PLOS ONE

RESEARCH ARTICLE

A Public Database of Memory and Naive B-Cell Receptor Sequences

William S. DeWitt[1☯], Paul Lindau[2,3☯], Thomas M. Snyder[1☯], Anna M. Sherwood[1], Marissa Vignali[1], Christopher S. Carlson[2], Philip D. Greenberg[2], Natalie Duerkopp[2], Ryan O. Emerson[1], Harlan S. Robins[1,2]*

1 Adaptive Biotechnologies, Seattle, United States of America, **2** Fred Hutchinson Cancer Research Center, Seattle, United States of America, **3** University of Washington, Seattle, United States of America

☯ These authors contributed equally to this work.
* hrobins@fredhutch.org

Abstract

The vast diversity of B-cell receptors (BCR) and secreted antibodies enables the recognition of, and response to, a wide range of epitopes, but this diversity has also limited our understanding of humoral immunity. We present a public database of more than 37 million unique BCR sequences from three healthy adult donors that is many fold deeper than any existing resource, together with a set of online tools designed to facilitate the visualization and analysis of the annotated data. We estimate the clonal diversity of the naive and memory B-cell repertoires of healthy individuals, and provide a set of examples that illustrate the utility of the database, including several views of the basic properties of immunoglobulin heavy chain sequences, such as rearrangement length, subunit usage, and somatic hypermutation positions and dynamics.

OPEN ACCESS

Citation: DeWitt WS, Lindau P, Snyder TM, Sherwood AM, Vignali M, Carlson CS, et al. (2016) A Public Database of Memory and Naive B-Cell Receptor Sequences. PLoS ONE 11(8): e0160853. doi:10.1371/journal.pone.0160853

Editor: Stephen J Turner, Monash University, Australia, AUSTRALIA

Received: April 5, 2016

Accepted: July 26, 2016

Published: August 11, 2016

Data Availability Statement: All underlying data for this manuscript is available from http://adaptivebiotech.com/pub/robins-bcell-2016 (http://doi.org/10.21417/B71018); as well as from the Dryad Digital Repository at http://dx.doi.org/10.5061/dryad.35ks2.

Funding: This work was funded by grant R01 CA175353 from the National Cancer Institute, National Institutes of Health to H.S.R., and by the W. M. Keck Foundation. Adaptive Biotechnologies, Corp. provided support in the form of salaries and stock options for authors W.S.D., T.M.S., A.M.S., M.V., and R.O.E., but did not have any additional role in the

Introduction

The diverse B-cell repertoire of a healthy individual allows the recognition of a wide range of antigenic epitopes, resulting in a robust adaptive humoral immune response against pathogens. The vast majority of B lymphocytes express a single unique B-cell antigen receptor (BCR), a heterodimeric protein complex composed of a heavy and a light immunoglobulin chain, each of which contains a highly diverse antigen-binding domain. The human immunoglobulin heavy chain (IgH) locus comprises approximately one megabase of chromosome 14, and contains at least 51 functional variable (V) region genes, 25 diversity (D) genes and 6 joining (J) genes that undergo a series of recombination events to assemble a functional heavy chain[1–3]. This recombination process creates a vast array of antigen-binding receptors through the random assortment of different V, D, and J segments (combinatorial diversity), and the insertion of non-templated (N) and palindromic nucleotides (P) at the junctions between V/D and D/J segments (junctional diversity). Productive in-frame VDJ rearrangements result in a functional heavy chain and lead to a permanent alteration of the genomic DNA sequence of a B cell, defining it as a clone. Similarly, the human immunoglobulin light chain κ and λ loci occupy approximately one megabase on chromosomes 2 and 22, respectively, and contain 30–40 V and 4–5 J

study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'Author Contributions' section.

Competing Interests: C.S.C. owns stock and receives consulting fees from Adaptive Biotechnologies. W.S.D., T.M.S., A.M.S., M.V., R.O. E., and H.S.R. are employees of Adaptive Biotechnologies with salary and stock options. There are no patents, products in development, or marketed products to declare. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

segments that can recombine to generate a light chain that is assembled with the heavy chain to form a functional receptor, jointly determining the specificity of recognition[3].

This initial BCR repertoire created in naive B cells through combinatorial and junctional diversity increases upon antigen encounter through the process of somatic hypermutation (SHM), which is mediated by activation-induced cytidine deaminase (AID)[4]. As a result, single base substitutions and occasional insertions or deletions occur throughout the rearranged BCR genes, generating a BCR with increased affinity for its antigen [5, 6]. Our understanding of SHM is limited by the relatively small number of BCR sequences from antigen-experienced B cells that have been available until recently.

The clonal diversity of the human BCR repertoire has been difficult to estimate. Early studies relied on extrapolation from the relatively small number of sequences obtained through low-throughput methods such as immunoscope or traditional Sanger-based sequencing (reviewed in [7, 8]). In recent years, high-throughput sequencing (HTS) methods have considerably increased the number of unique BCR sequences available to the scientific community. However, most of the sequences generated to date are not readily available in a centralized and curated database—the most widely used resource of immune loci (International ImMunoGeneTics, or IMGT®) currently contains approximately 50,000 rearranged human IgH sequences[9]. On the other hand, several other large datasets are publicly available: for example, the National Center for Biotechnology Information (NCBI) Sequence Read Database (SRA, http://www.ncbi.nlm.nih.gov/sra)) includes 454 pyrosequencing data from HIV-1 neutralizing antibodies from the Vaccine Research Center (SRP02639) and antibodies generated in response to influenza vaccination from dbGaP (SRP029381), as well as Illumina sequencing data from healthy donor repertoires from BioProject (SRP037774). In addition to this, a number of publications in the last few years have made considerable numbers of BCR sequences available to the scientific community[10–24]. As a consequence of this recent surge in the number of B-cell sequences available, centralized database and complex data processing and visualization tools are needed to analyze, visualize and interpret these large datasets of immune sequences.

Immunosequencing of the TCR and BCR repertoires has greatly improved our understanding of B- and T-cell biology[25], leading to the refinement and modification of B- and T-cell development models[26–29]. In addition, these data have resulted in multiple clinical advances. For example, immunosequencing has resulted in clinical tests for diagnosis and monitoring of minimal residual disease for lymphoid malignancies[23, 30], has guided the discovery of neutralizing antibodies against HIV[31], has been used to dissect the role of T-cells in autoimmuny[32, 33] vaccination[34] and transplant[35, 36], and to better understand the role of infiltrating T lymphocytes in ovarian cancer[37], melanoma[38] and glioblastoma[39].

Here, we present a public resource of more than 37 million unique immunoglobulin heavy chain (IgH) sequences resulting from the digital amplification and sequencing of the most variable region of the IgH gene from 10 million naive and 10 million memory B cells each from three healthy adult donors, using the immunoSEQ platform[18, 27, 40]. In addition, we have created a suite of software tools that facilitates the visualization and analysis of these data. Using many barcoded replicates for each sample, our method approximates single-molecule sequencing of BCRs at high-throughput, thus ensuring a faithful quantitative representation of nearly all clones present in the biological sample. Besides describing the study design, the specifics of the sequencing technology employed, and the resulting data set, we illustrate the use of the web-based tools developed to enable visualization and analysis of these data.

Finally, to further demonstrate the utility of this resource, we explore a few of the many potential biological questions that can be addressed through our data set: (1) we explored and compared the clonal diversity of naive and memory BCR repertoires at an hitherto unprecedented level of sequencing depth; (2) we confirmed V gene family usage patterns in healthy subjects using a bias-free approach; (3) we examined variations in the length of the third Complementarity Determining Region (CDR3) in naive and memory B-cell populations; (4) we analyzed SHM within the steady-state BCR repertoire; and (5) we deconvoluted patterns of SHM substitutions in V genes for naive and memory cells.

Materials and Methods

Sample source and B cell isolation procedure

Whole blood samples were collected from three 25–40 year old Caucasian males participating in a study of healthy human volunteers under approval of the Fred Hutchinson Cancer Research Center Institutional Review Board. The donors did not report any infections or vaccinations in the 6 months previous to sample collection. All donors provided written informed consent. All samples were processed less than 2 hours after venipuncture. Peripheral blood mononuclear cells were separated from 400 mL of whole blood by Ficoll (GE Healthcare) gradient density centrifugation at 400g and 22°C. Next, total B cells were enriched from PBMCs using CD19 MicroBeads and the autoMACS Pro Separator (Miltenyi Biotec). B cells were then stained with anti-CD19APC, anti-CD3FITC, anti-CD27PE, anti-IgM-APC750, and anti-IgD-PECy7 (all from BD BioSciences) and sorted using the BD FACS Aria II with FACSDiva v6.1.3 software (BD BioSciences). Naive (CD19$^+$, CD27$^-$, CD3$^-$, IgM$^+$, IgD$^+$) and memory (CD19$^+$, CD27$^+$, CD3$^-$) B cells were sorted to a purity of 97% or greater. Sort purity was assessed by passing a small sample of each sorted population back through the flow cytometer. We note that this memory B-cell sort contains all CD27$^+$ B cells, including both class-switched and IgM memory B cells. Representative flow cytometry plots of CD27 versus IgD expression on gated CD19$^+$ B cells and CD27 versus IgM expression on gated CD19$^+$CD27$^+$ B cells are shown in S1A and S1B Fig, respectively.

Sorted B-cell populations were pelleted at 300g at 4°C, and finally flash frozen in liquid nitrogen before being stored at -80°C. Genomic DNA was purified from sorted B cell populations using the QIAmp DNA Blood Mini Kit (Qiagen). Genomic DNA was normalized and the equivalent of 50,000 cells was dispensed each of 188 wells of 96-well plates.

