Tools for the analysis of B-cell clonal diversity in immune repertoires

A Thesis

Submitted to the Faculty

of

Drexel University

By

Bochao Zhang

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

Jan 2018

Table of Contents

Abstract

The lymphocytes of an organism harbor a diverse collection or repertoire of antigen receptors (antibodies (Ab) on B cells and T cell receptors (TCR) on T cells). This diversity of the adaptive immune repertoire is important for effective immune defense. In my thesis, I have developed computational methods to study the diversity and landscape of the immune repertoire using data from next-generation sequencing experiments and have applied these tools to the study human B cells in the blood and in different tissues. The first tool I developed was an efficient and accurate means of identifying the nearest germline-encoded variable (V) region gene in Ab and TCR rearrangements using anchoring sequences. I used this method to demonstrate that the length of the V gene sequence and its level of somatic mutation influenced the reliability of V gene assignment. Only with adequate V gene assignment can closely related sequences be reliably grouped together into clones (i.e. sequences from lymphocytes that likely derive from a common progenitor cell). With this method, I contributed to a computational pipeline that can process millions of antibody or TCR sequences in hours to days, unlike earlier methods. I used this method to identify and track expanded B-cell clones through the human body, creating an atlas of clonal connections between the blood, bone marrow, spleen, lung, mesenteric lymph node, jejunum, ileum and colon. To power the analysis to study clonal overlap between the tissues, I performed rarefaction analysis on biological replicates to determine how many replicates and how large the clones needed to be to have confidence in clones being present or absent in the different tissue sites. To quantify the level of overlap between two-tissue sites, I used the cosine statistic and showed that my analysis was robust to different measures of clones and clone size. To visualize the tissue distribution of large clones, I created line-circle plots, in which clones are displayed as lines with circles corresponding to the tissues. The sizes of the tissue circles were proportional to clone copy numbers and the colored wedges within the circles indicated the fraction of sequencing libraries that contained sequences from the clone. My analysis revealed the clones tended to partition into two large networks in the human body, one in blood-rich tissues such as the bone marrow, spleen and lung, and another network that was more separated from the blood, in the gastrointestinal tract. I also used methods to visualize and characterize the diversification within B-cell clones due to somatic hypermutation, including lineage tree analysis, the analysis of four-fold silent mutations (mutations that do not change the amino acid sequence), methods for studying intermingling of lineage tree branches (clumpiness) and the use of Bayesian modeling approaches. I used these methods to show that GI tract clones in the human body harbor higher levels of somatic hypermutation and that there is extensive sharing of sequence variants within individual clonal lineages between different sites within the GI tract. Finally, I studied selection of mutations within clonal lineages over time in patients with an autoimmune disease (Sjogren's syndrome) and showed that large clones that were resistant to B-cell depletion therapy were under negative selection. Together, my analysis tools provide a useful means to systematically and quantitatively characterize diversity at the population (repertoire) level and at the clonal level. These tools can be applied to future immune repertoire profiling to study immune responses to vaccines, cancer and infectious disease.

Keywords: B cell receptor, antibody, immune repertoire, next-generation sequencing

Introduction

Immune diversity

The collection of B or T cells (lymphocytes) in an organism is referred to as the immune repertoire. The B-cell repertoire in an adult human is estimated to exceed 100 billion different specificities [1] and similar diversity has been described for T cells [2]. This diversity of the repertoire is important to immune functions [3-5]. It allows the immune system to respond fairly quickly to almost any foreign antigen [6]. Quantifying repertoire diversity is useful for evaluating immune competence in health and disease. For example, limited repertoire diversity has been associated with frailty and old age [7] and oligoclonal T cell responses are associated with less effective anti-tumor immunity [8]. The expansion of cells that derive from a common precursor (e.g. clones) is important in malignant conditions such as lymphoma and leukemia [9, 10] as well as in pre-malignant conditions such as MGUS and MBL [11, 12] or in autoimmunity, where large clones may be pathogenic. Finally, the diversity of antigen-specific cells is useful for monitoring immunity to vaccines [13]. Hence, we wish to know if the immune system is functioning well by measuring the diversity of the immune repertoire.

Diversity of the B-cell immune repertoire is generated in two rounds of diversification and selection [3, 14, 15]. The first round occurs in the bone marrow, where the variable regions of antibodies are assembled through a process of V(D)J recombination that is carried out by recombinase activating genes (RAG1 and RAG2, [16]). First the diversity (D) and joining (J) gene segments are combined and then the upstream variable (V) region gene is rearranged to DJ. In addition to the combinatorial diversity of different V genes (45-50 in humans), D genes (27) and J gene (6) [17], there is variation at the junctions between the recombined gene segments. This junctional modification occurs by various mechanisms including addition of nucleotides by terminal deoxynucleotidyl transferase (TdT) [18], palindromic nucleotide additions (mediated by

RAG) and exonucleolytic nibbling (mediated by non-homologous end-joining machinery) [19]. Then, if the heavy chain is functional, the light chain is rearranged. Light chains come in two flavors, kappa and lambda. If kappa rearrangement fails to produce a useful or non-autoreactive antibody, the locus is deleted and rearrangement of lambda occurs (reviewed in [20]). Light chain rearrangement is also mediated by RAG but only involves V genes and J genes, no D gene. Also, TdT expression is much lower during light chain rearrangement, so there is less junctional diversity. Further combinatorial diversity of the antibody (and TCR) repertoire is achieved by the pairing of different heavy chains to light chains to form the receptor. The final product, in the case of the B cell, is an IgM antibody molecule, which is a tetramer, consisting of two identical heavy chains and two identical light chains. Only one kind of antibody or TCR is expressed on most lymphocytes (allelic exclusion), allowing selection of a lymphocyte to be coupled to its specific receptor.

The second round of diversification only occurs in B cells and consists of DNA point hypermutation (known as somatic hypermutation, SHM) of the antibody variable region during an immune response. SHM is carried out by the enzyme AID (activation-induced cytidine deaminase) [21] and, when coupled with selection for improved binding to antigen, results, over time, in affinity maturation (or improved antibody binding). SHM leads to intra-clonal diversification and can be used to study the dynamics of an immune response. Over time mutations accumulate, so the level of mutation can be used as a surrogate marker for how long an immune response has been going on. Furthermore, the nature of the mutation (whether it results in a change of the amino acid sequence) provides insight into selection of the clone. In general, B cells that are under positive selection have mutations in the complementarity determining regions (CDRs), or the parts of the antibody that are important for binding to the antigen, whereas they tend to have silent mutations in the framework regions (the conserved regions of the antibody that are important for maintaining its structure), reviewed in [22].

Repertoire diversity is thus composed of different expanded clones (inter-clonal diversity) and (for antibodies only) of sequence variants within individual clones (intra-clonal diversity). Therefore, to characterize B-cell repertoire receptor diversity we need to understand the diversity between and within B cell clones. Because the immune repertoire is so large, we need to collect massive data sets to be able to do this. Fortunately, with the advent of next-generation sequencing (NGS), we are able to collect data on immune repertoires ranging from hundreds of thousands to millions of cells [23]. However, even with this technology, we still are usually undersamping the true diversity of the immune system. Moreover, since we cannot usually track the development of clones over time, we must infer the genes the clone started out with from cross-sectional measurements of the repertoire. In this cross-sectional view of the repertoire in the blood, for example, some B-cell clones are newly formed (e.g. transitional B cells that are new bone marrow emigres) and have antibody V genes that are identical to the germline-encoded sequence, while other B cells are more mature (harboring somatic hypermutations, due to antigen exposure and selection). To complicate matters, NGS has an error rate of ~1% [24], so some of the sequences may appear to be different from germline due to sequencing error, or less commonly PCR error, rather than bona fide SHM.

Identifying Clones

The unit of immune selection is the clone, or a collection of cells with highly similar antigen receptors (Ab or TCR) that derive from a common progenitor cell. To study clones and, by extension, immune repertoires, we need to be able to reliably group their Ab or TCR sequences into clones. To do this, we identify the germline source of each sequence and divide them into clones based on V and J gene identity and similar CDR3 sequences. The CDR3, or third complementarity determining region, is the portion of the antibody (or TCR) heavy chain sequence where the V, D and J genes come together. Thus the CDR3 is the most diverse part of the antibody structure and is the most reliable part of the sequence to serve as a clonal "fingerprint". For the analysis of bulk populations of cells, we focus on the heavy chain because it has the most diverse CDR3 and, in the case of antibodies, it tends to harbor higher levels of SHM than the light chain. If we have data from single cells, we can use both the heavy chain and light chain sequences (paired) for clonal identification. After we have divided the sequences into clones, based upon their antibody sequences, we can determine sufficiency of sampling for clone detection and, we can analyze the diversity between clones and within clones.

V gene assignment

First, we need to identify the germline source of each sequence to group them into clones. The sequence data we get lack the information of the original germline gene. Thus, we need to identify the sequence and associate it with the closest germline gene in terms of the number of mutations [25, 26]. We assume that the germline gene with the fewest mutations compared to the sequence will be our best guess as to what the actual gene is. However, some germline genes are very similar to each other, especially if they derive from the same gene family [17], so our second best guess for the closest corresponding germline V gene may be only slightly worse or even indistinguishable from our first guess ("V-ties"). This issue is worsened if our sequences are more mutated and/or shorter in length. Thus, we need a method that can distinguish between the true mutation from the germline and quantify the differences.

VH replacement and footprints

During receptor editing, heavy chain or light chain V regions of self-reactive receptors can be modified by further gene rearrangement (reviewed in [20]). In the case of the antibody heavy chain, an upstream VH gene can invade into a pre-formed V(D)J rearrangement on the same chromosome via a cryptic heptamer that resides near the 3' end of ~90% of human VH genes. This type of editing rearrangement is called VH replacement. During VH replacement, sometimes the rearrangement can leave a “footprint” of the old V gene at the 3’ end (downstream of the cryptic heptamer), resulting in elongation of the framework region 3 (FWR3) sequence of the newly formed V region [27]. Since VH replacement cannot be directly observed in a normal immune repertoire, footprints are an indirect indicator of VH replacement. However, given the similarity of the FWR3 region among germline VH genes, it is not always possible to determine which VH gene was there initially, or if the altered sequence is even due to VH replacement at all.

B cell selection

As described above, the diversity of B cells is first established in the bone marrow by combining randomly selected V, D and J genes, pairing different heavy and light chains and through diversification at the junctions between the recombining gene segments. In the periphery (secondary lymphoid organs), only those clones having antibodies that recognize antigens that are stimulatory to the immune system are selected to proliferate. The selected clones are subject to an affinity maturation process, in which the cells go through somatic hypermutation and those with improved affinity proliferate. In both stages of antibody diversification, B cells expressing autoreactive receptors are also negatively selected [28]. In general, if we consider changes in repertoire structure from the perspective of clonal selection [29], the competition of clones can be divided into clonal shift (the competition between clones) and clonal drift (the competition among cells within a clone) [30].

Clonal shift

To study the clonal shift, our main limitation is the requirement of large coverage of repertoire. Only with decent coverage can we confidently quantify the diversity (or even the presence of absence) of clones. Otherwise, the lack of diversity could be the result of under-sampling. To deal with this issue, we require multiple replicate sampling and a method to measure the coverage. Rarefaction analysis has long been used in ecology and has proven to be a robust way to estimate diversity of a community [31], and thus is a good method to adapt to our immune diversity analysis. When quantifying clonal overlap, we need to take into account both the number and the sizes of the clones, and visualize their overlap within and between samples and anatomic sites. Further complicating the analysis, there are multiple ways to define the size of clone. We need to deploy clone-size thresholding that affords some correction for differences in sequencing depth (which can influence unique sequence numbers), while not relying exclusively on resampling to establish clone size cut-offs.

