aa), HIP2 (middle plot, $n = 15,413,214$ unique CDR3s with a median CDR3 length 16 aa) and HIP3 (right plot, $n = 7,081,314$ unique CDR3s with a median CDR3 length 15 aa). (b) Normalized frequency histogram of germline divergence values for HIP1 (left plot), HIP2 (middle plot) or HIP3 (right plot). Germline divergence was defined as 100 percent minus the percent nucleotide identity a read had with its closest matching germline variable (V) gene sequence. Median percent germline divergence values for HIP1, 2 or 3 were 3, 0 or 2 respectively. (c) Normalized frequency histogram of germline divergence values by isotype for subject HIP1 (left plot), HIP2 (middle plot) or HIP3 (right plot). The median germline divergence was 0 for all IgM data sets. All isotype data were obtained from the AbHelix sequencing method. (d) Heat map representation for unique V_H+J_H recombinations in subject HIP1, 2 or 3. The data from each set were transformed to obtain Z scores using the mean and standard deviation.

Extended Data Figure 2. Extent of sharing between Ig clonotypes belonging to HIP1–3.

(a) Normalized frequency histogram of HCDR3 sequence lengths belonging to V3J clonotypes from *All*/HIP1+2+3 (blue filled curve, $n = 30,156,947$ unique CDR3s with a median CDR3 length of 16 amino acids) and *Shared* HIP1+2+3 (grey bins, $n = 22,934$ unique CDR3s with a median CDR3 length of 13 aa). The medians were statistically different based on a two-tailed Mann-Whitney U test with a $P < 2.2 \times 10^{-16}$ (at an $\alpha = 0.05$).

(b) Normalized frequency histogram of CDR3 lengths belonging to all V3DJ clonotypes from HIP1 ($n = 1,750,325$ unique CDR3s with a median CDR3 length of 19 aa), HIP2 ($n = 3,889,527$ unique CDR3s with a median CDR3 length of 19 aa) and HIP3 ($n = 1,437,339$ unique CDR3s with a median CDR3 length of 19 aa).

(c) Cumulative distribution of normalized VDJ triple frequencies used for simulation: HIP1 ($n = 4,373$ unique VDJ triples), HIP2 ($n = 4,351$ unique VDJ triples) and HIP3 ($n = 4,372$ unique VDJ triples).

(d) Log-Log frequency plot between experimental and synthetic CDR3 lengths. The Pearson correlation

coefficient $r = 1.00$ with a $P < 2.2 \times 10^{-16}$ (at an $\alpha = 0.05$) ($n = 26$ CDR3 length bins for each set). **(e)** Normalized frequency histogram of V3DJ overlap counts between all three synthetic HIP distributions ($n = 3,641$ common clonotypes between sequenced repertoires). **(f)** V3J clonotypes with the largest numbers of somatic variants. Numbers in parenthesis denote counts for the number of unique somatic variants associated with a V3J clonotype for HIP1, HIP2 or HIP3. **(g)** Percentage overlaps for the Ig κ V3J clonotypes from the experimentally determined repertoires belonging to HIP1–3. **(h)** Percentage overlaps for Ig λ V3J clonotypes from the experimentally determined repertoires belonging to HIP1–3.

Extended Data Figure 3. Shared Ig heavy chain clonotypes for three cord blood samples.

(a) V3DJ clonotype overlaps from three cord blood samples, CORD1 ($n = 40,480$ unique V3DJ clonotypes), CORD2 ($n = 66,718$ unique V3DJ clonotypes) and CORD3 ($n = 105,555$ unique V3DJ clonotypes) (b) Cumulative distribution of normalized VDJ triple frequencies for CORD1 ($n = 2,273$ unique VDJ triples), CORD2 ($n = 2,788$ unique VDJ triples) and CORD3 ($n = 3,002$ unique VDJ triples). (c) Log-Log frequency plot between experimental and synthetic CDR3 lengths. The Pearson correlation coefficient $r = 1.000$ with a $P < 2.2 \times 10^{-16}$ (at an $\alpha = 0.05$) ($n = 21$ bins for each set). It should be noted that there were no V3DJ