PCR amplification and reduction of PCR bias

To amplify the CDR3 region of IgH, we used a 2-PCR reaction approach as previously described [23]. Briefly, the first step consists of a multiplex PCR that uses gene specific V-forward and J-reverse primers that bind to 47 V and 6 J functional genes as well as many of the pseudogenes for both V and J. The primers are designed for perfect complementarity to the germline V and J gene targets. In addition, the final five nucleotides of each primer were selected so as to bind to sequences that are much less likely to be affected by SHM[41]. The second PCR adds Illumina adaptor sequences and well-specific barcodes, for a total of 31 cycles of amplification.

Despite efforts to achieve consistent melting temperatures (T_m) between all the V and all the J primers, there is a wide variation in amplification efficiency. To remove this bias, we created a synthetic set of IgH receptors with universal flanking sequences that allow for direct sequencing on the Illumina platform[40]. The synthetic genes include all V-J combinations labeled with barcodes that allow for the ready identification of each template. This synthetic immune system is sequenced directly to precisely determine the abundance of each template. Then, multiplex PCR amplification with the V and J gene primers is performed on the synthetic pool and the resulting DNA is also sequenced. Comparing the known starting abundances with the resulting amplified sequences, we are able to assess the relative amplification efficiency of each V and J primer. We then modify the concentration of the primers that over and under amplify. The process is iterated several times until the majority of the bias is removed. We have shown that the results of this process are robust to variations in the length, GC-content, and overall abundance of the template.

Resolution of nucleotide sequences

To measure the amount of nucleotide assignment error in our analysis, we randomly selected molecules from the PCR amplified library of IgH receptor sequences, and sequenced them at a depth of at least 10 times the starting template quantity. In other words, since each well contained approximately 50,000 B cells, we aimed to sequence at least 500,000 molecules from each PCR library. This ensured that, even with some amplification variation and random sampling error, multiple copies of each template would be sequenced. Due to the very low error rate in Illumina sequencing (~.1%), the number of errors in a 130-basepair sequence is roughly distributed as $k_{error} \sim Bin(n = 130, p = .001)$, from which we compute $\Pr(k_{error} = 0) \cong .88$, and $\Pr(k_{error} = 1 \mid k_{error} > 0) \cong .94$. Thus, ~90% of all our templates result in no PCR or sequencing errors. Of the remaining ~10%, the large majority contain a single error. Given that these errors are not systematic, any particular error is almost always unique. Thus, we are able to readily correct these errors by identifying reads present once in the data set that differ by a single nucleotide from a sequence present multiple times, and collapsing them into the predominant clone. Additionally, since memory samples were found to have many more clones present in multiple wells, error correction was performed on data aggregated from all wells of a given sample. This ensures consistent consensus sequence assignment across wells. In terms of the diversity inference described below, this method of collapsing errors across wells is intrinsically conservative.

Germline annotation of nucleotide sequences and SHM detection

The CDR3 region was identified according to the standard previously determined by the IMGT collaboration[9]. Identification of the V, D, and J gene segments was performed using a scored alignment across a definition list of all known V, D, and J gene and allele members from IMGT. The most likely assignments (allowing for ties for similar gene sequences) for each gene segment were then added to the sequence reads as their germline annotation. Somatic hypermutation was calculated over just the V gene segment, based on sequence variations from the assigned germline gene/allele match.

Estimation of repertoire diversity from replicate occupancy data

To estimate clonal diversity, we derived an extension of an established sampling model in ecology and corpus linguistics: the Poisson abundance model[42–44]. This allows the construction of a likelihood function for replicate occupancy data parameterized by the richness and abundance distribution of the repertoire. Briefly, we synthesized the combinatorial probability of the replicate occupancy of a clone conditioned on sample abundance, with the Poisson abundance model of sample abundance conditioned on repertoire parameters. Analytically marginalizing over sample abundance as a latent variable, we formed the desired likelihood function and deployed tandem numerical and analytical optimizations facilitated by an asymptotic approximation for large richness. The full mathematical derivation and computational validation of this model can be found in the Supporting Information (S1 Method).

Results and Discussion

Immunosequencing of naive and memory B cells

In healthy adults, CD19^{+} B cells comprise 7–11% of lymphocytes circulating in peripheral blood[45]. This population is dominated by naive B cells, which correspond roughly to 65% of all peripheral B cells, while memory B cells account for about 30% of all circulating B cells[45]. To faithfully capture the breadth of the B-cell repertoire, we isolated naive (N, CD19$^+$ CD27$^-$ IgD$^+$ IgM$^+$) and memory (M, CD19$^+$ CD27$^+$) B cells from 400 mL of peripheral blood obtained from each of 3 healthy adult donors (D1, D2 and D3)[46]. Additionally, in order to estimate the reproducibility of the approach, we included two biological replicates of the naive B-cell sample from Donor 1 (i.e. D1-Na and D1-Nb).

These samples yielded 2–4 x 10^7 naive B cells and 1.5–2 x 10^7 memory B cells at greater than 97% purity from each donor. Considering that the approximately 5 L of peripheral blood of healthy adults is estimated to contain on average 6.5 x 10^8 naive B cells and 3.0 x 10^8 memory B cells[45], we calculate that by using a 400 mL sample, we captured 3.1–6.1% of the naive and 5–6.7% of the memory B cells circulating in peripheral blood, respectively.

Next, we sequenced a segment of the immunoglobulin heavy chain (IgH) gene from the naive and memory B cell populations purified from each donor that includes CDR3[18]. Since the CDR3 rearranges somatically during B cell development, the resulting sequences can be used to define unique B-cell clones, in the sense of descendants from a common naïve B cell; however, somatic hypermutation means that even among mature B cells that share a CDR3 by common descent, there can be additional sequence differences in e.g. the CDR1 and CDR2 regions.

In brief, for each of the samples, we extracted genomic DNA and we dispensed an amount corresponding to ~10^7 naive or memory B cells into 188 wells of two 96-well plates (the remaining wells were used for controls). This resulted in the allocation of the equivalent of approximately 50,000 cells per well (Fig 1A). We then performed a two-step PCR, including a multiplex step that uses V and J-specific primers to amplify a region of the IgH gene, followed by a second amplification that adds unique well-specific barcodes and Illumina adaptors. Next, we used a HiSeq instrument to sequence a 130 nt-long segment of the IgH gene that includes the CDR3[18]. This approach enabled us to sample the naive and memory repertoires of B cells of three healthy individuals to a depth much greater than other studies.

The value of the resulting dataset depends both on the accuracy of the IgH nucleotide sequences and the quantitation of the abundance of each B cell clone. Importantly, there are two major obstacles that hinder the quantitative immunosequencing of IgH genes. The first challenge, which is shared by other immune genes such as those encoding for T-cell receptors, arises from the process of gene rearrangement and the resulting intrinsic diversity of both types of immune receptors. The second challenge, unique to B cells, results from the additional level of divergence from the genomic sequence generated by SHM in antigen-experienced cells. Our approach to address these challenges is described in the Material and Methods section, and our analytical approach is described in detail in the S1 Method included in the Supporting Information section.

In brief, we used a digital counting method that yields counts of clones based on their presence or absence in each of the 188 wells, as diagrammed in Fig 1B. Quantitative accuracy is achieved by inclusively sequencing the receptors in each uniquely-barcoded well. We aimed for a minimum of 10-fold coverage of each BCR molecule in each well, and achieved an effective coverage that ranged from 8 to 12 average reads per template in the different samples. We also analyzed the distribution of the number of unique productive BCRs over the 188 wells for each sample, as shown in S2 Fig. Most of the samples had an average of 40,000 unique productive rearrangements per well, with the exception of the naive sample from Subject 2, which had a lower number of unique productive rearrangements per well.

Our method is binary, since we only consider presence or absence of each sequence in each well, and robust against a wide range of amplification efficiencies. The sequences in each well are identifiable by the presence of the unique barcode assigned to that well, and thus we report an "occupancy" value for each BCR sequence, which corresponds to the number of wells it was