Clonal drift

To study clonal drift, we need to differentiate between true diversification due to selection and random noise due to sequencing error of HTS [32]. Most of the sequencing errors can be removed by re-sequencing and filtering out singletons. The ability to detect selection from mutated sequences is a critical part of many studies. A major methodology to do so is to compare the levels of synonymous (no change in amino acid) and non-synonymous (results in amino acid change) mutations [33-36]. In neutral selection, the non-synonymous to synonymous ratio is about 3. Elevated levels indicate positive selection, while decreased levels indicate negative selection [37, 38]. It is expected that complementary determining regions (CDR) undergo positive selection while framework regions (FWR) undergo negative selection, as the CDR is where the receptor interacts with antigen and the FWR is important for maintaining the structure. Selection can be estimated using a binomial test [22]. To cancel the impact of selection on our measurement of mutation frequency (which can be used to study the number of cell divisions while the AID enzyme is engaged), we can use synonymous mutations with four-fold degenerate. Such four-fold silent mutations tolerate any point mutation at the third position of codons, and are least affected by selection. Thus, they are good indicators of baseline mutation burden. Lastly, we can use clumpiness [39] as a measure of the degree of intermingling of sequence variants among different tissues within clonal lineages.

Overview of Experimental Questions in the Thesis

The overarching theme of my thesis work was to analyze the clonal diversity in human B-cell repertoires. In general, if we consider changes in repertoire structure from the perspective of clonal selection [29], the competition of clones can be divided into clonal shift and clonal drift [30]. Unfortunately, both processes are difficult to identify. The large amount of data generated from NGS also calls for computationally efficient approaches for clone identification and characterization. Thus, my findings regarding immune repertoires can be divided into questions of clonal identification and the measurement of inter and intra clonal diversity.

In my first aim I developed methods for accurately and rapidly identifying the nearest corresponding germline V gene or genes to rearranged V region sequences. Currently existing methods have failed to adequately address cases in which multiple germline genes are equally possible [24]. In my first aim, I developed methods to categorize germline genes that cannot be differentiated under certain levels of somatic mutation and sequencing length. I also estimated the sampling level of the repertoire by using rarefaction analysis. By calculating the diversity of clones in different samples we can estimate the number of additional samples needed to achieve a certain level of coverage. These statistical tools were incorporated into ImmuneDB, which is a database for the storage, analysis and visualization of immune receptor repertoire data [40]. In addition to developing a method for germline V gene assignment and testing the adequacy of that assignment under different conditions of sequence length and SHM [41], I analyzed repertoire diversity in the FWR3. For the FWR3 analysis, I focused on studying diversity due to VH replacement and tested the hypothesis that longer FWR3 sequences harbor VH replacement “footprints”. I observed that these footprint-like sequences could not be differentiated from random FWR3 and CDR3 sequence diversification [42]. Additionally, I found that the number of footprints were positively correlated to CDR3 length in both in-frame and out-of-frame rearrangements and that they fit well with a Poisson distribution. These findings suggested that VHR footprints are not good indicator of VH replacement [42].

In my second aim, I characterized clonal shift by comparing the incidence and overlap of clones from different samples of a given tissue or across different tissues or anatomic sites. I quantified the clonal shift of the B-cell repertoire in different tissues by quantifying the level of clonal overlap within and across these tissues. The requirement for multiple replicate sampling was the main experimental constraint of this research. I was fortunate to collaborate with experimenters from University of Pennsylvania and Columbia University to overcome this limitation. They provided a massive sequence data set from eight different anatomic compartments in six different human organ donors [43]. I used rarefaction analysis to ensure good coverage and calculated cosine similarity within and between tissues. The cosine similarity analysis gave more weight to clones with larger size. I showed that large B-cell clones partitioned into two broad networks—one network spanning the blood, bone marrow, spleen and lung, while the other was restricted to tissues within the gastrointestinal (GI) tract (jejunum, ileum and colon).

In my third aim, I studied clonal drift of B cell clones in different human tissues. In B cells, we have a unique opportunity to study clonal drift since as they can undergo somatic mutation of their rearranged antibody genes. In my third aim, using the aforementioned dataset, I compared unique and four-fold synonymous mutations of clones within tissues, and found that GI tract clones displayed extensive sharing of sequence variants among different portions of the tract and had higher frequencies of somatic hypermutation, suggesting extensive and serial rounds of clonal expansion and selection in different tissues. Clumpiness analysis showed shared variants of clonal tree leaves in gut tissues, suggesting localized proliferation after SHM. Additionally, by studying big clones from a Sjögren’s syndrome patient across six time points and applying selection pressure analysis using baseline [44], I showed that the clone was under negative selection.

Articles

Article 1. Discrimination of germline V genes at different sequencing lengths and mutational burdens: A new tool for identifying and evaluating the reliability of V gene assignment

Published in final edited form as:
J Immunol Methods. 2015 December ; 427: 105–116. doi:10.1016/j.jim.2015.10.009.

Discrimination of germline V genes at different sequencing lengths and mutational burdens: a new tool for identifying and evaluating the reliability of V gene assignment

Bochao Zhang[1], **Wenzhao Meng**[2], **Eline T. Luning Prak**[2], and **Uri Hershberg**[1,3]

[1]School of Biomedical Engineering, Science and Health Systems, 711 Bossone Building, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA

[2]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, 405B Stellar Chance Labs, 422 Curie Boulevard, Philadelphia, PA 19104, USA

[3]Department of Microbiology and Immunology, College of Medicine, 2900 Queen Lane, Philadelphia, PA 19129, USA

Abstract

Immune repertoires are collections of lymphocytes that express diverse antigen receptor gene rearrangements consisting of Variable (V), (Diversity (D) in the case of heavy chains) and Joining (J) gene segments. Clonally related cells typically share the same germline gene segments and have highly similar junctional sequences within their third complementarity determining regions. Identifying clonal relatedness of sequences is a key step in the analysis of immune repertoires. The V gene is the most important for clone identification because it has the longest sequence and the greatest number of sequence variants. However, accurate identification of a clone's germline V gene source is challenging because there is a high degree of similarity between different germline V genes. This difficulty is compounded in antibodies, which can undergo somatic hypermutation. Furthermore, high-throughput sequencing experiments often generate partial sequences and have significant error rates. To address these issues, we describe a novel method to estimate which germline V genes (or alleles) cannot be discriminated under different conditions (read lengths, sequencing errors or somatic hypermutation frequencies). Starting with any set of germline V genes, this method measures their similarity using different sequencing lengths and calculates their likelihood of unambiguous assignment under different levels of mutation. Hence, one can identify, under different experimental and biological conditions, the germline V genes (or alleles) that cannot be uniquely identified and bundle them together into groups of specific V genes with highly similar sequences.

Corresponding author: Uri Hershberg. Address: School of Biomedical Engineering, Science and Health Systems, 711 Bossone Building, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA. uri.hershberg@drexel.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

high throughput sequencing; gene identification

1. Introduction

The diversity of the immune T cell receptor (TCR) and B cell receptor (BCR or antibody, Ig) repertoires allows T cells and B cells to respond to a wide variety of pathogens and establish protective immunity. Repertoire diversity is generated in a combinatorial fashion. Each antigen receptor is a tetramer made up of two heterodimers; each heterodimer consists of a heavy and a light chain. The variable portions of these chains (V regions) arise by recombination of individual members of variable (V), diversity (D only in the case of heavy chains), and joining (J) gene segments [1, 2]. V(D)J recombination of individual coding elements from the V, D, and J genes and junctional modifications result in considerable combinatorial diversity [3]. Lymphocytes are subjected to additional rounds of selection during the immune response, and B cells can undergo further diversification via somatic hypermutation of their antibodies [4].

Quantifying repertoire diversity is important in studies of inflammation [5], reaction to disease [6], vaccination, autoimmunity [7] and cancer [8]. By finding the dominant clones in a given repertoire or by studying the distribution of V gene usage, researchers can gain a better understanding of how the human immune system responds to a particular antigen or is perturbed by or during disease. Our ability to study the repertoire is greatly enhanced by the advent of high-throughput sequencing technologies [9]. However the resulting deluge of data has its own issues. Sequences are often partial and the error rates are significant [10]. Moreover, the sheer amount of data means that there is very little possibility to do quality control and analysis without the aid of computational means.

The clone (also referred to as clonotype) is the unit of selection of the immune response. Clones are collections of sequences that are associated with B cells that derive from a common precursor cell. To properly understand repertoire diversity, we first need to separate the sequences into clones [11]. Unique sequence variants are insufficient for this purpose because they often represent sequencing errors and even sequences with larger numbers of nucleotide differences than predicted by sequencing error may be clonally related. To assign sequence membership into clones, the V, (D) and J genes within each rearrangement need to be associated with their corresponding germline (unarranged) gene segments. Currently, there are several programs that have been developed to perform this function. The two most commonly used are IMGT's (ImMunoGeneTics) High V-Quest, which uses local alignment to find the best match between the sequence and V, D, and J gene segments [12], and IgBLAST, which breaks the sequence into a *k*-letter word list, scans the database for possible matching words, and evaluates the significance [13]. Good performance is also achieved by a three-dimensional dynamic programming algorithm for V(D)J segments called SoDA [14] and by applied statistical models, such as the hidden Markov model (HMM), used by iHMMune-align to obtain the optimized parameters fitting to the rearranged antibody [15]. IMGT's High V-Quest, and iHMMune-align will give multiple identifications when they do not have a conclusive identification. IgBLAST, on the other hand, will always give multiple identifications. (Features of each identification method are provided in Table 1).

Despite their many strengths, these methods do not take into account the *a priori* similarity of some germline V genes. Some V genes are more similar to each other than others. Thus there is some *a priori* non-uniform rate at which certain V genes can be confused with others due to point mutations, deletions or other errors [16]. In fact, some highly similar V genes (such as VH4-30-02 and VH4-30-04) are indistinguishable from each other even if unmutated. This problem is conpounded by two major issues of the high-throughput sequencing technology: (1) it generates short, partial sequences of the Ig genes and (2) it has a significant error rate [10]. Because of this in many high-throughput experiments there is less than the full component of positions with which to differentiate between germline V genes and even more of them are *a priori* indecipherable. None of the above methods takes into account this source of confusion. Even if they do score all possible good hits, they do not first calculate the likelihood that two germline V genes would show similar scores.

Here we therefore present two innovations: (1) a rapid heavy chain alignment method based on highly stable anchoring positions in V genes that are identical across all germline genes [12, 17] and seldom survive when mutated ([18, 19] and see below) and (2) a general framework of assessing confidence in V gene identification. Using the latter, we have calculated the mutation distances between V genes from the compiled IMGT list of the different human and murine germline V genes [12] and determined those V genes whose germline source cannot be discriminated at different V gene lengths and mutation levels.

2. Method and materials and calculation

2.1 Identifying which germline V genes are a priori too similar to discriminate

To identify which germline V genes can and which cannot be distinguished from each other, we use an alignment method to take a first pass at their identification. After the first round of identification, we make an initial estimate of the distribution of V gene lengths and mutation levels. Based on this estimation, we calculate the likelihood of two germline V genes/alleles being confused by chance for the estimated V gene lengths and levels of mutation (see **Calculation**). Those germline genes whose probability of being confused when assigned is above our pre-defined threshold ($P > 0.01$ in the examples shown here) will be identified as giving a mixed V gene identification (so-called "V-ties"). We use a sample-based estimate of mutation/ sequence length as we wish to compare clones and sequences across an entire experiment. This requires the assumption that in a single experiment or sample the mutation pattern/level and V gene length are consistent among sequences. We can then predict the V-ties we expect to find in the experiment while retaining a consistent set of common germline associations that we can use for clonal assignment and clonal diversity analysis throughout the experiment. If some sequences are suspected to be uncharacteristically mutated and thus skewing the estimated level of repertoire error/noise, they can be removed from analysis and V-ties can be reassigned.