clonotypes with CDR3 lengths less than 8 amino acids in length. **(d)** Normalized frequency histogram of V3DJ overlap counts between all three synthetic CORD distributions ($n = 45$ common clonotypes between all three sequenced repertoires). **(e)** V3J clonotypes identified in the adult subjects HIP1, 2, and 3 ("**All HIP1+2+3**") were combined with an independently derived set of Ig heavy chain V3J clonotypes for which sequences were publicly available⁸ ("**All Adaptive1+2+3**"). Starting from the combined set of 59,193,994 clonotypes from six adult Ig heavy chain repertoires, each of the three cord blood sets was scanned in a serial fashion, keeping only the common clonotypes. A total of 130 shared V3J clonotypes was identified.

Extended Data Figure 4. Schematic diagram showing bioinformatic sequence processing.

The flowchart shows how a typical sequencing run using paired ends (PE) reads from Illumina was processed using bioinformatics pipeline. Detailed descriptions for each of the programs used in the pipeline can be found in the supplementary methods.

Extended Data Figure 5. Schematic showing placement of primers.
Annotated example of a biological sequence obtained from the two-step barcoded library preparation protocol. The red and yellow regions show the placement of the first and second steps of PCR amplification. The cyan region shows the location of the RID tagged RT gene specific primer.

Extended Data Table 1.

Research subject demographics						
Donor and sample type	Donor	Gender	Race	Age (years)	Site of collection	Donor number from site
Healthy adult; leukapheresis collection	HIP1	F	Caucasian	47	Nashville, TN	VVC * 1051
	HIP2	M	Caucasian	22		VVC 657
	HIP3	M	Caucasian	29		VVC 1056
Neonate; cord blood	CORD1	M	Caucasian	Neonate	Pittsburgh, PA	NDRI [†] Donor 1135 (ND12279); birth weight 3,480 g
	CORD2	F	Caucasian	Neonate		NDRI Donor 1136(ND12280); birth weight 4,043g

Donor and sample type	Donor	Gender	Race	Age (years)	Site of collection	Donor number from site
	CORD3	F	Caucasian	Neonate		NDRI Donor 1137(ND12281); birth weight 3,500 g

* Vanderbilt Vaccine Center (VVC).

[†] National Disease Research Interchange (NDRI)

Extended Data Table 2.

Summary of sequencing methods and cell counts

Subject	Immune Repertoire Assay	Target	Number PBMCs Processed	Number B Cells Studied	Number T Cells Studied	NGS Platform	Sequencing Vendor
HIP1	Adaptive ImmunoSEQ® Human TCRA/b Kit	HCDR3	4×10^7	--	1.3×10^7	NextSeq SR-150	VANTAGE
	Clontech SMARTer® Human TCR Profiling Kit	Full-length	1×10^7	--	4×10^6	MiSeq PE-300	VANTAGE
	NEB AbSeq® Human B and T Cell Profiling Kit	Full-length	1×10^7	4.5×10^5	2×10^6	MiSeq PE-300	VANTAGE
	AbHelix® Human B and T-Cell Profiling Assay	Full-length	9×10^8	3.6×10^7	6×10^7	HiSeq PE-250	AbHelix, LLC
	Crowe Laboratory B Cell Profiling Assay	CDR1-FR4	4×10^8	2.7×10^7	--	HiSeq PE-250	HudsonAlpha
HIP2	Adaptive ImmunoSEQ® Human TCRA/b Kit	HCDR3	7.9×10^8	--	1.8×10^7	NextSeq SR-150	VANTAGE
	Crowe Laboratory B Cell MAF Profiling Assay	CDR1-FR4	2.9×10^8	2.1×10^7	--	HiSeq PE-250	VANTAGE
	AbHelix® Human B and T Cell Profiling Assay	Full-length	9×10^8	4.3×10^7	7.2×10^7	HiSeq PE-250	AbHelix, LLC
	Crowe Laboratory B Cell Profiling Assay	CDR1-FR4	4×10^8	2.7×10^7	--	HiSeq PE-250	HudsonAlpha
	Adaptive ImmunoSEQ® Human BCR Kit	HCDR3	6×10^9	1.9×10^7	--	HiSeq SR-150	Adaptive
HIP3	Adaptive ImmunoSEQ® Human TCRA/b Kit	HCDR3	8.1×10^8	--	1.8×10^7	NextSeq SR-150	VANTAGE
	AbHelix® Human B and T Cell Profiling Assay	Full-length	9×10^8	4.3×10^7	7.2×10^7	HiSeq PE-250	AbHelix, LLC
	Crowe Laboratory B Cell Profiling Assay	CDR1-FR4	4×10^8	2.7×10^7	--	HiSeq PE-250	HudsonAlpha
CORD1	Crowe Laboratory B Cell MAF Profiling Assay	CDR1-FR4	1.4×10^7	3.5×10^5	--	MiSeq PE-250	VANTAGE