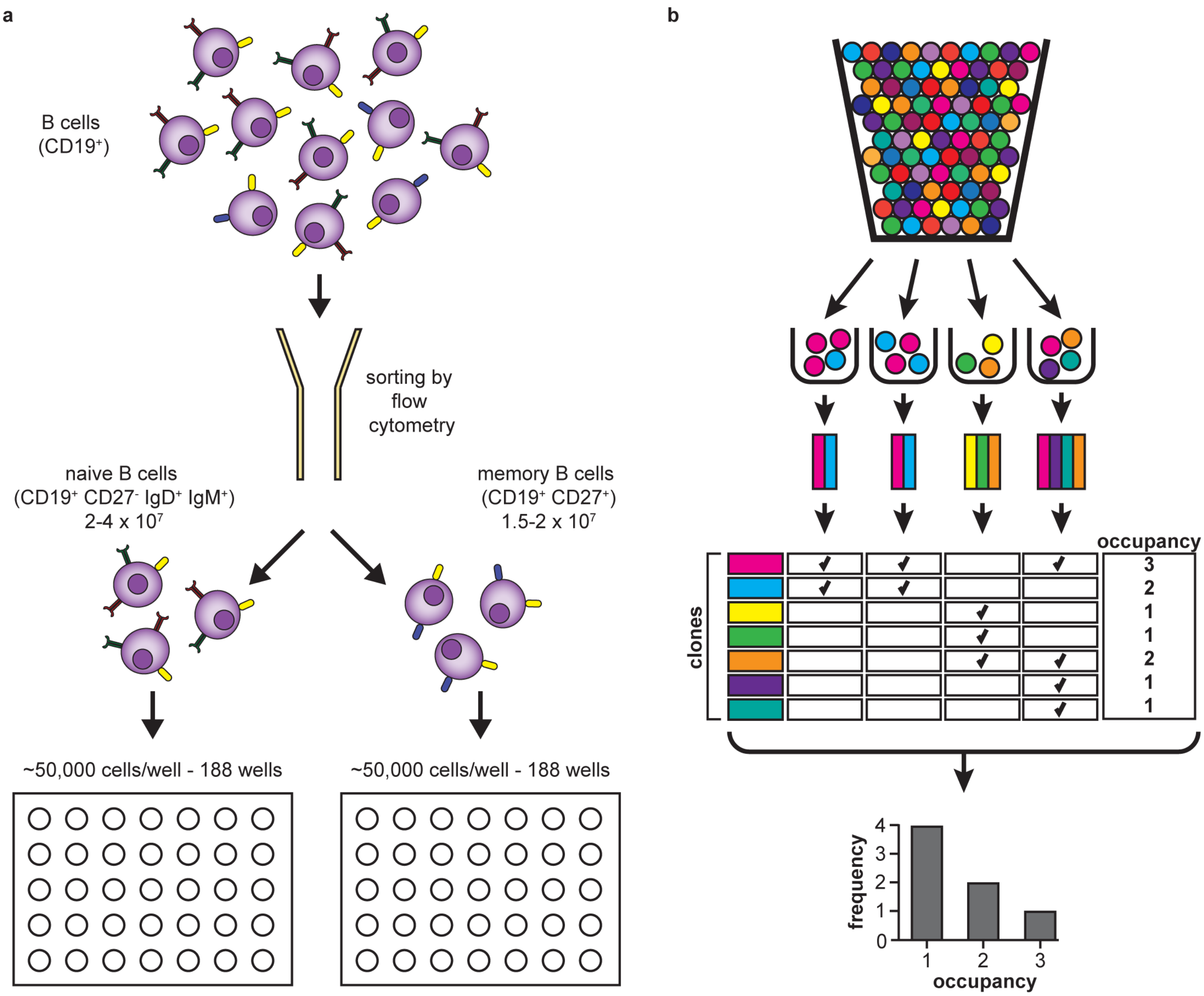

Fig 1. Experimental and informatic design. (a) Peripheral blood samples from three healthy donors were sorted using flow cytometry to isolate naive ($CD19^+$ $CD27^-$ IgD^+ IgM^+) and memory ($CD19^+$ $CD27^+$) B cells. For each sample, approximately 10^7 cells were distributed into two 96-well plates (i.e., into 188 wells, resulting in ~50,000 cells per well), and processed by immunosequencing. **(b)** Schematic of the 'urn sampling' quantitation method. Cells are represented by colored balls, with each color indicating a different clone identity. Each ball (cell) is randomly allocated to a sample bin (well). Occupancy is calculated after censoring count information, and thus is expressed as presence or absence. The majority of clones are present in just one out of 188 wells, indicating that they were almost certainly represented by a single cell in the original sample.

doi:10.1371/journal.pone.0160853.g001

observed in. Clones with abundance in the repertoire of less than 1:1,000,000 B cells (i.e. the vast majority of all B-cell clones) will rarely be present more than once in any well. Therefore, for most B cells, their sample abundance will be equal to the number of wells they are observed in. We determined that the vast majority of clones have an occupancy value equal to 1 (Fig 2A). Since multiple cells of the same clone are unlikely to appear in any given well, this strongly implies that a single cell out of the initial 10^7 expressed that particular BCR sequence. As occupancy increases, this metric becomes a decreasingly precise (and increasingly negatively biased) estimator for sample abundance, since the incidence of multiple occurrences of a given clone in a single well becomes more probable.

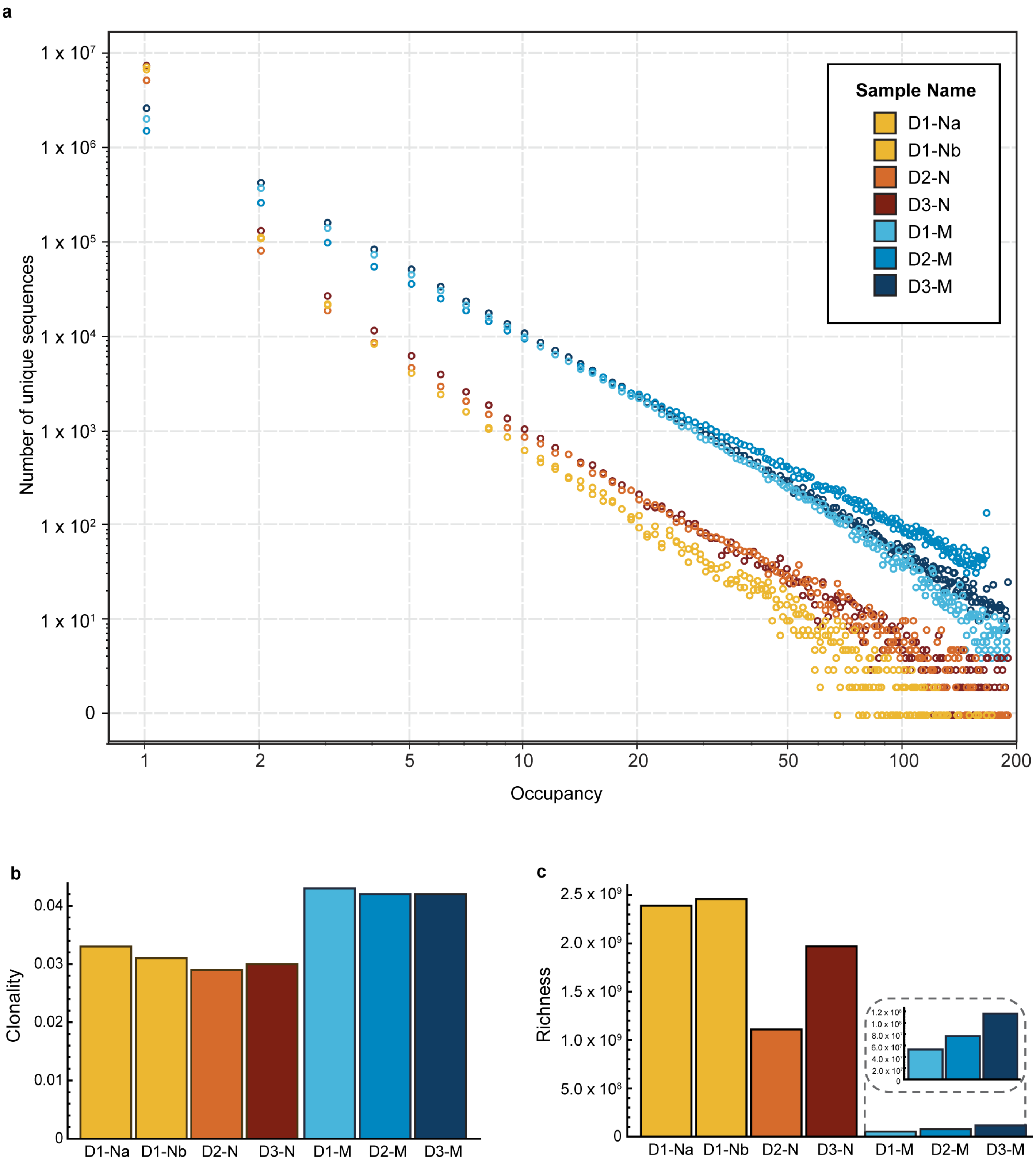

Fig 2. Inference of diversity in the naive and memory B-cell repertoires. (a) The graph shows the distribution of unique sequences, as the number of unique sequences (y-axis) versus their occupancy (x-axis) for the naive (orange) and memory (blue) samples for the three donors (D1, D2 and D3, including two technical replicates for the naive sample from Donor 1). The vast majority of the sequences have occupancy of 1. **(b)** Clonality index for all samples. **(c)** Richness index for all samples. While the clonality index is higher for memory samples, the richness index is higher for the naive samples.

doi:10.1371/journal.pone.0160853.g002

Diversity of the naive and memory B cell receptor repertoires

We first compared the overlap between the naive and memory B-cell repertoires of the three donors studied (Table 1). For this analysis, we only considered exact sequence matches.

For each sample obtained from each of the donors (D1-Na, D1-Nb and D1-M; D2-N and D2-M; and D3-N and D3-M), the table indicates the pairwise overlap between repertoires, computed as the fraction of the unique sequences for each sample in the rows labeled to the left that are also found (with no mismatches allowed) in the each of the samples listed in the columns. The color gradient of the cells indicates the degree of overlap, with higher overlaps indicating a darker shade of red.

Due to the intrinsically large size and diversity of the B-cell repertoire, the overall overlap between samples is small. However, as expected, it is higher between the two independent replicates of the naive repertoire of Donor 1 than between those of different donors. Also, the naive and memory B-cell populations of each donor are more similar to each other than to those of different donors.

Next, for each sequence present in the data we computed the maximum well occupancy among all samples (a measure of clonal abundance), and also the number of subjects the sequence was observed in. S3 Fig shows the distribution of maximum occupancy among sequences found in only 1 subject, in any two subjects, and in all three subjects. We observe that shared sequences (those present in two or three subjects) tend to have higher maximum occupancy. This could be the result of shared memory cells resulting from common pathogen exposures among subjects, or alternatively, the consequence of recurrent generation of high-probability V(D)J recombinations that are identical by state but not by descent in different individuals.