It is important to note that the assignment of germline V-ties is a specific one as the potential confusion of germline V genes assigned will always be the same specific *small* subset of all the V genes. The amount of V gene sequence positions we observe changes which germlines will be V-ties. Lack of sequence information can be divided into two types: (1) Partial sequence reads and (2) mutation/sequencing error. Tables showing the list of potentially confused V genes given specific V gene identifications at 100, 150, 200 nucleotide and full sequences length with 0.05, 0.15 and 0.30 mutation frequencies are found in the supplemental materials (Supplemental Table 1–4).

The precision of V gene identification we consider for the sake of clonal identification is usually at the level of the gene or in some cases, if mutation rates are low and we have full sequences, the allele. Using this method, we can consider a set of unique sequences or clones and assign to each the germline gene for which we have adequate confidence. Some V genes can be fully differentiated at the gene level, some at the allele level and some, for a given dataset of specific sequencing quality or level of mutation, can only be assigned at the level of V-ties with one or two other potential germline V genes. The issue we are pinpointing here is one of germline similarity. Some germline genes/ alleles are so similar that when we query their mutant progeny we cannot discriminate between them with adequate confidence as random error may confuse them. Thus re-sequencing (if it does not remove error) will not change the type of V-ties we identify. PCR error (and selection) can skew the distribution such that the more mutated but more "false germline like" sequence will be more prevalent.

A set of Matlab codes for calculating V-ties can be found on-line at: https://github.com/DrexelSystemsImmunologyLab/ConservedIdentification.git. It can take germline aligned and V(D)J gene associated sequences from any identification method and calculate V-ties base on these identifications.

2.2. Description of the conserved anchor method of germline association

In addition to providing an assessment of our confidence in germline VH gene assignment, we describe herein a novel method of VH gene identification and alignment to IMGT numbering. Our method utilizes consistencies of VH gene structure to make alignment much faster without any loss in accuracy. We show here how it applies to human B cell VH genes and show how it can be modified for use on human VL genes and on murine VH and VL genes. Our human VH germline identification method starts with JH gene identification and then continues to anchor the VH gene and align it. First, we find JH genes by exact match of nucleotides. The nucleotides we use for the JH gene are shown in Table 2 and are located at positions 46 to 63 of the JH gene alignment according to IMGT numbering [20]. If no match is found, the nucleotide strings used will be reduced by one codon from the 3' end and new strings are then used to find the match. This process is repeated until we can find a match. However, a minimum of twelve matching nucleotides is required to ensure the accuracy of matching. If still no match is found, then the reverse complement of original sequence is used and the aforementioned steps are repeated to find matches. If no match is found in either the original sequence or its reverse complement, then the sequence is put in a separate file (Figure 1). Second, we pinpoint the position of the human VH gene using the highly conserved amino acid sequence 'DXXXXXC' which starts with an aspartic acid (D) residue at position 98 (by IMGT numbering) and ends with a Cysteine (C) residue at position 104. These positions are highly conserved at both the amino acid and the nucleotide level as the D is encoded by the nucleotides 'GAC' at position 292 to 294 in all but two alleles of functional and non-partial human heavy chain genes while the C is encoded by TGT at positions 310–312 in all but 5 alleles of one gene (Table 3). In these 5 alleles, the V gene ends out of frame with a TG at positions 310–311.

We identify GACNNNNNNNNNNNNNNNTGT in the sequence. If the sequences lack the GAC nucleotide or it's synonymous mutation or if we find multiple 'DXXXXXC' we search for the nucleotides encoding YYC, at amino acid positions 102–104. These too are highly conserved at the nucleotide level and will most commonly take on the form of TATTACTGT (Table 3). If neither of the primary nucleotide motifs encoding DXXXXXC or YYC is found, we will search for other nucleotide combinations that can encode them (Figure 1). If after these steps still no V gene anchor is found, the sequence is put into another separate rejection file for sequences with an identified J and no V.

To test how frequently these positions were mutated, we sent a set of 150,000 sequences (as described in section 2.4) [21] to be aligned using High V-Quest [12]. This analysis resulted in 92,491 unique sequences. We found 'D' mutated synonymously 457 (0.49%) times and non-synonymously 3850 (4.16%) times, while 'C' is found 785 (0.85%) and 2921 (3.16%) times respectively (Table 4). As we would expect from negative selection and random error these ratios of mutation either match or fall below the ratios of 8 to 1 and 7 to 1 expected for D or C (under uniform patterns of mutation or error and remembering that C can mutate to stop). The combination of both anchoring sites 'DXXXXXC' was found mutated 732 (0.54%) times. The other possible source of confusion, in which DXXXXXC occurs more than once in a V gene sequence happened 1293 times. But in all these cases only one of the pair had the second anchor YYC. The first 'Y' at position 102 was found mutated synonymously 807 times and non-synonymously 2934 times, and the second 'Y' at position 103 was found 1921 and 7989 was 'YHC' (IGHV3-20*01) at those positions and 383 whose germline was 'YCC' (IGHV4-34*11) at those positions (Table 4). The combination of 'YYC' was found mutated simultaneously 292 times. The mutation frequencies of D at position 98, Y at position 102 and C at position 104 are the lowest among all positions (Figure 2). It is important to note that the above data are from high-throughput sequencing data of DNA, and include sequencing errors. In similar experiment, studying mutated B cell receptors taken from human lymph nodes, the genes were sequenced from barcoded mRNA, where consensus alignments were used to create the sequences of B cell receptors and most if not all of the sequencing errors were fixed [19]. In this barcoded data set [19] we found the nucleotides encoding C mutated 79 times synonymously and zero times non-synonymously in 3,017 sequences with copy number greater than one.. This leads us to predict that, with proper sequencing error correction, the amount of unique sequences lost by relying upon the Anchor method, should drop below the 2% level we observe here (see **Results section 4.3**).

After the relative position of the sequence and germline are determined by the anchor(s), all the alleles of all V genes in the IMGT database are compared with the sequences. The allele(s) with the fewest mismatches will be assigned as the germline source of the sequence. If multiple germline genes are identified as being equally distant from the query sequence, they will be both labeled as possible germline sources. It is important to note that such a confusion of identity can happen with any two germline genes but is much rarer than the appearance of V-ties and implies either a lack of information or high mutation levels. The method described here does not identify insertion/deletions in sequence. However, we apply an insertion/deletion control after the identification. If there are more than 9 mutations in a sliding window of 15 nucleotides, we label the sequences as having potential insertion/deletion(s) and identify their germline source using local alignment. A set of Matlab codes for the Anchoring method of germline association can be viewed online at: https://github.com/DrexelSystemsImmunologyLab/ConservedIdentification.git.

Similar sequence anchor points are found in human light chains and TCR, and in the murine V genes for BCRs and TCRs (Table 5). In human and murine light chains, the dominant codon encoding ‘D’ at position 98 is ‘GAT’ instead of ‘GAC’ in heavy chains and TCR. However, in murine TCR α chains the ‘D’ is at IMGT position 100 and ‘YYC’ at position 104. In TCR there is no clearly dominant amino acid combiantion at positions 102–104. They have relatively equal codon usage encoding ‘YYC’, ‘YLC’ and ‘YFC’ at these positions. The anchor variations for these chains are found in the supplemental materials (Supplemental Tables 5–12).

2.3. Simulated validation of the consistent mis-identification of V-ties

To validate our method of V gene alignment and germline origin identification, we used simulated mutant sequences. The method for generating these sequences is described in Section 2.4. These simulated sets of mutant sequences were compared to a reference set of all functional and non-partial alleles [12] (see Supplemental Table 13). We compared our identification of the simulated dataset using the anchor method with those from two commonly used V gene identification tools: IMGT’s High V-Quest and NCBI’s IgBLAST. We used the downloadable IgBLAST (version 2.2.28) with the same set of germlines as in our method. High V-Quest uses the entire germline dataset in IMGT of which our set is a subset.

High V-Quest uses a global pairwise alignment without insertions or deletions [12] and outputs multiple V gene identifications without a metric for preference (insertion/deletions are fixed at later stage). IgBLAST makes a *k*-letter word list (*k*=9 for V gene identification by default), scans the database for possible matching words and evaluates the significance [13] (Table 1). In this way they generate a list of possible V gene identifications with their significance. For this reason we consider the top five hits as identifications from IgBLAST. To compare both methods, we also ran IgBLAST with different word sizes: the default word size 9 and the minimum word size allowed of 4, which we determined would give optimal identification results.

We determined if identification occurred correctly at the gene level only, not at the level of alleles. We have divided the results into three categories to evaluate the performance of each method.

Category 1: the gene is distinguished and the unique identification is correct

Category 2: the correct gene is identified along with the expected confusing gene (as described in **section 2.1**).

Category 3: other, unpredicted misidentifications

2.4 Germline and mutant sequences used

(i) Germline sequences analyzed for V-ties: All germlines and alleles analyzed are from the IMGT database version 3.1.2, also in Supplemental Table 14. The exact identification of V-ites will depend on list of known germline genes that is queried.

(ii) Human IgH rearrangements: Peripheral blood B cell DNA was enriched using a dual step PCR based amplicon capture as described previously [21].

(iii) Simulated sequences: Simulated sequences were made with V genes randomly mutated with a 0.02, 0.05, 0.10, 0.15 and 0.3 mutation frequencies, uniformly spread across the sequence, but not the anchoring points. We generated 500 mutated sequences for each of the 192 known functional and non-partial human VH gene alleles or a total of 96,000 sequences. We did not mutate the anchor sites described in 2.2. This is because we are attempting to test the miss-assignment of expected sets of germlines that form V-ties. The mutation of anchor sites does not in any way change the likelihood of confusing germline V genes as the anchor sites are identical in all germline V genes and thus have no power in determining them. These sequence datasets can be provide in fasta format on request.

3. Calculating the likelihood of confusing two V genes

We calculated the similarities of the BCR and TCR V genes and alleles by how many nucleotide differences they have at certain sequence lengths. The probability *p* of two specific V alleles, of the same or different V genes, to be confused at a particular length and mutation frequency is given by the following equation:

$$p=\int_{K/2}^{K} \mathrm{hype}(x, M, K, N) \times \mathit{0.33}^{x}$$

Where *p* is the hypergeometric probability of each value of *x* (from *K/2* to *K*) using the corresponding size of the population, *M*, number of items with the desired characteristic in the population, *K*, and number of samples drawn, *N*. In this distribution, if we assume that mutations have equal probability of targeting at each position, *M* stands for the length of alignment, *K* the differences between two genes/alleles and *N* is the number of estimated mutations in the sequence. The estimated mutation number *N* is calculated from the average alignment length *L* and *r* is the average fraction of sequence positions that are mutated. 0.33 is the probability of one nucleotide mutating into others, assuming equal chance (no bias). Although this form of calculation ignores known patterns of mutation it allows us to generalize the method across species and include species where a good model of mutation does not exist. Also we do not set a prior probability for V gene or allele usage. The knowledge of exact germline gene usage across the population in humans is still very limited and is at present beyond the scope of this analysis.

4. Results

4.1. Germline V gene similarity

We counted the number of different nucleotides between any two human heavy chain variable alleles when counting from the 3' end (Figure 3). We found genes from the same families were often highly similar to each other, especially in the VH1 and VH4 families. In addition, certain genes in the VH3 family are very similar. Certain alleles in IGHV3-30 and IGHV3-33 only have two nucleotide differences. IGHV3-30-5*01 and IGHV3-30*18 have exactly the same nucleotides. We also calculated the differences within BCR light chains and both β and α TCR V genes (Supplemental Figures 1–4).