Subject	Immune Repertoire Assay	Target	Number PBMCs Processed	Number B Cells Studied	Number T Cells Studied	NGS Platform	Sequencing Vendor
CORD2	Crowe Laboratory B Cell MAF Profiling Assay	CDR1-FR4	7.5×10^6	6.1×10^5	--	MiSeq PE-250	VANTAGE
CORD3	Crowe Laboratory B Cell MAF Profiling Assay	CDR1-FR4	1.3×10^7	9.8×10^5	--	MiSeq PE-250	VANTAGE

Extended Data Table 3.

One-step RT-PCR primers used in this study

Primer	Application	Sequence
Human IgH cDNA synthesis and reverse PCR primer		
J _H	Human IgH RT primer and reverse PCR primer	NNNNCTTACCTGAGGAGACGGTGACC
Human IgH forward PCR primer mix		
V _H 1-FR1	Human multiplex forward IgH PCR primer	NNNNGGCCTCAGTGAAGGTCTCCTGCAAG
V _H 2-FR1	Human multiplex forward IgH PCR primer	NNNNGTCTGGTCCTACGCTGGTGAACCC
V _H 3-FR1	Human multiplex forward IgH PCR primer	NNNNCTGGGGGGTCCCTGAGACTCTCCTG
V _H 4-FR1	Human multiplex forward IgH PCR primer	NNNNCTCGGAGACCCTGTCCCTCACCTG
V _H 5-FR1	Human multiplex forward IgH PCR primer	NNNNCGGGGAGTCTCTGAAGATCTCCTGT
V _H 6-FR1	Human multiplex forward IgH PCR primer	NNNNTCGCAGACCCTCTCACTCACCTGTG
Human IgK cDNA synthesis and reverse PCR primer mix		
J _K 1	Human IgK RT primer and reverse PCR primer	NNNNTTTGATATCCACCTTGGTCCC
J _K 2	Human IgK RT primer and reverse PCR primer	NNNNTTAATCTCCAGTCGTGTCCC
Human IgK forward PCR primer mix		
V _K 1-2-FR1	Human multiplex forward IgK PCR primer	NNNNATGAGGSTCCCYGCTCAGCTGCTGG
V _K 3-FR1	Human multiplex forward IgK PCR primer	NNNNCTCTTCTCCTCTGCTACTCTGGCTCCCAG
V _K 4-FR1	Human multiplex forward IgK PCR primer	NNNNATTCTCTGTTGCTCTGGATCTCTG
Human Igλ cDNA synthesis and reverse PCR primer mix		
J _λ 1	Human Igλ RT primer and reverse PCR primer	NNNNAGGACGGTGACCTTGGTCCC
J _λ 2	Human Igλ RT primer and reverse PCR primer	NNNNAGGACGGTCAGCTGGGTCCC
Human Igλ forward PCR primer mix		