We also estimated the clonal diversity of the repertoires–i.e. the number of distinct somatically rearranged receptors present in each repertoire and their relative abundances–which defines the search space available for immune recognition and is therefore essential for the quantitative characterization of the BCR repertoire. For each sample, we inferred two diversity indices: *richness*, defined as the number of distinct clones, and *clonality*, a measure of abundance uniformity that ranges from 0 (maximally uniform) to 1 (most disparate, or clonally dominated; see the Materials and Methods and S1 Method for a detailed description of these indices). Fig 2B shows the maximum likelihood estimates of clonal diversity. Using either diversity metric, the samples cluster distinctly by cell type, and these results were consistent across individuals. As expected, our results indicate that memory clones have more disparate repertoire abundances (higher clonality) than naive clones, and that naive clones are extremely diverse.

Our replicate PCR well methodology accurately assesses the abundance of nearly all B-cell clones in each sample. A small number of memory clones are present at high frequency, and

Table 1. Overlap among the naive and memory repertoires of the three donors.

		1			2		3	
		Na	Nb	M	N	M	N	M
1	Na		4.18E-03	7.87E-04	4.41E-04	8.34E-05	6.96E-04	1.07E-04
	Nb	4.55E-03		8.21E-04	4.44E-04	8.42E-05	6.92E-04	1.08E-04
	M	1.75E-03	1.68E-03		5.42E-05	9.48E-06	8.90E-05	2.03E-05
2	N	6.55E-04	6.06E-04	3.61E-05		2.38E-03	7.06E-04	1.15E-04
	M	2.42E-04	2.25E-04	1.24E-05	4.67E-03		2.65E-04	5.43E-05
3	N	6.66E-04	6.09E-04	3.83E-05	4.56E-04	8.74E-05		4.07E-03
	M	1.94E-04	1.80E-04	1.65E-05	1.40E-04	3.38E-05	7.70E-03	

doi:10.1371/journal.pone.0160853.t001

thus are found in all or nearly all of the replicate PCR wells. This is expected to cause negative bias in the clonality inferences for the memory populations. Despite this conservative bias, the memory and naive populations cluster distinctly.

The inferred richness of the naive B-cell repertoire is of a similar magnitude to the expected abundance of naive B cells in the peripheral blood (~1x10^9)[45], suggesting that the typical naive clone does not undergo proliferation prior to antigen encounter. In contrast, the richness of the memory B-cell population is consistent with each clone undergoing several divisions on average. The relatively higher clonality observed for memory cells as compared to naive cells indicates that a small percentage of these clones experience significant proliferation. Our conclusion that the typical naive B-cell clone undergoes no proliferation prior to antigen encounter raises questions regarding previous calculations that suggested that naive B cells in the peripheral blood of adults undergo approximately 1.9 cycles of homeostatic proliferation on average [47]. However, it is important to point out that the study by Van Zelm *et al.* uses an indirect method of estimating the replication history based on deletion circles, and that, unlike our approach, it does not have the ability to resolve distinct clones. On the other hand, we do not measure replication history and instead calculate it from the diversity metric and estimates of the number of B cells in the periphery reported in the literature. Thus, both sets of results are not directly comparable and do not necessarily contradict each other.

In summary, our data confirm that the naive repertoire of a healthy adult is extremely rich, and thus suggests that the typical naive B-cell clone undergoes no proliferation prior to antigen encounter, while we observe that memory B-cell clones undergo several cycles of division on average. Future studies will focus on mining this extremely deeply-sequenced data to further understand ongoing maturation of clones within the memory compartment at steady state. The assay also has the potential to determine whether the different subsets of cells contained in the memory compartment (i.e. switched memory cells, unswitched memory cells, as well as any plasmablasts or plasma cells present due to ongoing immune responses) possess different distributions of mutation rates.

Examples of possible explorations of this dataset

To demonstrate that our data are accurate and of high quality, we made use of these tools to answer several fundamental questions about the B-cell repertoire in healthy individuals. In addition to the clonal diversity inferences described above, we provide a set of four examples that illustrate the utility of the data set and the related analysis tools. For each of these examples, we created a dashboard in the immunoSEQ Analyzer workspace (http://adaptivebiotech.com/link/publicBCellResource) so that the analysis of each example and the accompanying visualizations that follow can be reproduced by the user.

Example 1: Characterization of IGHV family and gene usage. The IGH V locus contains over 50 functional genes (depending on the individual's haplotype) that are classified into 7 families based on nucleotide sequence homology[48]. Each gene segment has a certain likelihood of undergoing rearrangement and being incorporated into a mature immunoglobulin molecule, and in addition the process of negative selection of immature B cells further restricts V gene segment use, resulting in an unequal representation of V gene families in the naive B-cell repertoire. Similarly, the positive selection of naive B cells to populate the memory compartment results in variations in V gene segment representation[13, 49].

Traditionally, standard measurements of TCR usage in T cells have utilized PCR-based V beta spectratyping (reviewed in reference [7]), but no equivalent approach exists for the analysis of V gene usage in B cells. However, recent immunosequencing approaches have begun to shed light on B-cell gene usage[18, 19]. To assess the broad similarities and differences in gene

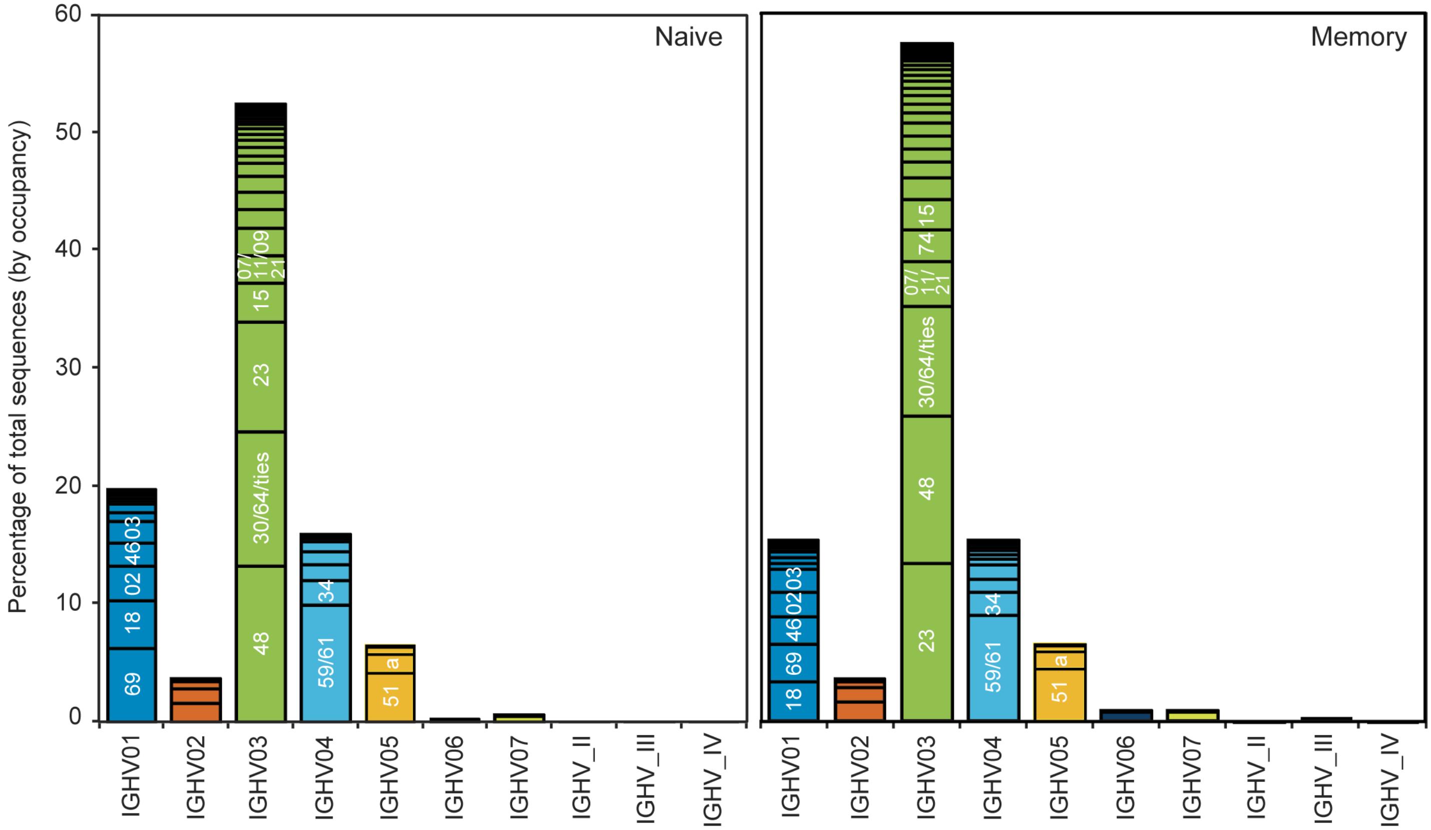

Fig 3. V family and V gene usage patterns. The histograms show the relative percent of total sequences (by occupancy) for each of the IGHV families (as shown under the graphs), for the naive (left panel) and memory (right panel) samples, aggregated for the three donors. Within each family, discrete bands represent each of the individual genes. The most abundant genes within each family are indicated (e.g., 69 in IGHV01 refers to the gene IGHV01-69). Overall, memory samples contain fewer IGHV01 and more IGHV03 family sequences than naive samples, with some gene-level differences evident as well.

doi:10.1371/journal.pone.0160853.g003

usage between the naive and memory B-cell repertoires, we compared the IGHV family and gene usage in naive and memory B cells in three healthy donors (Fig 3). In agreement with previous reports, we found that the IGHV3 gene family is utilized most commonly in both repertoires[49, 50]. Moreover, we observed that, in these subjects, IGHV3-48 is the most commonly used V gene in the naive repertoire followed by IGHV3-30 or IGHV3-64, two genes that are indistinguishable over the region covered by the sequence reads. In the memory repertoire, IGHV3-23 is used most commonly, followed by IGHV3-48. We found that the second most commonly expressed gene family in the naive repertoire of these subjects corresponds to IGHV1, followed by IGHV4. In contrast, the memory repertoire has equivalent representation of the IGHV1 and IGHV4 gene families. At the gene specific level, we observed a decrease in the relative frequency of IGHV1-69 and IGHV1-18 within the IGHV1 family in memory compared to naive B cells, consistent with previous studies[13]. Taken together, these data, which were obtained from a single experiment, reproduce observations from several previously published studies [13, 18, 49–51], validating the utility of this dataset.