Using the hypergeometric calculation described in the **Methods and Calculations** sections, we found that the likelihood of failed V gene assignment increases, as one would predict, with higher mutation frequencies and shorter read lengths (Figure 4 and Supplementary Zip file). As shown in Figure 4, we calculated at 150 nucleotide length and 0.05 mutation frequency, that some VH genes (Supplemental Table 2) have more than 0.01 probability of being confused with each other and thus mutants from these germline V genes cannot be definitively distinguished. We would call these germline V genes at this sequence length/ mutation level V-ties. Mutated sequences that are assigned either of these germline V genes should instead be assigned the V-ties they belong to (Supplemental Table 2).

We have generated sets of tables delineating exactly which VH genes cannot be unequivocally uniquely identified at 0.05, 0.15 and 0.3 mutation frequencies and 100, 150, 200 and full-sequence lengths (Supplemental Tables 1–4 and Supplementary Zip file). We have done so for human BCRs and TCRs (Supplementary Zip file). While TCRs do not mutate, in many cases very short reads are used and there can be a sequencing error of ~ 1–3% [22]. We therefore only created such tables with 0.03 error frequencies. TCR α chain V gene germlines can be clearly distinguished even at 100 nucleotide length. The set of V genes that cannot be distinguished for TCR β chain can be seen in Supplemental Table 15. We have also calculated the probabilities of gene pairs confusing with each other at aforementioned mutation levels and sequence lengths for human BCR and TCRs so one can set his/her own threshold instead of 0.01 used in this paper (Supplementary Zip file). Finally we supply a Matlab code (https://github.com/DrexelSystemsImmunologyLab/ConservedIdentification.git) that can filter sets of genes with identified germline sequences so that the V genes that cannot be uniquely identified are explicitly identified the appropriate germline V-ties are assigned.

4.2. Identifying V-ties with the Anchor method, with High V-Quest and IgBLAST

To see if V-ties appeared where they were predicted to appear we compared the V assignments using two standard algorihems for V gene assignment and our novel Anchor method. To do so we used simulated sequences (mutated as described in **Methods**) as only with those could we *know* a-priori their actual germline source. To our surprise we observed that not all the methods did a good job of identifying germlines. IgBLAST at its default settings has very poor VH gene identification. However, all three methods can give reasonable results by optimizing the sequence feature parameters (Figure 5 at 150 nucleotide length and 0.15 mutation frequency and Supplementary Figures 5–15 for 100, 150, 200 nucleotide and full length and with 0.05. 0.15 and 0.30 mutation frequencies).

Most importantly, in terms of our predication of V-ties, for all methods, whenever sequences are partial length or somatically mutated, a consistent subset of genes will be misidentified in the way we predicted (category 2 identification – green bars in Figure 5 and see **Calculations and Methods** section). It is important to note that such misidentifications could not be distinguished from correct identifications or other types of error in a non-simulated set of sequences. Thus, as we suggest in **Methods**, the only solution for these consistently miss-assigned germline V genes is to combine them with their appropriate confounding germline V genes (Figure 5). These sequences should not be confused with sequences that are equidistant from two different germline genes by their mutation count. The germline V genes identified as V-ties are clearly identified at the level of the specific V-tie, and will most often be identified as being closest to one specific germline V of those associated by the V-tie. However, their chance of being randomly identified or misclassified as a specific *other* V gene is significant (considered to be $p > 0.01$ in our **Calculations** and **Methods** section) and either the real gene or the one we predict to confuse it with are identified; red is the fraction of other incorrect identifications that cannot be explained by V-ties.

4.3 Comparing the Anchoring method to other alignment methods

Computational efficiency—To test the efficiency of our germline association method, we selected 10,000 sequences (299±6 nucleotides in length that have a partial VH gene sequence, all of the junctional sequences and a partial JH gene sequence) with an estimated 0.03 mutation frequency [21] (see Section 2.4). It took our method 35.54 seconds to finish the identification while IgBLAST needed 347.58 seconds using the default word size of 9 nucleotides and 3,877.59 seconds using the minimum word size (4 nucleotides). This analysis shows that our method is less computationally intensive than IgBLAST, especially when using the more accurate minimum word size there. Although High V-Quest is clearly the standard high-throughput alignment program used in our field it does not have a stand-alone version and jobs need to be queued on the IMGT server. This makes that and not the processing speed the relevant time limiting step in using it. Thus from what we can compare we can conclude that the Anchoring method outperforms IgBLAST (×10 compared to the default word size and ×100 when using the more accurate minimal word size).

Sequence loss—To test to what extent the use of the Anchor positions causes us to lose sequence data, we compared a single IMGT High V-Quest run of 150,000 sequences to our alignment of the same set of genes. The sequences were taken at random from a set of 1.8 million B cell heavy chain V(D)J gene sequences sequenced from human blood [21]. From these 150,000 sequences, we were able to identify 141,496 (94.33%) using the V Anchor method discussed in **Section 2.2**, while with High V-Quest we identified 146,821 (97.88%). Discounting duplicated sequences, we identified 88,209 unique sequences using the Anchor method described here and High V-Quest identified 92,490. 87,958 of these unique sequences were in complete agreement with respect to their copy number and V identity between two methods. There were also a few sequences identified only by one method. The V Anchor method identified 154 unique sequences and High V-Quest 4,460 unique sequences that the other method did not.

It is important to note that much of the extra diversity that was only identified by High V-Quest (5% more sequences) is probably due to sequencing error. To remove unique sequence types generated through sequencing error we next considered only unique sequences with copy number >1 [11]. The removal of singleton sequences indeed improves our level identification in comparison to High V-Quest. Both methods agree on the identity of 15,252 unique sequences of copy number >1 and High V-Quest identifies only 357 additional sequences (or 2% more). There are also a few genes that are identified by both methods where they do not agree as to the V gene assignment and/or the copy number. The discrepancies between the two methods are interesting as they reveal the minor issues with each method. One reason for these differences is that we are comparing partial sequences of unequal length. With High V-Quest, alignments are performed not only to functional V genes but also to pseudogenes and incomplete V genes in the IMGT database. Hence there are 15 unique sequences to which the Anchor method assigns specific V genes that IMGT High V-Quest considers to most closely resemble partial sequences or pseudogenes. In addition there are 13 unique sequences that IMGT's High V-Quest assigns to a pseudogene whose component sequences we considered to be two different V genes. Finally there are 4 sequences (out of 150,000) that were identified as containing indels by High V-Quest and did not pass our threshold (Supplemental Table 16–18). In all instances it is hard to categorically state which method was correct. However, these examples pinpoint potential limitations in our method of insertion/deletion detection and in High V-Quest user operability, which does not allow the user to select which subset or version of the database of germline genes they wish to compare to the mutant sequences.

5. Discussion

The BCR and TCR repertoires play critical roles in immune function and pathogenesis [23, 24]. One of the first steps in studying immune responses is to study how immune repertories shift in response to antigens, vaccines and pathogens. Identification of germline genes that comprise the building blocks of the antigen receptor is a crucial first step in repertoire analysis. Unfortunately, this step is difficult because germline genes can be highly similar and can undergo somatic mutation (only in the case of BCRs) and be subject to sequencing error (BCR and TCR). High-throughput sequencing methods generate large numbers of sequences at a low cost, providing a way to essentially map the immune repertoire, but can use short read lengths and have high sequence error rates. For this reason it is now critical to categorize, as much as possible, the reason for uncertainty in germline gene assignment.

As we have shown in the **Results**, at all sequence lengths, the V genes from the same gene family are quite similar and can have differences as low as one or two nucleotides. Some alleles of different V genes are even identical to each other over short lengths. For example, IGHV3-30*07 and IGHV3-33*04 are identical in the last 119 nucleotides. As the sequence read length increases, V genes can be better differentiated, but even with full-length sequence data if the mutation frequency is above 0.1, some germline V genes are effectively indistinguishable.

This raises the problem that some germline V genes cannot be well discriminated. However, they are not unknowable as they are similar only to other specific V genes and can be discriminated from most other germline V genes. In our method, after the first round of identification, we estimate the expected variability in our data. Based on the alignment length and mutation frequency of the first round identification, we can calculate the likelihood of error due to mutation (or other sequence changes) using the simple hypergeometric test described in **Calculations**. V genes that we calculate as being impossible to distinguish at a given length/mutation/error rate will be identified as a single germline source. When we construct clones in later steps these V genes will be put in the same clone as they are indistinguishable. In this way, we can have more reliable identifications and know at what level of categorization (family, gene or allele) the identifications are definitive. This is especially useful in studies of inflammation wherein B cellclones can be highly mutated.

Figure 6 shows four sequences [21] that would not be associated into a single clone if we did not consider V-ties. Applying V-ties, they were identified as IGHV3-30, IGHV3-33 and/or IGHV3-NL1, and put in the same clone, as these genes have a $p > 0.01$ to be confused at the length (90.45 nucleotides) and mutation frequency (0.02) found in this dataset. Despite the low mutation rate in the sample in general, these sequences had two mutations (C210T and C354A) in common and the same CDR3 (CARDRASCPDYW) confirming that they probably belonged to the same clone, which would have been missed if standard V gene alignment practices had been followed [11]. Figure 7 shows another example where allowing for a more ambiguous V gene assignment helps to identify the full clone. The identical CDR3 and five common mutations from the germline sequences (T168C, C276T, G301A, G303A and C366A) found in all three sequences suggest they are in the same clone. However, if we had not considered the inherent ambiguity in assigning them to IGHV 3-11 or 3-48, we would have considered them to comprise two separate clones and have assigned them up to 3 erroneous mutations. By giving them the hybrid V-ties assignment we consider only mutations we are sure of and correctly assign these sequences to a single clone (Figure 7c).

Beyond explicit indication of the specificity of V gene identification, our germline identification Anchor method performs as well or better than existing human germline identification and clonal assignment methods. Specifically, at high mutation levels IgBLAST (with the default word size) does not work well. As shown in the **Results**, as mutation frequency increases, the performance of IgBLAST quickly worsens (Figure 5 and Supplemental Figures 5–15). This error rate can be corrected by shortening the word size but then computing time balloons to 100-fold longer than the Anchor method described here. High V-Quest and the Anchor method have equally reliable V gene identifications. The Anchor method allows for command line alignment of sequences and control over the members of the germline V gene database used to compare with the query sequences. High V-Quest identifies ~ 2% more unique sequences than the Anchor method. Lack of control of the gene database used to compare with the query sequences is problematic when we have knowledge of our input beforehand. For example we know the V genes are all functional when we extract the sequence from immunized patient blood, but if we compare sequences to non-functional germline genes as well we could assign them as germline source if mutant sequences happened to exhibit more similarity to them and not their germline source. Finally, the V Anchor method allows us to ensure that all sequences are aligned in terms of IMGT numbering even if we are uncertain of their exact germline source. In this way we can both retain information on CDR3 structure and at least some measure of intra-clonal diversity analysis can be achieved even if we are not sure of the precise germline source of the clone.

In summary, we have developed a new methodology (Anchor method) to rapidly identify the originating germline genes of rearranged antibody and TCR sequences, based on conserved sites. In addition, we have analyzed the similarity of V genes at different lengths, and created matrices of small V gene groups (V-ties) that cannot be discriminated because of sequence similarity of their germline genes for different levels of sequence information. The exact identification of V-ties will depend on the list of known germline genes that is queried. Herein we have identified V ties for the current, most widely used human and murine IMGT heavy chain germline V gene lists. The exact groups of V-ties we show here derive from the genes in these lists. The V-ties would change if we added new V gene alleles and of course will also change when we look at the immune systems of other species. However, the methodology and consequences of our analysis remain the same: We will identify some germline genes/ alleles that cannot be uniquely associated or discriminated when querying their mutant progeny.