Primer	Application	Sequence
V _λ 1-FR1	Human multiplex forward Igλ PCR primer	NNNNGGTCCTGGGCCCAGTCTGTGCTG
V _λ 2-FR1	Human multiplex forward Igλ PCR primer	NNNNGGTCCTGGGCCCAGTCTGCCCTG
V _λ 3-FR1	Human multiplex forward Igλ PCR primer	NNNNGCTCTGTGACCTCCTATGAGCTG
V _λ 4+5-FR1	Human multiplex forward Igλ PCR primer	NNNNGGTCTCTCTCSCAGCYGTGCTG
V _λ 6-FR1	Human multiplex forward Igλ PCR primer	NNNNGTTCTTGGGCCAATTTTATGCTG
V _λ 7-FR1	Human multiplex forward Igλ PCR primer	NNNNGGTCCAATTCYCAGGCTGTGGTG
V _λ 8-FR1	Human multiplex forward Igλ PCR primer	NNNNGAGTGGATTCTCAGACTGTGGTG

For details on published primer sets see Methods and Supplementary methods.

Two-step RT-PCR primers used in this study

Primer	Application	Sequence
Human IgH cDNA synthesis primer mix		
MAF_J _H	Human IgH RT primer with RID	TTGGCACC CGAGAATTCCTGHHHHHACAAHHHHNCTTACCTGAGGAGACGGTGACC
Human IgK cDNA synthesis primer mix		
MAF_J _K 1	Human IgK RT primer with RID	TTGGCACC CGAGAATTCCTGHHHHHACAAHHHHNNTTTGATATCCACCTTTGGTGCC
MAF_J _K 2	Human IgK RT primer with RID	TTGGCACC CGAGAATTCCTGHHHHHACAAHHHHNNTTTAATCTCCAGTCGTGTCGCC
Human Igλ cDNA synthesis primer mix		
MAF_J _λ 1	Human Igλ RT primer with RID	TTGGCACC CGAGAATTCCTGHHHHHACAAHHHHNAGGACGGTGACCTTGGTGCC
MAF_J _λ 2	Human Igλ RT primer with RID	TTGGCACC CGAGAATTCCTGHHHHHACAAHHHHNAGGACGGTCAGCTGGGTGCC
First PCR amplification		
IgH, K, λ forward PCR primer	Step-out primer, anneals on the IgH, K, λ RT primer	ACTGGAGTTCC TTGGCACCCGAGAAATTCCTCACTG
Human IgH reverse PCR primer mix		
MAF_V _H 1-FR1	Human multiplex reverse IgH FID PCR primer	CGTTCA GAGTTCTACAGTCCGACGATCHHHHACHHHHNGCAGGGCCTCAGTGAAGGTCTCTCTGCAAG
MAF_V _H 2-FR1	Human multiplex reverse IgH FID PCR primer	CGTTCA GAGTTCTACAGTCCGACGATCHHHHACHHHHNGCAGGTCTGGTCTACGCTGGTGAACCC
MAF_V _H 3-FR1	Human multiplex reverse IgH FID PCR primer	CGTTCA GAGTTCTACAGTCCGACGATCHHHHACHHHHNGCAGCTGGGGGGTCCCTGAGACTCTCTG
MAF_V _H 4-FR1	Human multiplex reverse IgH FID PCR primer	CGTTCA GAGTTCTACAGTCCGACGATCHHHHACHHHHNGCAGCTTCGGAGACCCCTCTCCCTCACCTG
MAF_V _H 5-FR1	Human multiplex reverse IgH FID PCR primer	CGTTCA GAGTTCTACAGTCCGACGATCHHHHACHHHHNGCAGCGGGGAGTCTCTGAAGATCTCTGT
MAF_V _H 6-FR1	Human multiplex reverse IgH FID PCR primer	CGTTCA GAGTTCTACAGTCCGACGATCHHHHACHHHHNGCAGTCGCAGACCCCTCTCACCTGCTGTG
Human IgK reverse PCR primer mix		
MAF_V _K 1-2-FR1	Human multiplex reverse IgK FID PCR primer	CGTTCA GAGTTCTACAGTCCGACGATCHHHHACHHHHNGCAGATGAGGSTCCCYGCTCAGCTGTGG
MAF_V _K 3-FR1	Human multiplex reverse IgK FID PCR primer	CGTTCA GAGTTCTACAGTCCGACGATCHHHHACHHHHNGCAG CTCTCTCTCTGCTACTCTGGTCCCAG

For details on published primer sets see Methods and Supplementary methods.