Example 2. Measurement of CDR3 length distribution. The immunoglobulin CDR3 is the most important determinant of antibody-antigen recognition[52, 53]. Its length varies mostly due to recombination, and can also change slightly from SHM. Therefore, we compared the CDR3 length distribution of the naive and memory repertoires to understand both the limits and flexibility of the antigen-binding capacity of B cells. We found the average CDR3 length

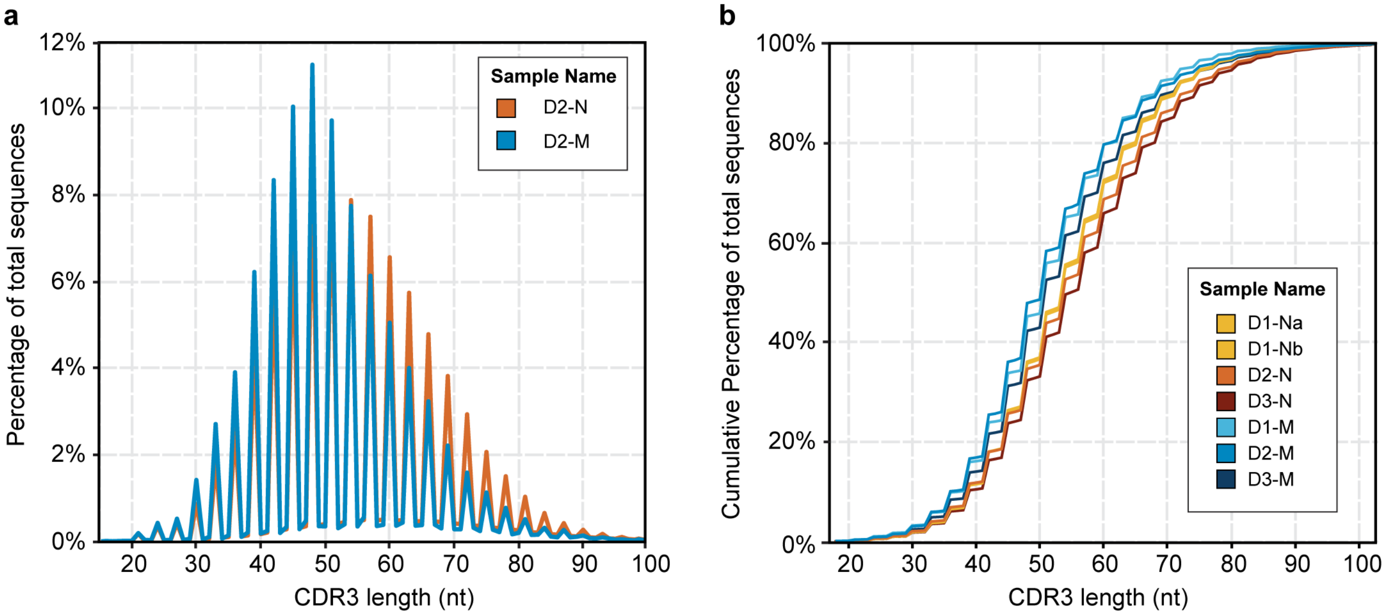

Fig 4. Comparison of CDR3 lengths in naive versus memory B-cell samples. (**a**) The graph shows the normalized percentage of total sequences for the naive (orange) and memory B cells (blue) from donor D2. (**b**) The graph shows the cumulative percentage of total sequences at a given CDR3 length for all naive and memory samples, as indicated in the inset. The technical replicates for donor D1 overlap closely and are not distinguishable in this figure. The memory repertoire is consistently 3 nucleotides (or 1 amino acid) shorter than the naive repertoire at the same cumulative frequency.

doi:10.1371/journal.pone.0160853.g004

in the naive B-cell repertoire to be 48 nucleotides, while the memory B cells had, on average, a CDR3 length of 45 nucleotides (Fig 4). Unproductive CDR3 sequences have an even longer average size (~60 nt) than that seen for productive sequences in naive or memory cells. These two facts suggest that, while the B-cell recombination process generates long and highly diverse CDR3 regions, functional clones that become part of the memory repertoire are biased towards shorter CDR3 sequences. In addition, we observe that there is a greater variability in CDR3 length in naive cells compared to memory cells, suggesting that the naive repertoire has the potential to bind a wider range of antigens than are actually encountered by the donors in this study. These data agree with previous findings [13, 18, 54], further confirming the validity of our dataset.

Example 3: Assessment of purity of flow-cytometry sorted cell populations. Since SHM occurs during antigen-induced maturation, a naive B cell is characterized by the absence of substitutions in its germline V gene[5]. Thus, to examine the purity of our sorted B-cell populations, we determined the rate of substitutions in the V genes of the naive and memory B-cell repertoires (Fig 5). Approximately 95% of sorted naive B cells displayed no V gene substitutions, and had low clonal abundances, which are typical of naive cells. In contrast, memory B cells harbored an average of 3–4 substitutions per 100 nt in the V genes, and additionally displayed a much broader range of clonal abundances, as expected of antigen-experienced B cells. Taken together, these analyses suggest that our method accurately and faithfully captures the circulating B-cell populations.

Example 4: Analysis of somatic hypermutation in memory B cells. Affinity maturation, including somatic hypermutation and class switching, is critical to the production of functional antibodies[55–60]. We were able to easily define somatic hypermutation sites by identifying variations from germline sequences within the sequenced region of the V gene. While a certain number of single nucleotide variations in the V gene may result from inherited SNPs, a review

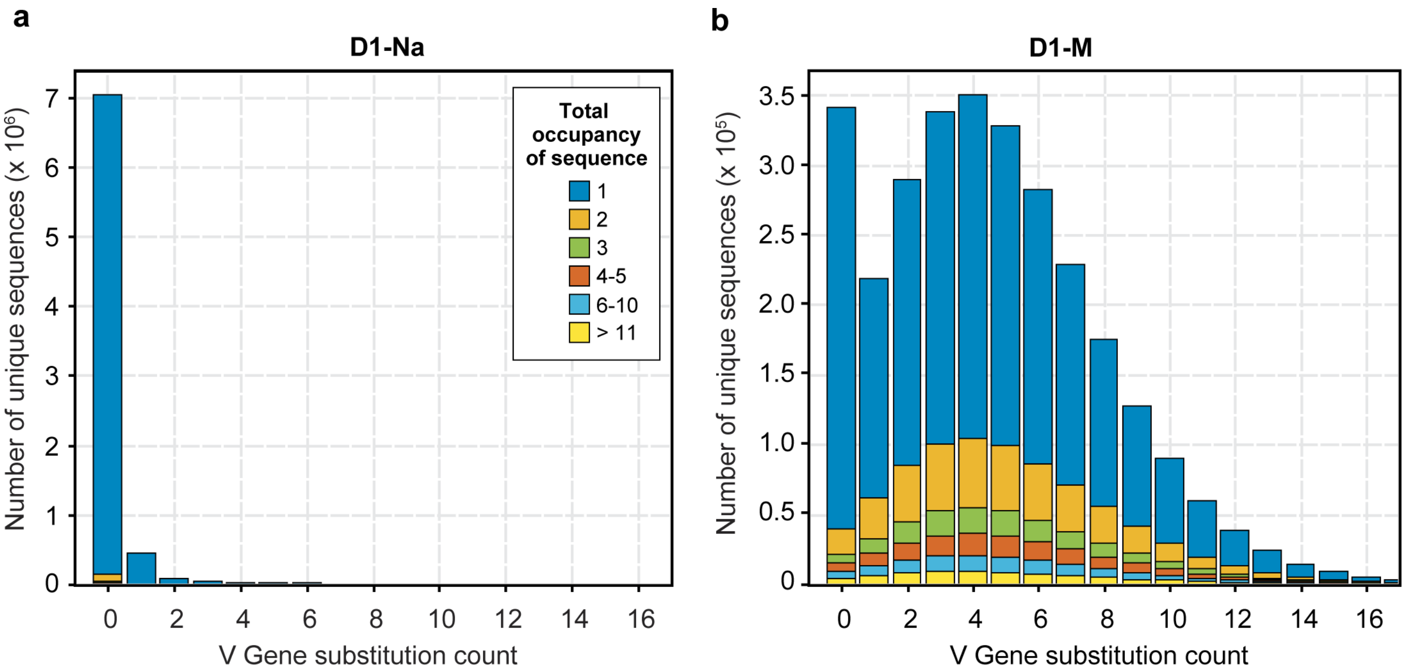

Fig 5. Comparison of Somatic Hyper Mutation in paired naive and memory B-cell samples from the same donor. The figure shows data for the naive (**a**) and memory sample (**b**) from Donor 1, which is representative of all three donors. The x-axis corresponds to the number of substitutions differing from the germline V gene sequence, and the y-axis indicates the number of unique sequences that display that number of substitutions. The colors indicate different total well occupancies, with blue indicating singletons present in just one well, and the other colors showing progressively higher well occupancy, as indicated in the figure. The majority of the sequences in the naive B-cell sample have 0 substitutions and correspond to low abundance clones observed in a single well (blue). In contrast, the memory B cell sample from the same individual shows a much broader distribution of substitutions, as well as many more sequences with occupancy greater than 1.

doi:10.1371/journal.pone.0160853.g005

of the V gene sequences observed in naive cells in the same individual makes it easy to exclude this possibility in most cases.