To allow scientists to identify the V-ties in their data, given the set of relevant germline genes in the repertoires they are analyzing, we have created a simple set of programs (in Matlab). These programs implement the Anchor alignment method and post process any BCR alignment results, from High V-Quest or IgBLAST or our own conserved site Anchor method, so as to assign V genes to their appropriate V-ties under different sequencing lengths and levels of mutation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors would like to thank Gregory Schwartz and Aaron Rosenfeld for their insightful comments and helpful arguments. Research reported in this publication was supported by the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health under Award Number P01AI106697 and by NIH P30-CA016520. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Alt, Frederick W.; Baltimore, David. Joining of Immunoglobulin Heavy Chain Gene Segments: Implications from a Chromosome with Evidence of Three D-JH fusions. Proceedings of the National Academy of Sciences. 1982; 79(13):4118–4122.

2. Petrie, Howard T.; Livak, Ferenc; Burtrum, Douglas; Mazel, Svetlana. T Cell Receptor Gene Recombination Patterns and Mechanisms: Cell Death, Rescue, and T Cell Production. The Journal of Experimental Medicine. 1995; 182(1):121–127. [PubMed: 7790812]
3. Tonegawa, Susumu. Somatic Generation of Antibody Diversity. Nature. 1983; 302(5909):575–581. [PubMed: 6300689]
4. Grimaldi, Christine M.; Hicks, Ruthmarie; Diamond, Betty. B Cell Selection and Susceptibility to Autoimmunity. The Journal of Immunology. 2005; 174(4):1775–1781. [PubMed: 15699102]
5. Goronzy, Jörg J.; Weyand, Cornelia M. Ageing, Autoimmunity and Arthritis: T-cell Senescence and Contraction of T-cell Repertoire Diversity – Catalysts of Autoimmunity and Chronic Inflammation. Arthritis Research and Therapy. 2003; 5(5):225–234. [PubMed: 12932282]
6. Abe, Jun; Kotzm, Brian L.; Melssner, Cody; Melish, Marian E.; Takahashi, Masato; Fulton, David; Romagne, Francois; Malissen, Bernard; Leung, Donald Y. Characterization of T Cell Repertoire Changes in Acute Kawasaki Disease. The Journal of Experimental Medicine. 1993; 177(3):791–796. [PubMed: 8094737]
7. Hershberg, Uri; Meng, Wenzhao; Zhang, Bochao; Haff, Nancy; St Clair, E William; Cohen, Philip L.; McNair, Patrice D.; Li, Ling; Levesque, Marc C.; Luning Prak, Eline T. Persistence and Selection of an Expanded B-cell Clone in the Setting of Rituximab Therapy for Sjögren's Syndrome. Arthritis Research & Therapy. 2014; 16(1):R51. [PubMed: 24517398]
8. Houghton, Alan N. Cancer Antigens: Immune Recognition of Self and Altered Self. The Journal of Experimental Medicine. 1994; 180(1):1–4. [PubMed: 8006576]
9. Berglund, Eva C.; Kiialainen, Anna; Syvänen, Ann-Christine. Next-generation Sequencing Technologies and Applications for Human Genetic History and Forensics. Investig Genet. 2011; 2(2011):23.
10. Liu, Lin; Li, Yinhu; Li, Siliang; Hu, Ni; He, Yimin; Pong, Ray; Lin, Danni; Lu, Lihua; Law, Maggie. Comparison of Next-Generation Sequencing Systems. BioMed Research International. 2012; 2012
11. Hershberg, Uri; Luning Prak, Eline T. The Analysis of Clonal Expansions in Normal and Autoimmune B Cell Repertoires. Philosophical Transactions of the Royal Society B. 2015; 307(1676):20140239.
12. Brochet, Xavier; Lefranc, Marie-Paule; Giudicelli, Véronique. IMGT/VQUEST: the Highly Customized and Integrated System for IG and TR Standardized V-J and V-D-J Sequence Analysis. Nucleic Acids Research. 2008; 36(suppl 2):W503–W508. [PubMed: 18503082]
13. Ye, Jian; Ma, Ning; Madden, Thomas L.; Ostell, James M. IgBLAST: an Immunoglobulin Variable Domain Sequence Analysis Tool. Nucleic Acids Research. 2013; (2013):gkt382.
14. Volpe, Joseph M.; Cowell, Lindsay G.; Kepler, Thomas B. SoDA: Implementation of a 3D Alignment Algorithm for Inference of Antigen Receptor Recombinations. Bioinformatics. 2006; 22(4):438–444. [PubMed: 16357034]
15. Gaëta, Bruno A.; Malming, Harald R.; Jackson, Katherine JL.; Bain, Michael E.; Wilson, Patrick; Collins, Andrew M. iHMMune-align: Hidden Markov Model-based Alignment and Identification of Germline Genes in Rearranged Immunoglobulin Gene Sequences. Bioinformatics. 2007; 23(13):1580–1587. [PubMed: 17463026]
16. Kepler, Thomas B. Reconstructing a B-cell Clonal Lineage. I. Statistical Inference of Unobserved Ancestors. F1000Res. 2013; 2
17. Schwartz, Gregory W.; Hershberg, Uri. Conserved Variation: Identifying Patterns of Stability and Variability in BCR and TCR V Genes with Different Diversity and Richness Metrics. Physical Biology. 2013; 10(3):035005. [PubMed: 23735612]
18. Schwartz, Gregory W.; Hershberg, Uri. Germline Amino Acid Diversity in B Cell Receptors is a Good Predictor of Somatic Selection Pressures. Frontiers in Immunology. 2013; 4
19. Stern, Joel NH.; Yaari, Gur; Vander Heiden, Jason A.; Church, George; Donahue, William F.; Hintzen, Rogier Q.; Huttner, Anita J.; Laman, Jon D.; Nagra, Rashed M.; Nylander, Alyssa; Pitt, David; Ramanan, Sriram; Siddiqui, Bilal A.; Vigneault, Francois; Kleinstein, Steven H.; Hafler, David A.; O'Connor, Kevin C. B Cells Populating the Multiple Sclerosis Brain Mature in the Draining Cervical Lymph Nodes. Science Translational Medicine. 2014; 6(248):248ra107–248ra107.

20. Lefranc, Marie-Paule; Pommie, Christelle; Ruiz, Manuel; Giudicelli, Veronique; Foulquier, Elodie; Truong, Lisa; Thouvenin-Contet, Valerie; Lefranc, Gerard. IMGT Unique Numbering for Immunoglobulin and T cell Receptor Variable Domains and Ig Superfamily V-like Domains. Developmental & Comparative Immunology. 2003; 27(1):55–77. [PubMed: 12477501]

21. Meng, Wenzhao; Jayaraman, Sahana; Zhang, Bochao; Schwartz, Gregory W.; Daber, Robert D.; Hershberg, Uri; Garfall, Alfred L.; Carlson, Christopher S.; Luning Prak, Eline T. Trials and Tribulations with VH Replacement. Frontiers in Immunology. 2014; 5

22. Marshall, Brendan; Schulz, Ruth; Zhou, Min; Mellor, Andrew. Alternative Splicing and Hypermutation of a Nonproductively Rearranged TCR Alpha-chain in a T cell Hybridoma. The Journal of Immunology. 1999; 162(2):871–877. [PubMed: 9916710]

23. Mauri, Claudia; Bosma, Anneleen. Immune Regulatory Function of B cells. Annual Review of Immunology. 2012; 30(2012):221–241.

24. Kronenberg, Mitchell; Siu, Gerald; Hood, Leroy E.; Shastri, Nilabh. The Molecular Genetics of the T-cell Antigen Receptor and T-cell Antigen Recognition. Annual Review of Immunology. 1986; 4(1):529–591.

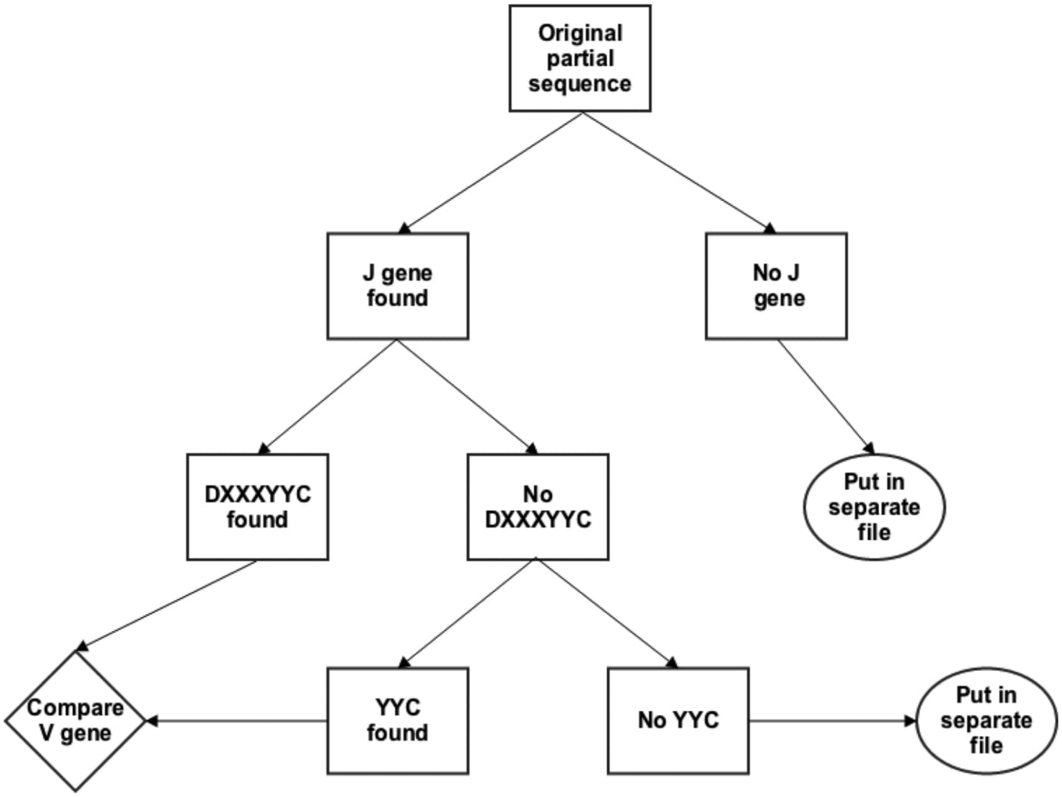


Figure 1. Workflow of germline association process

(1) Search partial sequences for germline J gene signature. Sequences with no J gene found are put to a separate file. (2) Sequences with identified J gene are checked for germline V anchor DXXXYYC. (3) Counting from the anchor positions, sequences are compared to all germlines and minimally distant germline(s) is/are assigned. (4) Sequences without this anchor will be checked for second anchor YYC and similarly compared. (5) Sequences without any anchor sites are put to a separate file. (For further details see text, **Methods Section 2.2**)