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. We thank Yashasri Umareddy for assistance with the R. We thank Samuel B. Day for assistance with artwork. We thank scientists at the VANTAGE core of Vanderbilt University Medical Center (VUMC), Adaptive Biotechnologies, the Genomic Services Lab at the Hudson Alpha Institute for Biotechnology (Huntsville, AL), and Douglas Zhang and team at Abhelix. We thank New England BioLabs for early access to pre-release Abseq reagents. We thank Karen Trochez and Jill Janssen of the Clinical Trials Center at VUMC and staff and physicians of the Vanderbilt University Medical Center leukapheresis clinic for assistance with large-scale human cell collections. We thank Simon Mallal and Mark Pilkinton (Vanderbilt), Richard Scheuermann (JCVI), and Wayne Koff, Ted Schenkelberg and the Advisory Board of the Human Vaccines Project for helpful discussions. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, Nashville, TN and the San Diego Supercomputer Center at the University of California, San Diego. We acknowledge the use of cord blood cells procured by the National Disease Research Interchange (NDRI) with support from NIH grant U42 OD11158. This work was supported by a grant from the Human Vaccines Project, and institutional funding from Vanderbilt University Medical Center.

Competing Financial Interests. J.E.C has served as a consultant for Sanofi and Pfizer, is on the Scientific Advisory Boards of CompuVax and Meissa Vaccines, is a recipient of research grants from Takeda, Sanofi and Moderna, and is founder of IDBiologics. All other authors declare no conflicts of interest.

REFERENCES

1. Ezkurdia I et al. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum Mol Genet* 23, 5866–5878, (2014). [PubMed: 24939910]
2. The Adaptive Immune Receptor Repertoire Community of the Antibody Society <<https://www.antibodysociety.org/the-airr-community/>>
3. Zalocusky KA et al. The 10,000 Immunomes Project: Building a Resource for Human Immunology. *Cell Rep* 25, 513–522 e513, (2018). [PubMed: 30304689]
4. Ye J, Ma N, Madden TL & Ostell JM IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* 41, W34–40, (2013). [PubMed: 23671333]
5. Hsieh TC, Ma KH & Chao A iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods Ecol Evol* 7, 1451–1456, (2016).
6. Kaplinsky J & Arnaout R Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. *Nat Commun* 7, 11881, (2016). [PubMed: 27302887]
7. Trepel F Number and distribution of lymphocytes in man. A critical analysis. *Klin Wochenschr* 52, 511–515 (1974). [PubMed: 4853392]
8. DeWitt WS et al. A Public Database of Memory and Naive B-Cell Receptor Sequences. *PLoS One* 11, e0160853, doi:10.1371/journal.pone.0160853 (2016). [PubMed: 27513338]
9. Arnaout R et al. High-resolution description of antibody heavy-chain repertoires in humans. *PLoS One* 6, e22365, doi:10.1371/journal.pone.0022365 (2011). [PubMed: 21829618]
10. Boyd SD et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci Transl Med* 1, 12ra23 (2009).
11. Correia BE et al. Proof of principle for epitope-focused vaccine design. *Nature* 507, 201–206, (2014). [PubMed: 24499818]
12. Jardine JG et al. HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen. *Science* 351, 1458–1463, (2016). [PubMed: 27013733]
13. Briney B et al. Tailored Immunogens Direct Affinity Maturation toward HIV Neutralizing Antibodies. *Cell* 166, 1459–1470 e1411, (2016). [PubMed: 27610570]
14. Crowe JE Jr. Principles of Broad and Potent Antiviral Human Antibodies: Insights for Vaccine Design. *Cell Host Microbe* 22, 193–206, (2017). [PubMed: 28799905]