After identifying likely somatically hypermutated residues in the V gene segments, we created a set of tools to view these data for all genes and samples over the sequenced V gene region. Fig 6 shows an example of the resulting data for gene IGHV1-69. Our analysis and visualization tools allow a clear visualization of SNPs and transition/transversion rates (top panel), as well as overall SHM rates by position (middle panel) gleaned from our very deep sampling of memory B-cell sequence data. In addition, several reported hotspot and coldspot AID targeting motifs[61] can be evaluated (bottom panel). The most frequently reported hotspot motif (most generally described as GYW/WRC on the two strands[6]) accounts for many of the observed positions with high SHM levels, while some nucleotides that display SHM, such as nucleotide 267 in several V genes including V01-69 and V03-23, are not part of a known hotspot motif. It is possible that mutations of this position, which flanks the CDR3, might have increased functional importance for improved antibody binding, despite the absence of known AID-targeting motifs.

Conclusions

In this study, we provide the research community with an accurate and rich dataset of BCRs, as well as a set of straightforward tools to enable its in-depth study. By combining flow cytometry purification of peripheral B cells with high-throughput immunosequencing of 10 million naive and 10 million memory B cells from each of three healthy adult donors, we generated a BCR

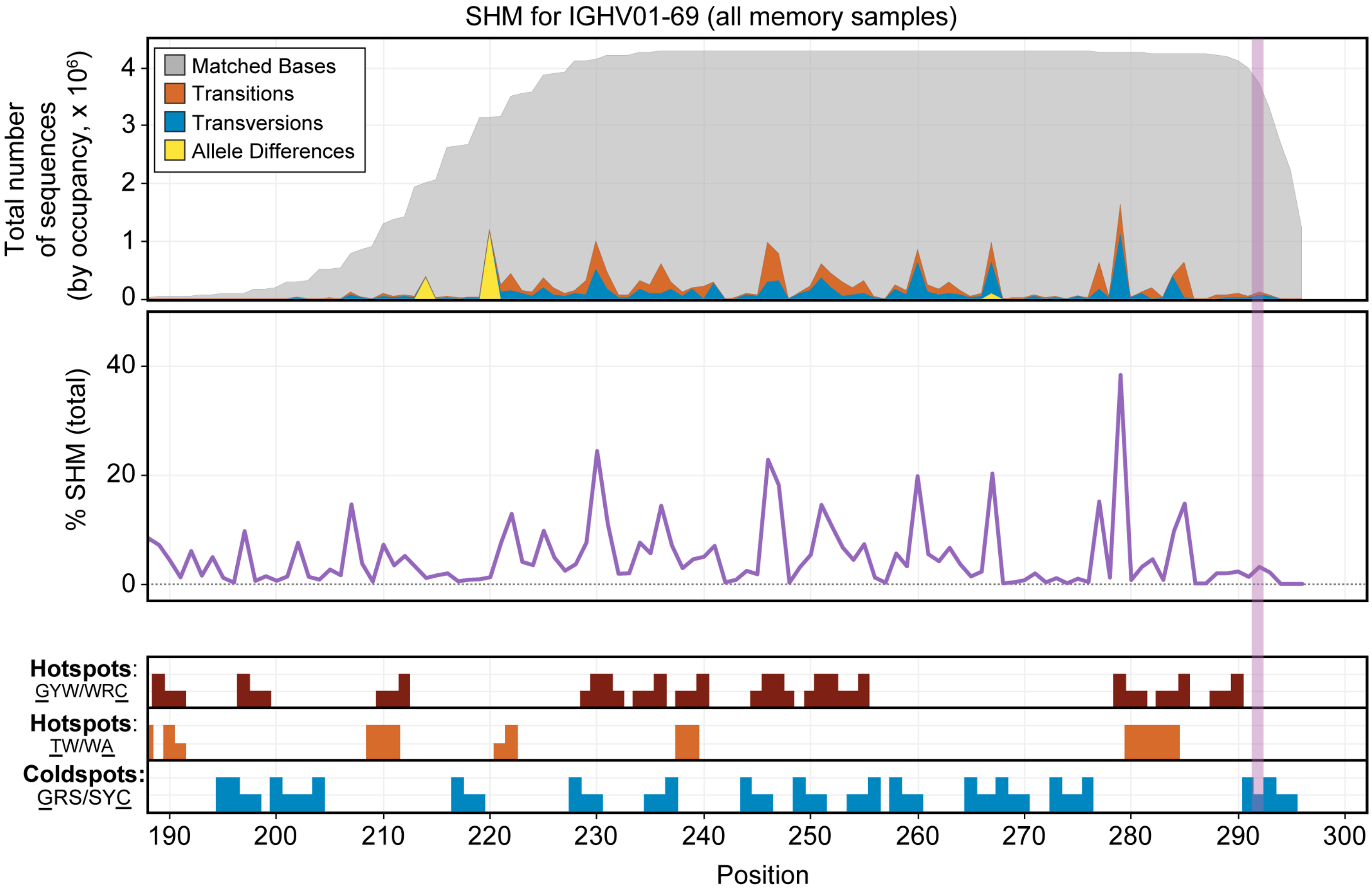

Fig 6. Somatic hypermutation pattern observed over the sequenced region of the IGHV01-69 gene. The figure includes combined data from the memory B-cell population for all 3 donors. The top panel shows the total distribution of sequenced bases by occupancy for the primary allele of IGHV01-69. Nucleotides that match the germline sequence are displayed in gray. Transitions are shown in orange and transversions in blue. Allelic differences, which are also seen in the naive samples, are indicated in yellow. The vertical dotted line marks the average start of the CDR3 region. The middle panel shows the normalized percentage SHM by base for this gene across the memory B cell samples for all three donors. The bottom panel shows suspected SHM hotspot (red and orange bars) and coldspot (blue bars) motifs present in the sequence of this gene over the region assayed. Positions with higher bars indicate bases targeted within the motif (underlined in the legend to the left). The GYW/WRC pattern (red) explains most of the significant sites of SHM for this gene, but some spots of high mutation are not captured by the displayed motifs. In the data viewer, this view can be generated for any V gene and for any combination of data sets.

doi:10.1371/journal.pone.0160853.g006

sequence library containing more than 37 million unique BCR sequences. Whereas some of the currently existing databases, such as IMGT[9], contain a large number of curated IgH sequences from many individuals, this method allowed us to probe the B-cell repertoire of a small number of individuals at an unprecedented depth. In parallel, we developed set of tools tailored to analyze and visualize the resulting data set, which can be accessed from http://adaptivebiotech.com/pub/robins-bcell-2016 (please follow the 'Advanced Visualizations' link).

As an example of the utility of our dataset, we assessed a fundamental property of the BCR repertoires, i.e. their clonal diversity. To do this, we approximated high throughput digital cell counting using a multi-replicate experimental design, and we inferred the clonal diversity of the memory and naive BCR repertoires of three healthy adults using a novel likelihood model.

To further illustrate the utility of these data and the associated tools, we present several other examples that assess general properties of B-cell repertoires that have been previously investigated at a smaller scale, including V gene family usage patterns; the length of CDR3 regions; the numbers of SHM substitutions, and the patterns and types of SHM in naive and memory B cells. Importantly, our observations match previous reports and thus confirm the robustness of our dataset.

Finally, the many-replicate experimental design employed in this study, in which each of the 188 PCR wells corresponds to a replicate sample, constitutes a sample abundance probe robust to the inherent stochasticity of PCR amplification. Moreover, this approach represents a crucial quantitative advance over previous sequencing studies of antigen receptor repertoire diversity, which have been limited by either poor quantitation or by the lower throughput of single-cell methods[27, 62, 63]. We expect that these data will be used by other experts in the field of immunology to address additional fundamental questions about BCR development and *in vivo* antigen binding in humans.

Data Availability

Access to the data set resulting from the experiments described in this study (both at the well level and at the sample level), as well as a link to the tools we developed to enable the analyses presented herein, can be found at http://adaptivebiotech.com/pub/robins-bcell-2016. We have also assigned a unique identifier to this dataset: http://doi.org/10.21417/B71018. The immunoSEQ Analyzer interface includes several tools that can be used to perform further analyses of the data. The "Advanced Visualization" link found in the landing page for this dataset enables access to Fig 2 to Fig 6 in this study, and each of them is followed by a set of interactive dashboards that allow viewing different aspects the data, such as Occupancy (data underlying Fig 2), VDJ tools (data underlying Fig 3), CDR3 tools (data underlying Fig 4), Substitutions tools (data underlying Fig 5), and SHM tools (data underlying Fig 6). Most dashboards include a sample selection option: data are coded by sample type (naive vs. memory) and for each of the three donors studied (including the two repeats for the naive sample from donor 1). Several of the dashboards include filters that allow viewing subsets of the data (e.g. sequences for productive vs. non-productive rearrangements, out-of-frame sequences or sequences with STOP codons). The code for the tools developed for the analysis can be downloaded from the Public B cell dataset code link.