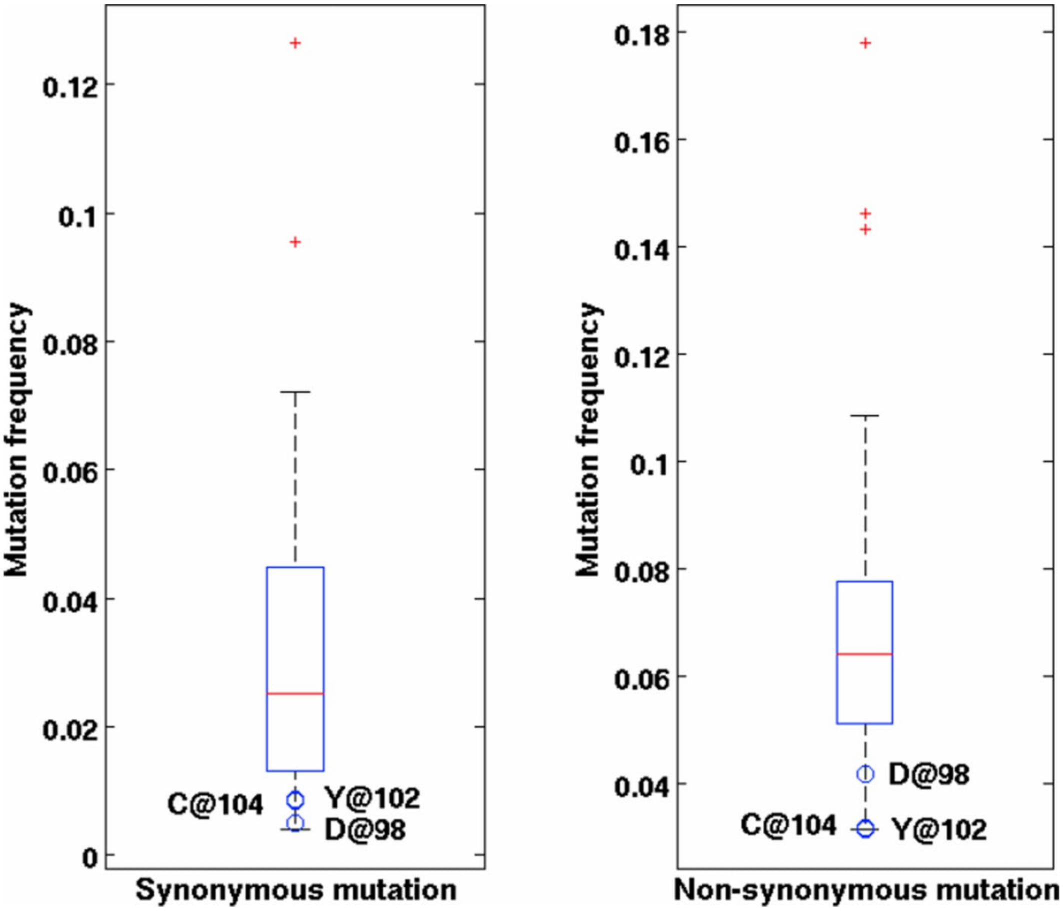


Figure 2. Boxplots of mutation frequency of amino acid positions 76–105 in [21]

Red line indicates median. Blue box indicates 25% and 75% quartile. Whiskers indicate the furthest data not considered outlier. Red dots indicate outliers. Blue circles indicate mutation frequencies of D, Y and C at position 98, 102 and 104. (A) Synonymous mutations; (B) Non-synonymous mutations.

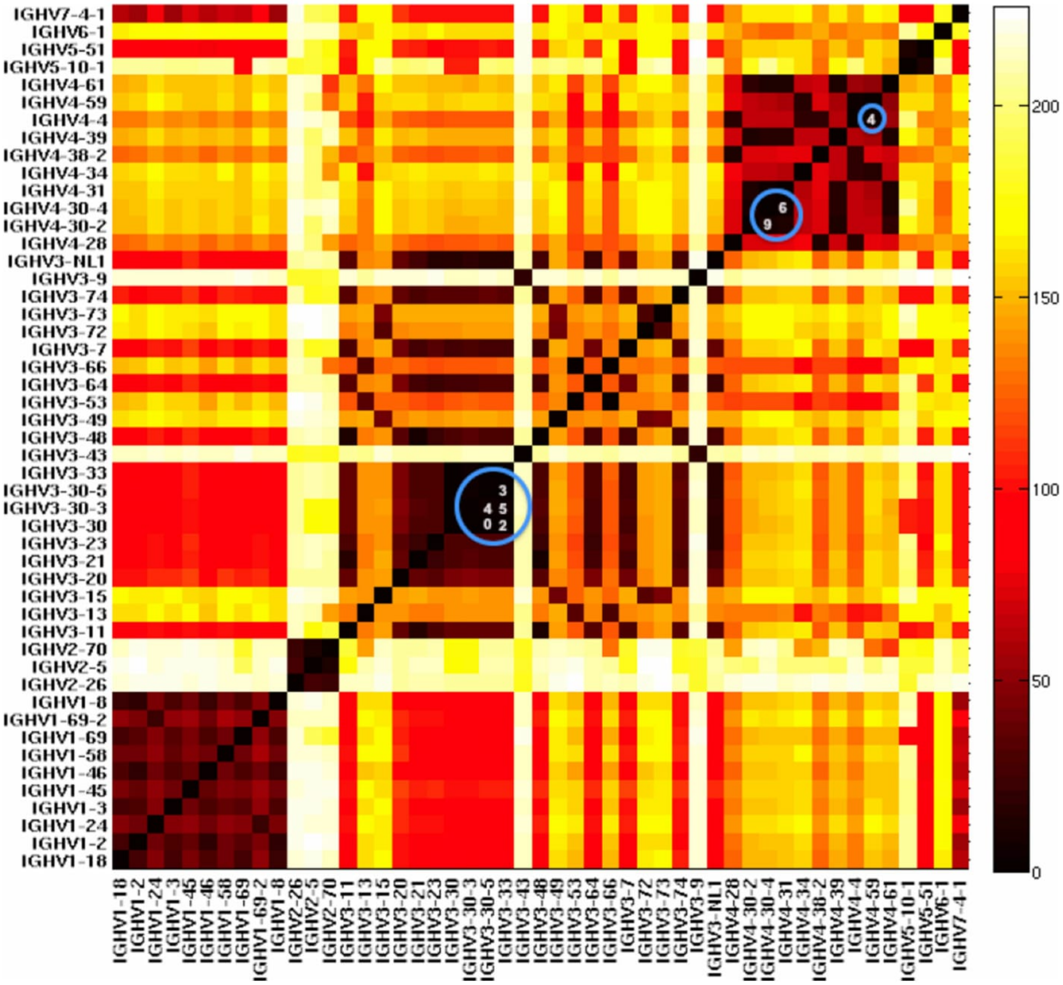


Figure 3. Heat map of the minimal sequence difference of each human germline VH gene pair [12] at full sequence length

Minimal nucleotide differences among all alleles pairs between genes. Distances range from 0 (black) through red (100) and yellow (200) to white (>200). The comparison begins at the 3' end of each full-length VH gene. The blue numbers in the circles represent the numbers of mismatched nucleotides in the most similar pairs. These pairs of VH genes cannot be discriminated from each other at almost any length or mutation frequency.

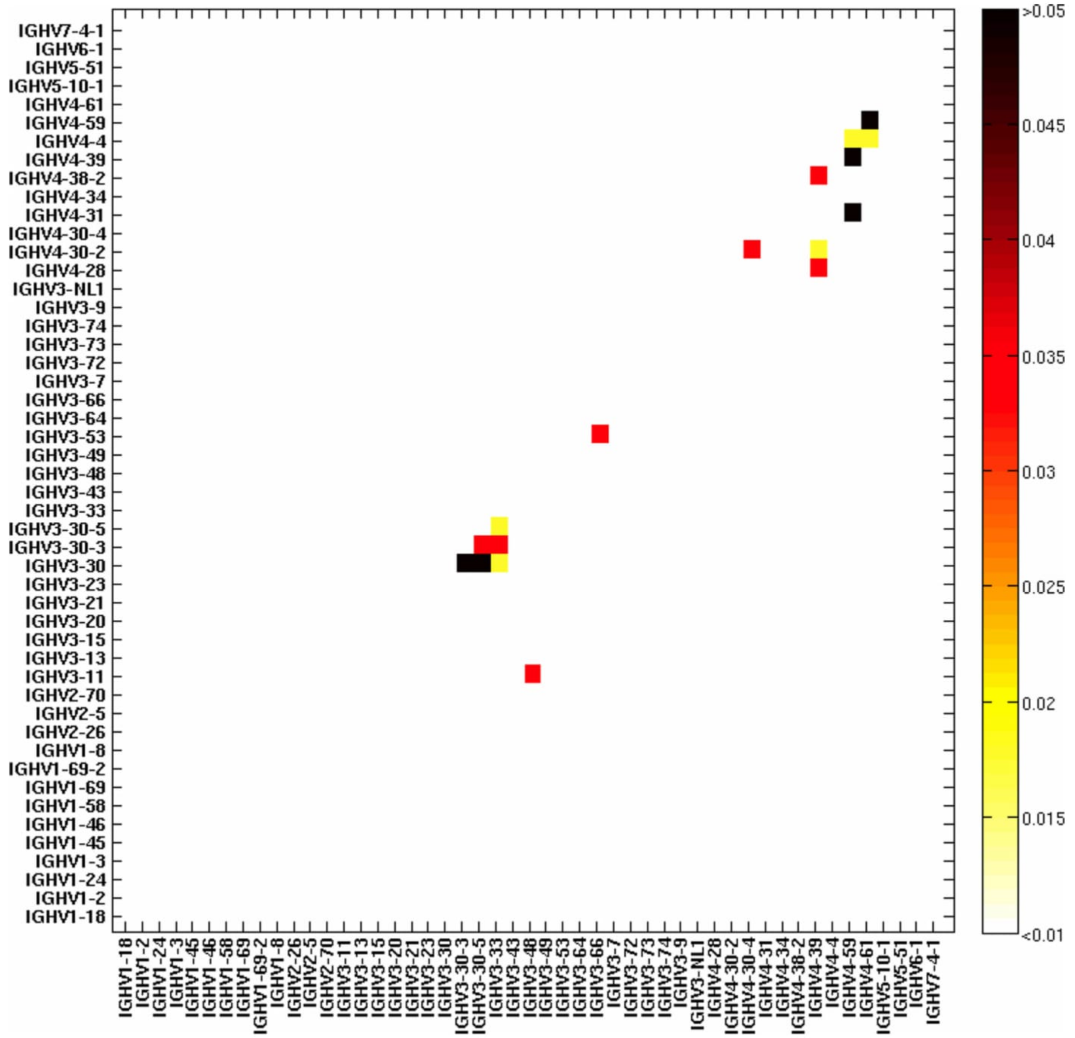


Figure 4.

The probabilities of confusing the germline association of a mutated sequence between any two germline VH genes at 150 nucleotide length and 0.05 mutation frequency (calculated as described in Calculations). Likelihoods range from white *p<0.01* to black *p>0.05*).

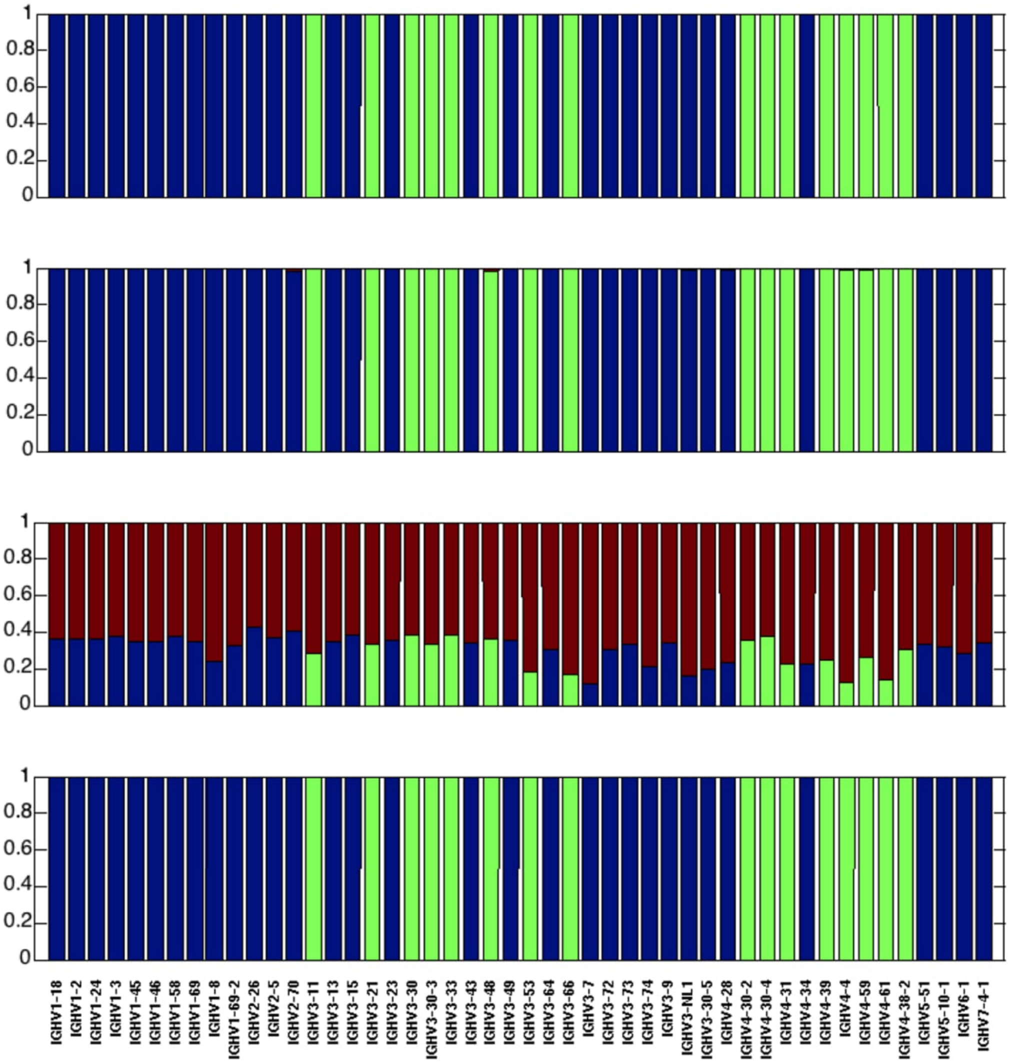


Figure 5. Comparison of human VH identification results using three different sequence identification methods at n=150 nucleotide read length and 0.15 mutation frequency