15. Krause JC et al. Epitope-specific human influenza antibody repertoires diversify by B cell intraclonal sequence divergence and interclonal convergence. *J Immunol* 187, 3704–3711, (2011). [PubMed: 21880983]
16. Xu R et al. A recurring motif for antibody recognition of the receptor-binding site of influenza hemagglutinin. *Nat Struct Mol Biol* 20, 363–370, (2013). [PubMed: 23396351]
17. de Bourcy CFA, Dekker CL, Davis MM, Nicolls MR & Quake SR Dynamics of the human antibody repertoire after B cell depletion in systemic sclerosis. *Sci Immunol* 2, (2017).
18. Pederson T The immunome. *Mol Immunol* 36, 1127–1128 (1999). [PubMed: 10698314]
19. Briney BS, Willis JR, Finn JA, McKinney BA & Crowe JE Jr. Tissue-specific expressed antibody variable gene repertoires. *PLoS One* 9, e100839, doi:10.1371/journal.pone.0100839 (2014). [PubMed: 24956460]
20. DeKosky BJ et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol* 31, 166–169, (2013). [PubMed: 23334449]
21. DeKosky BJ et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat Med* 21, 86–91, (2015). [PubMed: 25501908]
22. Crooks GE, Hon G, Chandonia JM & Brenner SE WebLogo: a sequence logo generator. *Genome Res* 14, 1188–1190, (2004). [PubMed: 15173120]
23. Diss TC, Liu HX, Du MQ & Isaacson PG Improvements to B cell clonality analysis using PCR amplification of immunoglobulin light chain genes. *Mol Pathol* 55, 98–101 (2002). [PubMed: 11950958]
24. Smith K et al. Rapid generation of fully human monoclonal antibodies specific to a vaccinating antigen. *Nat Protoc* 4, 372–384, (2009). [PubMed: 19247287]
25. van Dongen JJ et al. Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98–3936. *Leukemia* 17, 2257–2317, (2003). [PubMed: 14671650]
26. Khan TA et al. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci Adv* 2, e1501371, (2016). [PubMed: 26998518]
27. Andrews S FastQC: A quality control tool for high throughput sequence data., <<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>>
28. Edgar RC & Flyvbjerg H Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* 31, 3476–3482, (2015). [PubMed: 26139637]
29. Roehr JT, Dieterich C & Reinert K Flexbar 3.0 - SIMD and multicore parallelization. *Bioinformatics* 33, 2941–2942, (2017). [PubMed: 28541403]
30. Rognes T, Flouri T, Nichols B, Quince C & Mahe F VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4, e2584, doi:10.7717/peerj.2584 (2016). [PubMed: 27781170]

Figure 1. Estimates of V3J clonotype diversity from three healthy adult subjects, designated HIP1, 2, or 3.

Interpolation (thin curves) and extrapolation (thick curves) of species diversity values were obtained using the program iNEXT⁵. The endpoint diversity estimates are represented by the symbols (○) for interpolation and (●) for extrapolation, respectively. The 95% confidence limits were all within $\pm 0.05\%$ of the end-point estimates. The program Recon⁶ was used to estimate the number of unobserved or “missing” V3J clonotypes. The observed frequency of clonotype group sizes and their theoretical fits obtained using Recon are represented by the symbols (○) or (×), respectively. Only the first 25 clonotype group sizes are shown on the plot for clarity. **(a) (Left panel)** Experimental sequencing yielded about 10.7 million Ig heavy V3J clonotypes for HIP1. The species richness endpoint estimate was 10,715,954. Extrapolation gave a species richness estimate of 12,590,751. **(Right panel)** Recon estimates suggested a total of 9.4 million missing clonotypes. **(b) (Left panel)** Experimental sequencing yielded about 17.1 million Ig heavy V3J clonotypes for subject HIP2. The species richness endpoint estimate was 17,110,333. Extrapolation gave a species richness estimate of 20,210,426. **(Right panel)** Recon estimates suggested a total of 15.7 million missing clonotypes. **(c) (Left panel)** Experimental sequencing yielded about 9.0 million V3J clonotypes for HIP3. The endpoint species richness estimate was 8,989,812. Extrapolation gave a species richness estimate of 11,984,340. **(Right panel)** Recon estimates suggested a total of 5.6 million missing clonotypes. **(d)** A summary of estimates for repertoire size based on clonotype frequencies. Species richness values obtained from experimental sequencing were rounded to nearest hundred thousand. **(e)** Clustering of Ig heavy chain V3J clonotypes in the HCDR3s reduced the total number of unique clonotypes by 35 to 46%.