Finally, the full dataset can also be downloaded from the Public B cell dataset link, as well as from the Dryad Digital Repository at http://datadryad.org/resource/doi:10.5061/dryad.35ks2.

Supporting Information

S1 Fig. Representative contour plots of peripheral blood B-cell subsets.
(PDF)

S2 Fig. Distribution of the number of unique sequences across 188 wells for each sample used in this study.
(PDF)

S3 Fig. Distribution of maximum occupancy among sequences found in only 1 subject, in any two subjects, and in all three subjects.
(PDF)

S1 Method. Replicate immunosequencing as a robust probe of antigen receptor repertoire diversity.
(PDF)

Acknowledgments

We are grateful to Dr. Christopher Tipton for helpful suggestions, and to Katie Moran for help with editing. W.S.D. thanks Kameron Decker Harris for helpful mathematical discussions and

Erick Matsen for comments and corrections to the S1 Method included in the Supporting Information.

Author Contributions

Conceived and designed the experiments: WSD PL TMS AMS CSC PDG ND ROE HSR.

Performed the experiments: PL AMS.

Analyzed the data: WSD PL TMS MV ROE.

Contributed reagents/materials/analysis tools: HSR.

Wrote the paper: WSD PL TMS MV HSR.

References

1. Davis MM, Calame K, Early PW, Livant DL, Joho R, Weissman IL, et al. An immunoglobulin heavy-chain gene is formed by at least two recombinational events. Nature. 1980; 283(5749):733–9. Epub 1980/02/21. PMID: 6766532.
2. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein J, et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. Am J Hum Genet. 2013; 92(4):530–46. Epub 2013/04/02. PMID: 23541343; PubMed Central PMCID: PMC3617388. doi: 10.1016/j.ajhg.2013.03.004
3. Murphy K, Travers P, Walport M. Janeway's Immunobiology. 7th ed. New York, NY: Garland Science; 2008.
4. Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y, Honjo T. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell. 2000; 102(5):553–63. Epub 2000/09/28. PMID: 11007474.
5. Jacob J, Kelsoe G, Rajewsky K, Weiss U. Intraclonal generation of antibody mutants in germinal centres. Nature. 1991; 354(6352):389–92. Epub 1991/12/05. doi: 10.1038/354389a0 PMID: 1956400.
6. Pham P, Bransteitter R, Petruska J, Goodman MF. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. Nature. 2003; 424(6944):103–7. Epub 2003/06/24. doi: 10.1038/nature01760 PMID: 12819663.
7. Calis JJ, Rosenberg BR. Characterizing immune repertoires by high throughput sequencing: strategies and applications. Trends Immunol. 2014; 35(12):581–90. PMID: 25306219. doi: 10.1016/j.it.2014.09.004
8. Six A, Mariotti-Ferrandiz ME, Chaara W, Magadan S, Pham HP, Lefranc MP, et al. The past, present, and future of immune repertoire biology—the rise of next-generation repertoire analysis. Front Immunol. 2013; 4:413. PMID: 24348479; PubMed Central PMCID: PMC3841818. doi: 10.3389/fimmu.2013.00413
9. Giudicelli V, Duroux P, Ginestoux C, Folch G, Jabado-Michaloud J, Chaume D, et al. IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. Nucleic Acids Res. 2006; 34(Database issue):D781–4. doi: 10.1093/nar/gkj088 PMID: 16381979; PubMed Central PMCID: PMC1347451.
10. Arnaout R, Lee W, Cahill P, Honan T, Sparrow T, Weiand M, et al. High-resolution description of antibody heavy-chain repertoires in humans. PLoS One. 2011; 6(8):e22365. Epub 2011/08/11. PMID: 21829618; PubMed Central PMCID: PMC3150326. doi: 10.1371/journal.pone.0022365
11. Boyd SD, Gaeta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J Immunol. 2010; 184(12):6986–92. Epub 2010/05/25. PMID: 20495067. doi: 10.4049/jimmunol.1000445
12. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, Sahaf B, et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci Transl Med. 2009; 1(12):12ra23. Epub 2010/02/18. PMID: 20161664; PubMed Central PMCID: PMC2819115.
13. Briney BS, Willis JR, McKinney BA, Crowe JE Jr. High-throughput antibody sequencing reveals genetic evidence of global regulation of the naive and memory repertoires that extends across individuals. Genes Immun. 2012; 13(6):469–73. Epub 2012/05/25. PMID: 22622198. doi: 10.1038/gene.2012.20
14. DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat Biotechnol.

2013; 31(2):166–9. PMID: 23334449; PubMed Central PMCID: PMCPMC3910347. doi: 10.1038/nbt.2492

15. DeKosky BJ, Kojima T, Rodin A, Charab W, Ippolito GC, Ellington AD, et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. Nat Med. 2015; 21(1):86–91. PMID: 25501908. doi: 10.1038/nm.3743

16. Galson JD, Truck J, Clutterbuck EA, Fowler A, Cerundolo V, Pollard AJ, et al. B-cell repertoire dynamics after sequential hepatitis B vaccination and evidence for cross-reactive B-cell activation. Genome Med. 2016; 8(1):68. PMID: 27312086; PubMed Central PMCID: PMCPMC4910312. doi: 10.1186/s13073-016-0322-z

17. Jackson KJ, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K, et al. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. Cell Host Microbe. 2014; 16(1):105–14. PMID: 24981332; PubMed Central PMCID: PMCPMC4158033. doi: 10.1016/j.chom.2014.05.013

18. Larimore K, McCormick MW, Robins HS, Greenberg PD. Shaping of human germline IgH repertoires revealed by deep sequencing. J Immunol. 2012; 189(6):3221–30. PMID: 22865917. doi: 10.4049/jimmunol.1201303

19. Laserson U, Vigneault F, Gadala-Maria D, Yaari G, Uduman M, Vander Heiden JA, et al. High-resolution antibody dynamics of vaccine-induced immune responses. Proc Natl Acad Sci USA. 2014; 111(13):4928–33. PMID: 24639495; PubMed Central PMCID: PMC3977259. doi: 10.1073/pnas.1323862111

20. Prabakaran P, Chen W, Singarayan MG, Stewart CC, Streaker E, Feng Y, et al. Expressed antibody repertoires in human cord blood cells: 454 sequencing and IMGT/HighV-QUEST analysis of germline gene usage, junctional diversity, and somatic mutations. Immunogenetics. 2012; 64(5):337–50. PMID: 22200891. doi: 10.1007/s00251-011-0595-8

21. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. Proc Natl Acad Sci USA. 2013; 110(33):13463–8. PMID: 23898164; PubMed Central PMCID: PMC3746854. doi: 10.1073/pnas.1312146110

22. Strauli NB, Hernandez RD. Statistical inference of a convergent antibody repertoire response to influenza vaccine. Genome Med. 2016; 8(1):60. PMID: 27255379; PubMed Central PMCID: PMCPMC4891843. doi: 10.1186/s13073-016-0314-z

23. Wu D, Emerson RO, Sherwood A, Loh ML, Angiolillo A, Howie B, et al. Detection of minimal residual disease in B lymphoblastic leukemia by high-throughput sequencing of IGH. Clin Cancer Res. 2014; 20(17):4540–8. PMID: 24970842. doi: 10.1158/1078-0432.CCR-13-3231

24. Cortina-Ceballos B, Godoy-Lozano EE, Tellez-Sosa J, Ovilla-Munoz M, Samano-Sanchez H, Aguilar-Salgado A, et al. Longitudinal analysis of the peripheral B cell repertoire reveals unique effects of immunization with a new influenza virus strain. Genome Med. 2015; 7:124. PMID: 26608341; PubMed Central PMCID: PMCPMC4658769. doi: 10.1186/s13073-015-0239-y

25. Robins H. Immunosequencing: applications of immune repertoire deep sequencing. Curr Opin Immunol. 2013; 25(5):646–52. Epub 2013/10/22. doi: 10.1016/j.coi.2013.09.017 PMID: 24140071.

26. Elhanati Y, Murugan A, Callan CG Jr., Mora T, Walczak AM. Quantifying selection in immune receptor repertoires. Proc Natl Acad Sci USA. 2014; 111(27):9875–80. Epub 2014/06/20. PMID: 24941953; PubMed Central PMCID: PMC4103359. doi: 10.1073/pnas.1409572111

27. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, et al. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood. 2009; 114(19):4099–107. PMID: 19706884; PubMed Central PMCID: PMC2774550. doi: 10.1182/blood-2009-04-217604

28. Robins HS, Srivastava SK, Campregher PV, Turtle CJ, Andriesen J, Riddell SR, et al. Overlap and effective size of the human CD8+ T cell receptor repertoire. Sci Transl Med. 2010; 2(47):47ra64. PMID: 20811043; PubMed Central PMCID: PMC3212437. doi: 10.1126/scitranslmed.3001442

29. Sherwood AM, Desmarais C, Livingston RJ, Andriesen J, Haussler M, Carlson CS, et al. Deep sequencing of the human TCRgamma and TCRbeta repertoires suggests that TCRbeta rearranges after alphabeta and gammadelta T cell commitment. Sci Transl Med. 2011; 3(90):90ra61. Epub 2011/07/08. PMID: 21734177; PubMed Central PMCID: PMC4179204. doi: 10.1126/scitranslmed.3002536

30. Wu D, Sherwood A, Fromm JR, Winter SS, Dunsmore KP, Loh ML, et al. High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. Sci Transl Med. 2012; 4(134):134ra63. PMID: 22593176. doi: 10.1126/scitranslmed.3003656

31. Doria-Rose NA, Schramm CA, Gorman J, Moore PL, Bhiman JN, DeKosky BJ, et al. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. Nature. 2014; 509(7498):55–62. Epub 2014/03/05. PMID: 24590074. doi: 10.1038/nature13036

32. Muraro PA, Robins H, Malhotra S, Howell M, Phippard D, Desmarais C, et al. T cell repertoire following autologous stem cell transplantation for multiple sclerosis. J Clin Invest. 2014; 124(3):1168–72. PMID: 24531550; PubMed Central PMCID: PMCPMC3934160. doi: 10.1172/JCI71691

33. Schneider-Hohendorf T, Mohan H, Bien CG, Breuer J, Becker A, Gorlich D, et al. CD8(+) T-cell pathogenicity in Rasmussen encephalitis elucidated by large-scale T-cell receptor sequencing. Nat Commun. 2016; 7:11153. PMID: 27040081; PubMed Central PMCID: PMCPMC4822013. doi: 10.1038/ncomms11153

34. DeWitt WS, Emerson RO, Lindau P, Vignali M, Snyder TM, Desmarais C, et al. Dynamics of the cytotoxic T cell response to a model of acute viral infection. J Virol. 2015. doi: 10.1128/JVI.03474-14 PMID: 25653453.