From top to bottom **(A)** The Anchor method (this paper); **(B)** High V-Quest; **(C)** IgBLAST using the default word length (9 characters); **(D)** IgBLAST using the minimal word length (4 characters). In all cases the blue color represents the fraction of correctly identified sequences in which a VH gene is uniquely identifiable; green is the fraction of identifications when a gene is not distinguishable from at least one other gene (using the calculation described in the **Calculation and Methods** section) and either the real gene or the one we predict to confuse it with are identified; red is the fraction of incorrect identifications.

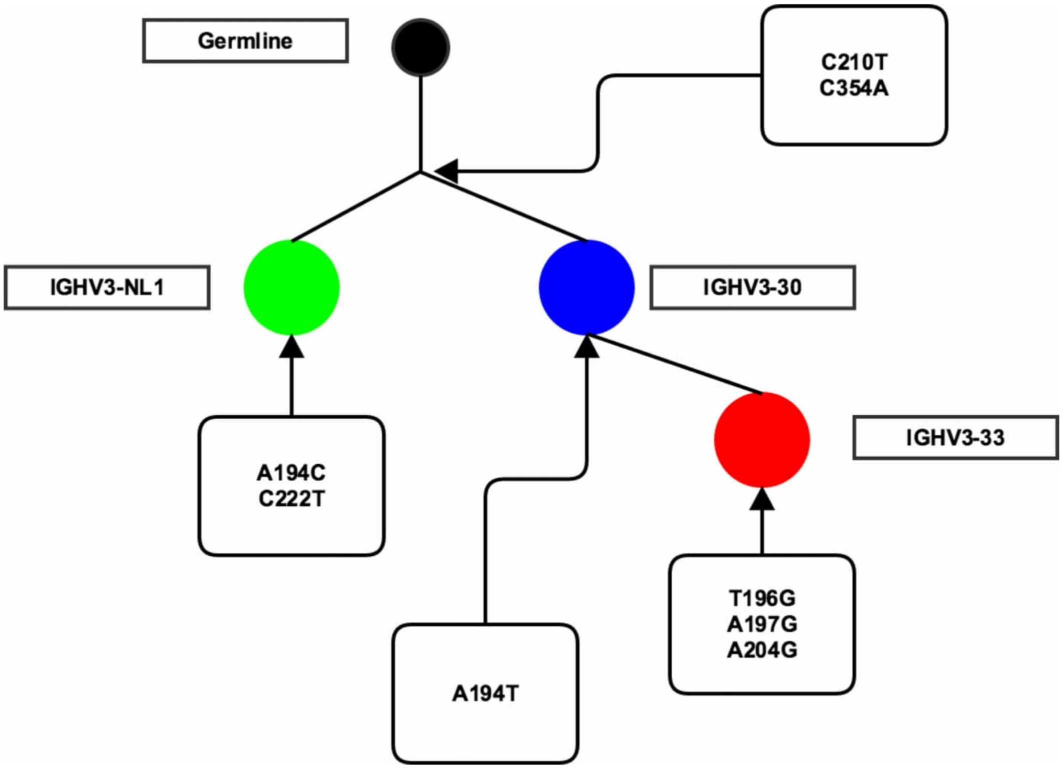


Figure 6. Part of a phylogenetic tree of a clone from naïve B cells in human

Each circle represents a sequence. Different colors show what V gene the sequences were originally identified. The germline was made by ignoring the different positions of tied V genes. The mutations on each branching level are shown. The numbers in each circle show how many additional mutations each sequence has.

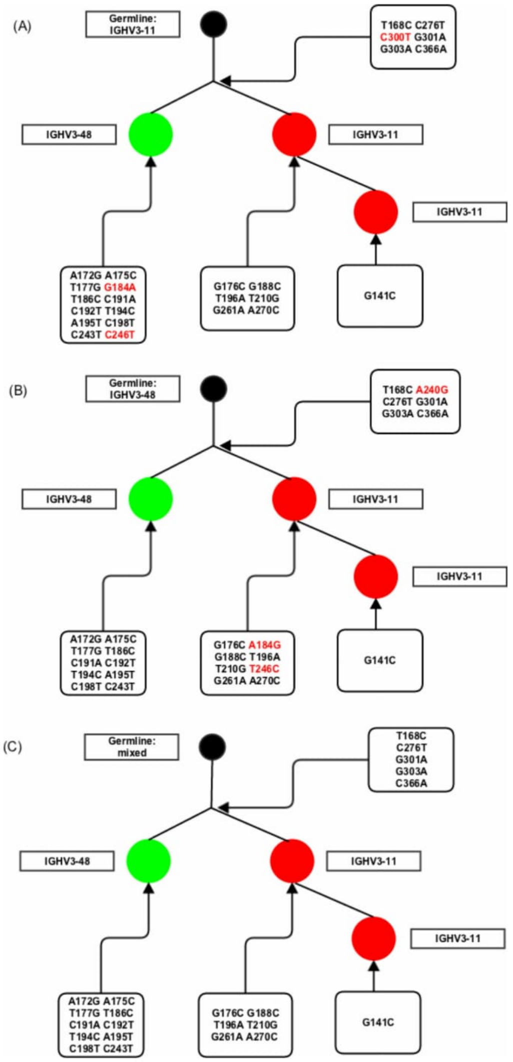

(A)
Germline: IGHV3-11
T168C C276T C300T G301A G303A C366A
IGHV3-48
IGHV3-11
IGHV3-11
A172G A175C T177G G184A T186C C191A C192T T194C A195T C198T C243T C246T
G176C G188C T196A T210G G261A A270C
G141C
(B)
Germline: IGHV3-48
T168C A240G C276T G301A G303A C366A
IGHV3-48
IGHV3-11
IGHV3-11
A172G A175C T177G T186C C191A C192T T194C A195T C198T C243T
G176C A184G G188C T196A T210G T246C G261A A270C
G141C
(C)
Germline: mixed
T168C C276T G301A G303A C366A
IGHV3-48
IGHV3-11
IGHV3-11
A172G A175C T177G T186C C191A C192T T194C A195T C198T C243T
G176C G188C T196A T210G G261A A270C
G141C

Figure 7. Phylogenetic tree of a clone from a plasmablast dataset of a lupus patient

Each circle represents a sequence. Different colors and tags show what V gene the sequences were originally identified. The germline was made by **(A)** filling in IGHV3-11*01 germline sequence, **(B)** filling in IGHV3-48*01 germline sequence or **(C)** ignoring the different positions of these V genes. The mutations of each hypothetical and real node are shown. The additional mutations in (A) and (B) compared to (C) are shown in red.

Table 1

Comparison of existing methods

Program	Algorithm	Stand alone version	Give multiple identifications?	Control of germline database
IMGT/High V-Quest	Local alignment	No	Yes (no quality score)	No
IgBLAST	BLAST searches performed against a user-selected germline V gene database	Yes	Yes	Yes
SoDA	Local alignment and 3D dynamic programming	Yes	No	Yes
iHMMune-align	Hidden Markov model	Yes	Yes	Yes
Conserved Anchor method	Hamming distance after finding conserved anchors	Yes	Yes (when tied)	Yes

Table 2

Nucleotides used to identify human JH genes, at positions 46–63 by IMGT numbering

J gene	Nucleotide
IGHJ1/4/5	TGGTCACCGTCTCCTCAG
IGHJ2	TGGTCACTGTCTCCTCAG
IGHJ3	TGGTCACCGTCTCTTCAG
IGHJ6	CGGTCACCGTCTCCTCAG

Table 3

Nucleotide variations on ‘TATTACTGT’ & ‘GAC’ in human VH genes and alleles

Incomplete YYC			
Name	AA position	NT sequence	AA sequence
IGHV2-70*02	102–104	TATTACTG	YY
IGHV2-70*03	102–104	TATTACTG	YY
IGHV2-70*04	102–104	TATTACTG	YY
IGHV2-70*06	102–104	TATTACTG	YY
IGHV2-70*07	102–104	TATTACTG	YY
IGHV2-70*08	102–104	TATTACTG	YY
Different nucleotide or AA sequences			
Name	AA position	NT sequence	AA sequence
IGHV2-70*13	102–104	TATTATTGT	YYC
IGHV3-20*01	102–104	TATCACTGT	YHC
IGHV4-31*10	102–104	GACTACTGT	DYC
IGHV4-34*11	102–104	TATTGCTGT	YCC
IGHV4-4*01	102–104	TATTGCTGT	YCC
IGHV7-4-1*05	102–104	TGTTACTGT	CYC
IGHV3-30*05	98	GGC	G
IGHV4-31*05	98	GCG	A

Table 4

Numbers of synonymous and non-synonymous mutations of the anchor positions in 92,491 unique DNA sequences identified using High V-Quest [21] and 3,017 mRNA sequences with copy number >1 [19].

	DNA				mRNA			
Position	**98**	**102**	**103**	**104**	**98**	**102**	**103**	**104**
Amino acid	D	Y	Y (H/C)	C	D	Y	Y (H/C)	C
Nucleotide	GAC	TAT	TAC	TGT	GAC	TAT	TAC	TGT
Synonymous	457	807	1921	785	14	99	178	79
Non-synonymous	3850	2934	7381	2921	54	0	0	0

Table 5

Anchors in other BCR and TCR genes. The default amino acid at position 98 is D. The default amino acids at position 102–104 are YYC. Leucine (L) and Phenylalanine (F) are very common variants at position 103.