Figure 2. Shared clonotypes between three healthy adult subjects (HIP1, 2 and 3).

(a) Shared V3J clonotypes from sequenced Ig heavy chains. (b) (Left panel) Shared V3DJ clonotypes from sequenced Ig heavy chains with HCDR3 lengths from 3 to 28 amino acids. (Right panel) Shared V3DJ clonotypes from synthetic HIP repertoires with HCDR3 lengths from 3 to 28 amino acids. The percentage overlaps were based on the average of 1,000 comparisons from bootstrap testing involving synthetic HIP repertoires. The average and standard error of the mean (s.e.m.) for the percentage overlaps was 0.03% (5.0×10^{-5}) between simHIP1 and simHIP2, 0.03% (4.9×10^{-5}) between simHIP1 and simHIP3, and 0.02% (6.0×10^{-5}) between simHIP2 and simHIP3. The average and s.e.m. for the percentage overlap between simHIP1 and simHIP2 and simHIP3 was 0.0004% (6.9×10^{-6}). The V3DJ overlap count between all three sequenced repertoires ($n = 3,641$ common clonotypes) ranked highest in the 1,000 comparisons giving a $P = 1.0 \times 10^{-4}$ (see Extended Data Fig. 2e for normalized histogram of common clonotypes between synthetic sets). (c)

Fold change in V_H+J_H usage between *Shared* HIP1+2+3 ($n = 29,062$ unique clonotypes) and all HIP subjects (designated: *All* HIP1+2+3, $n = 36,064,712$ unique clonotypes). **(d)** Common motifs in *Shared* V3J clonotypes with long CDR3s shown as a WebLogo²². **(e)** Somatic variant count for V3J clonotypes from the *Shared* HIP1+2+3 collection whose somatic variants had identical CDR1 and CDR2 amino acid sequences plotted in order of decreasing frequency. Numbers in parenthesis denote V3J clonotypes having the largest number of somatic variant counts.

Figure 3. Occurrence of public V3J clonotypes that are shared in adult and cord blood repertoires.

(a) (*Left panel*) Normalized frequency histogram of HCDR3 sequence lengths from V3J clonotypes belonging to CORD1 (*top plot*, n = 229,478 unique CDR3 sequences with a median CDR3 length 14 aa), CORD2 (*middle plot*, n = 243,497 unique CDR3 sequences with a median CDR3 length 15 aa) or CORD3 (*bottom plot*, n = 322,882 CDR3 sequences with a median CDR3 length 16 aa). (*Right panel*) Normalized frequency histogram of germline divergence values for CORD1, 2 and 3 and adult subjects HIP1, 2, and 3. The shaded area corresponds to the CORD1–3 data set. Germline divergence was defined as 100 percent minus the percent nucleotide identity a read had with its closest matching germline variable (V) gene sequence. **(b)** Heat map representation for unique V_H+J_H recombinations in CORD1, 2 or 3. **(c)** Shared V3J clonotypes between CORD samples. **(d)** Schematic

illustration showing shared V3J clonotypes common to all six subjects. Starting from the *Shared* HIP1+2+3 set, the three CORD sets were compared sequentially to determine the presence of 51 common clonotypes. (e) Shared V3J clonotypes between all six subjects. The V_H and J_H germline gene for each clonotype appears directly above the CDR3 amino acid sequence. Identical CDR3 sequences appearing within multiple clonotypes appear in blue. Clonotypes with the same CDR3 length and one amino acid difference appear in green text; the amino acid change is denoted in red. All underlined text denotes the location of the assigned D_H germline gene. Histograms above each column provide frequencies for the number of matching clonotypes from *A//HIP1+2+3* that were 1, 2 or 3 mismatches (from left to right) from one of the shared clonotypes appearing in the column directly below.