35. Morris H, DeWolf S, Robins H, Sprangers B, LoCascio SA, Shonts BA, et al. Tracking donor-reactive T cells: Evidence for clonal deletion in tolerant kidney transplant patients. Sci Transl Med. 2015; 7 (272):272ra10. doi: 10.1126/scitranslmed.3010760 PMID: 25632034; PubMed Central PMCID: PMCPMC4360892.

36. Emerson RO, Mathew JM, Konieczna IM, Robins HS, Leventhal JR. Defining the alloreactive T cell repertoire using high-throughput sequencing of mixed lymphocyte reaction culture. PLoS One. 2014; 9 (11):e111943. PMID: 25365040; PubMed Central PMCID: PMCPMC4218856. doi: 10.1371/journal.pone.0111943

37. Emerson RO, Sherwood AM, Rieder MJ, Guenthoer J, Williamson DW, Carlson CS, et al. High-throughput sequencing of T-cell receptors reveals a homogeneous repertoire of tumour-infiltrating lymphocytes in ovarian cancer. J Pathol. 2013; 231(4):433–40. PMID: 24027095. doi: 10.1002/path.4260

38. Tumeh PC, Harview CL, Yearley JH, Shintaku IP, Taylor EJ, Robert L, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. Nature. 2014; 515(7528):568–71. PMID: 25428505; PubMed Central PMCID: PMCPMC4246418. doi: 10.1038/nature13954

39. Hsu MS, Sedighim S, Wang T, Antonios JP, Everson RG, Tucker AM, et al. TCR Sequencing Can Identify and Track Glioma-Infiltrating T Cells after DC Vaccination. Cancer Immunol Res. 2016; 4(5):412–8. PMID: 26968205; PubMed Central PMCID: PMCPMC4873445. doi: 10.1158/2326-6066.CIR-15-0240

40. Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung MW, Parsons JM, et al. Using synthetic templates to design an unbiased multiplex PCR assay. Nat Commun. 2013; 4:2680. PMID: 24157944. doi: 10.1038/ncomms3680

41. Chothia C, Gelfand I, Kister A. Structural determinants in the sequences of immunoglobulin variable domain. J Mol Biol. 1998; 278(2):457–79. doi: 10.1006/jmbi.1998.1653 PMID: 9571064.

42. Fisher RA, Corbet AS, Williams CB. The relation between the number of species and the number of individuals in a random sample of an animal population. J Anim Ecol. 1943; 12:42–58.

43. Sanathanan L. Estimating the size of a truncated sample. J Am Statist Assoc. 1977; 72(356):669–72.

44. Rodrigues J, Milan LA, Leite JG. Hierarchical bayesian estimation for the number of species. Biom J. 2001; 43(6):737–46.

45. Morbach H, Eichhorn EM, Liese JG, Girschick HJ. Reference values for B cell subpopulations from infancy to adulthood. Clin Exp Immunol. 2010; 162(2):271–9. Epub 2010/09/22. PMID: 20854328; PubMed Central PMCID: PMC2996594. doi: 10.1111/j.1365-2249.2010.04206.x

46. Agematsu K, Nagumo H, Yang FC, Nakazawa T, Fukushima K, Ito S, et al. B cell subpopulations separated by CD27 and crucial collaboration of CD27+ B cells and helper T cells in immunoglobulin production. Eur J Immunol. 1997; 27(8):2073–9. Epub 1997/08/01. doi: 10.1002/eji.1830270835 PMID: 9295047.

47. van Zelm MC, Szczepanski T, van der Burg M, van Dongen JJ. Replication history of B lymphocytes reveals homeostatic proliferation and extensive antigen-induced B cell expansion. J Exp Med. 2007; 204(3):645–55. Epub 2007/02/22. doi: 10.1084/jem.20060964 PMID: 17312005; PubMed Central PMCID: PMC2137914.

48. Matsuda F, Ishii K, Bourvagnet P, Kuma K, Hayashida H, Miyata T, et al. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. J Exp Med. 1998; 188 (11):2151–62. PMID: 9841928; PubMed Central PMCID: PMC2212390.

49. Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. Blood. 2010; 116(7):1070–8. Epub 2010/05/12. PMID: 20457872; PubMed Central PMCID: PMC2938129. doi: 10.1182/blood-2010-03-275859

50. Wu YC, Kipling D, Dunn-Walters DK. The relationship between CD27 negative and positive B cell populations in human peripheral blood. Front Immunol. 2011; 2:81. Epub 2011/01/01. PMID: 22566870; PubMed Central PMCID: PMC3341955. doi: 10.3389/fimmu.2011.00081

51. Glanville J, Kuo TC, von Budingen HC, Guey L, Berka J, Sundar PD, et al. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. Proc Natl Acad Sci U S A. 2011; 108(50):20066–71. PMID: 22123975; PubMed Central PMCID: PMC3250199. doi: 10.1073/pnas.1107498108

52. Xu JL, Davis MM. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. Immunity. 2000; 13(1):37–45. Epub 2000/08/10. PMID: 10933393.

53. Rock EP, Sibbald PR, Davis MM, Chien YH. CDR3 length in antigen-specific immune receptors. J Exp Med. 1994; 179(1):323–8. PMID: 8270877; PubMed Central PMCID: PMC2191339.

54. Mroczek ES, Ippolito GC, Rogosch T, Hoi KH, Hwangpo TA, Brand MG, et al. Differences in the composition of the human antibody repertoire by B cell subsets in the blood. Front Immunol. 2014; 5:96. PMID: 24678310; PubMed Central PMCID: PMC3958703. doi: 10.3389/fimmu.2014.00096

55. Herzenberg LA, Black SJ, Tokuhisa T, Herzenberg LA. Memory B cells at successive stages of differentiation. Affinity maturation and the role of IgD receptors. J Exp Med. 1980; 151(5):1071–87. Epub 1980/05/01. PMID: 6966317; PubMed Central PMCID: PMC2185844.

56. Weiss U, Rajewsky K. The repertoire of somatic antibody mutants accumulating in the memory compartment after primary immunization is restricted through affinity maturation and mirrors that expressed in the secondary response. J Exp Med. 1990; 172(6):1681–9. Epub 1990/12/01. PMID: 2124253; PubMed Central PMCID: PMC2188767.

57. Kocks C, Rajewsky K. Stepwise intraclonal maturation of antibody affinity through somatic hypermutation. Proc Natl Acad Sci USA. 1988; 85(21):8206–10. Epub 1988/11/01. PMID: 3263647; PubMed Central PMCID: PMC282396.

58. Zhang Y, Meyer-Hermann M, George LA, Figge MT, Khan M, Goodall M, et al. Germinal center B cells govern their own fate via antibody feedback. J Exp Med. 2013; 210(3):457–64. Epub 2013/02/20. PMID: 23420879; PubMed Central PMCID: PMC3600904. doi: 10.1084/jem.20120150

59. Peron S, Laffleur B, Denis-Lagache N, Cook-Moreau J, Tinguely A, Delpy L, et al. AID-driven deletion causes immunoglobulin heavy chain locus suicide recombination in B cells. Science. 2012; 336 (6083):931–4. Epub 2012/04/28. PMID: 22539552. doi: 10.1126/science.1218692

60. Victora GD, Nussenzweig MC. Germinal centers. Annu Rev Immunol. 2012; 30:429–57. Epub 2012/01/10. PMID: 22224772. doi: 10.1146/annurev-immunol-020711-075032

61. Bransteitter R, Pham P, Calabrese P, Goodman MF. Biochemical analysis of hypermutational targeting by wild type and mutant activation-induced cytidine deaminase. J Biol Chem. 2004; 279(49):51612–21. Epub 2004/09/17. doi: 10.1074/jbc.M408135200 PMID: 15371439.

62. Greene J, Birtwistle MR, Ignatowicz L, Rempala GA. Bayesian multivariate Poisson abundance models for T-cell receptor data. J Theor Biol. 2013; 326:1–10. PMID: 23467198; PubMed Central PMCID: PMC3972257. doi: 10.1016/j.jtbi.2013.02.009

63. Rempala GA, Seweryn M, Ignatowicz L. Model for comparative analysis of antigen receptor repertoires. J Theor Biol. 2011; 269(1):1–15. PMID: 20955715; PubMed Central PMCID: PMC3006491. doi: 10.1016/j.jtbi.2010.10.001