			D at 98		**YYC at 102–104**		
Number of alleles	**V gene type**	**Total No. of germline genes**	**With NT variation**	**With AA variation**	**With NT variation**	**With AA variation (YLC/YFC)**	**With AA change at both positions**
Human	□	49	2	1	6	5	0
	κ	61	0	0	29	2	0
	λ	65	20	0	31	4	0
	α	88	25	3	0	78 (24/48)	0
	β	106	34	28	0	106 (51/41)	0
Mouse	Heavy	258	22	3	99	50	1
	κ	120	20	0	74	31	0
	α	206	40	0	0	165 (36/124)	0
	β	46	17	16	0	46 (19/21)	0

Article 2. Trials and Tribulations with VH Replacement

frontiers in
IMMUNOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 30 January 2014
doi: 10.3389/fimmu.2014.00010

Trials and tribulations with VH replacement

Wenzhao Meng[1], Sahana Jayaraman[1], Bochao Zhang[2], Gregory W. Schwartz[2], Robert D. Daber[1,3], Uri Hershberg[2,4], Alfred L. Garfall[5], Christopher S. Carlson[6] and Eline T. Luning Prak[1]*

[1] Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[2] School of Biomedical Engineering, Science and Health Systems, Drexel University, Philadelphia, PA, USA
[3] Center for Personalized Diagnostics, University of Pennsylvania Health System, Philadelphia, PA, USA
[4] Department of Microbiology and Immunology, College of Medicine, Drexel University, Philadelphia, PA, USA
[5] Division of Hematology-Oncology, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[6] Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Edited by:
Zhixin Zhang, University of Nebraska Medical Center, USA

Reviewed by:
Paolo Casali, University of Texas Health Science Center, USA
Zhixin Zhang, University of Nebraska Medical Center, USA

***Correspondence:**
Eline T. Luning Prak, Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
e-mail: luning@mail.med.upenn.edu

VH replacement (VHR) is a type of antibody gene rearrangement in which an upstream heavy chain variable gene segment (VH) invades a pre-existing rearrangement (VDJ). In this Hypothesis and Theory article, we begin by reviewing the mechanism of VHR, its developmental timing and its potential biological consequences. Then we explore the hypothesis that specific sequence motifs called footprints reflect VHR versus other processes. We provide a compilation of footprint sequences from different regions of the antibody heavy chain, and include data from the literature and from a high throughput sequencing experiment to evaluate the significance of footprint sequences. We conclude by discussing the difficulties of attributing footprints to VHR.

Keywords: V(D)J recombination, VH replacement, VH, DH, and JH gene segments, receptor editing

CONTEXT, DEFINITION, AND POTENTIAL MECHANISMS OF VH REPLACEMENT

Antibodies are heterotetrameric proteins comprised of two heavy chains and two light chains that are formed through V(D)J recombination to generate a highly diverse repertoire of antigen binding receptors expressed by B cells. The *recombinase activating gene* encoded proteins, RAG1 and RAG2, target conserved heptamer and nonamers within recombination signal sequences (RSSs) to cleave the DNA that flanks recombining gene segments that join together to form the variable regions of antibody heavy and light chains [reviewed in Ref. (1)]. Typical V(D)J recombination generates a signal joint and a coding joint, and the latter is further diversified at the junction between the recombining gene segments by mechanisms including P-addition, N-addition, and exonucleolytic nibbling [reviewed in Ref. (2)]. Occasionally atypical rearrangements occur, generating hybrid joints, open-and-shut joints, or joints between RSSs that ordinarily do not recombine (2–5).

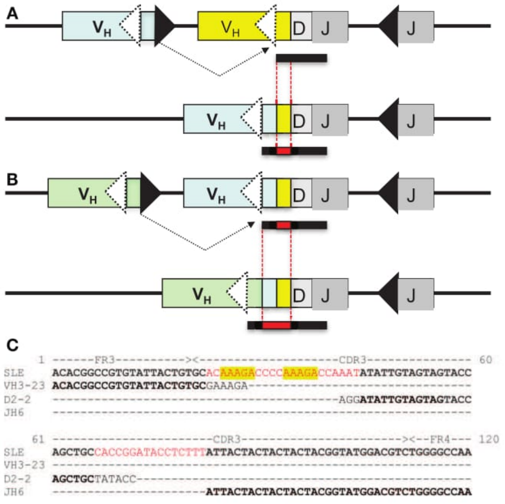
Antibodies can be further revised and diversified through receptor editing of the light chain, somatic hypermutation, gene conversion, and VH replacement (VHR). Receptor editing typically involves RAG-dependent leapfrogging rearrangements on the same allele as the defective or autoreactive light chain, rearrangement on other alleles (κ or λ) and/or RS deletion [which renders preceding κ rearrangement non-functional, reviewed in Ref. (6)]. Somatic hypermutation is DNA point hypermutation carried out by activation induced cytidine deaminase (AID) (7), and typically signifies a T-cell dependent antibody response. Gene conversion, in which homologous sequences from other V genes are grafted into the functional V gene, is a common method of gene diversification in chickens (8), rabbits and more recent examples have been described in horses and humans (9), and appear to be AID-dependent (10). The final category of antibody gene diversification is VHR, which is the focus of this article. Replacement involves the transfer (or invasion) of some or most of another V gene into an existing gene rearrangement.

Darlow and Stott have reviewed the literature on VHR and envision two broad mechanistic classes of V replacement (11). The first, also termed “classical” VHR, consists of invasion of an existing VDJ rearrangement by an upstream VH. In classical VHR there is RAG-mediated cleavage at a cryptic RSS (cRSS) located in the 3′ end of the previously rearranged VH gene. The cRSS has a DNA sequence that differs from the conventional heptamer that flanks the DH gene segment by one nucleotide, bolded in the sequence that follows: 5′-**T**ACTGTG-3′ (12) and is found in ~70% of murine VHs and over 90% of human VHs (13). Occasionally other heptamers containing the 3′ GTG nucleotides can be used, suggesting that the last three nucleotides of the cRSS motif are critical (14, 15). The TGT within the cRSS is the codon encoding the conserved cysteine at the junction between FR3 and CDR3. The second class of replacement, according to Darlow and Stott, involves the transfer of other sequences of homology between different V genes at different sites, many of which appear to also resemble cRSSs. Examples of this second category of VHR have been described in antibodies cloned from single B cells in human tonsils (16), in antibodies cloned from

Abbreviations: CDR3, third complementarity determining region; cRSS, cryptic recombination signal sequence; DH, antibody heavy chain diversity gene segment; footprint 5-mer, pentameric nucleotide sequence attributed to VHR; GVHD, graft versus host disease; IgH, antibody heavy chain; JH, antibody heavy chain joining gene segment; RSS, recombination signal sequence; SLE, systemic lupus erythematosus; VH, antibody heavy chain variable gene segment; VHR, VH replacement.

synovial tissue of patients with rheumatoid arthritis (17), and in antibodies cloned from human mucosa associated lymphoid tissue lymphomas (18). Alternatively or in addition to RAG-mediated rearrangement, replacements in this second category may arise due to AID-mediated homologous recombination events that are unrelated to the putative cRSSs (11). However, the mechanism of type 2 replacement is far from resolved as recently a non-AID-dependent form of replacement has been described at the κ locus using human pre-B cell lines (19). As the molecular mechanism of type 2 replacement remains to be fully elucidated, we will focus the remainder of our analysis in this manuscript on classical VHR (which we refer to hereafter as "VH replacement").

During VHR, an upstream VH gene invades into the cRSS, replacing all but the last few nucleotides of the previously rearranged VH gene (**Figure 1A**). The remaining 3′ nucleotides of the VH, DH, and JH gene segments are retained in the new rearrangement. The extra nucleotides from the 3′ end of the previous VH gene are sometimes referred to as a "footprint." Nearly all human VH genes have between five and nine nucleotides in the potential footprint, between the cRSS and the RSS. Most primary RSS rearrangements delete several of these nucleotides from the 3′ end, so the potential footprint may not be easily recognizable. Moreover, during VHR, additional nucleotides can be deleted, so the footprint from the primary VH can be entirely lost during VHR. It is also possible for more than one replacement rearrangement to occur on the same heavy chain allele, a process referred to as "serial" or "successive" VHR (**Figure 1B**) (20). An example of a heavy chain rearrangement with more than one footprint sequence is given in **Figure 1C**.



```
     1 ------FR3------------><-----------------CDR3---------------- 60
SLE    ACACGGCCGTGTATTACTGTGCACAAAGACCCCAAAGACCAAATATATTGTAGTAGTACC
VH3-23 ACACGGCCGTGTATTACTGTGCGAAAGA--------------------------------
D2-2   -----------------------------------------AGGATATTGTAGTAGTACC
JH6    ------------------------------------------------------------

    61 ----------------------CDR3------------------------><--FR4--- 120
SLE    AGCTGCCACCGGATACCTCTTTATTACTACTACTACTACGGTATGGACGTCTGGGGCCAA
VH3-23 ------------------------------------------------------------
D2-2   AGCTGCTATACC------------------------------------------------
JH6    ----------------------ATTACTACTACTACTACGGTATGGACGTCTGGGGCCAA
```

FIGURE 1 | (A) VH replacement: an upstream VH gene invades by rearranging into a pre-existing rearrangement. RAG cleaves the conventional recombination signal sequence (black triangle) of the invading VH (light blue VH gene) and cleaves at a cryptic heptamer sequence (cRSS, dashed white triangle) of the invaded VH gene (yellow VH gene). The resulting rearrangement is shown on the second line of the diagram and includes the DH and JH genes of the previous rearrangement and the new VH gene. Also included in the VH replacement product is a remnant or "footprint" of the preceding VH gene (denoted by a yellow box). Often the products of VH replacement exhibit CDR3 elongation, due to the retention of the footprint sequence (the CDR3 is indicated by the bar under the sequence and the added length of the new CDR3 sequence including the footprint (in red) is indicated by the black bar below the sequence.) **(B)** Serial VH replacement. The same conventions are used as in **(A)** and a longer CDR3 is generated, via the accumulation of footprint sequences. In both panels, boxes denote exons, lines introns, triangles RSS and the rearrangement is indicated by dashed black lines. This diagram is not drawn to scale. **(C)** Long CDR3 sequence with possible VH replacement(s). Shown is the nucleotide sequence of an expanded B cell clone that was recovered from peripheral blood DNA of a patient with systemic lupus erythematosus (SLE) that reveals a 91 nucleotide CDR3. Kowal et al. described an anti-dsDNA H chain sequence comprised of a VH3-N-DH2-2-JH6, which has similar features to this junction, although it was shorter (21). Sequences in black font match the corresponding germline gene segments. Red font denotes possible N-additions and yellow shading highlights potential footprint sequences. Dashes indicate regions where sequences do not overlap. FR3, framework region 3; CDR3, third complementarity determining sequence; FR4, framework region 4.

DEMONSTRATION OF VH REPLACEMENT IN CELL LINES AND MOUSE MODELS

VH replacement was initially discovered in two different transformed B cell lines (12, 22). In both of these early studies, B cells with non-functional heavy chain gene rearrangements (VDJ−) were able to generate functional heavy chains (VDJ+) by undergoing further heavy chain rearrangement into the cRSS. Continued VHR could also convert a functional VDJ+ rearrangement into a non-functional one through the incorporation of an upstream pseudo-VH gene (12).

The development of antibody heavy chain (IgH) knock-in mice provided a formal demonstration of VHR in B cells *in vivo*. VHR was documented in hybridomas derived from the 3H9 heavy chain knock-in mouse (13). VHR and invasion of upstream DH gene also occurred in a knock-in for the T15 heavy chain (15). Furthermore, B cells from quasi-monoclonal mice, which have an anti-(4-hydroxy-3-nitrophenyl) acetyl (NP) heavy chain knock-in and can only produce λ light chains, due to homozygous engineered κ deficiency, can lose reactivity to NP by VHR. Strikingly, most secreted antibodies in the quasi-monoclonal mouse appear to arise through VHR (23). VHR was also observed in mice that were genetically engineered to contain two non-productively rearranged heavy chain alleles. In these VDJ−/VDJ− mice, IgHs were generated via VHR in a RAG-dependent manner (crossing the VDJ−/VDJ− mice onto a RAG2 deficient background resulted in a failure to generate IgM+ B cells) (24). The ability of RAG1 and RAG2 to bind to the cRSS was also demonstrated by electrophoretic mobility shift assays using VH4-34 cRSS versus consensus 12-RSS sequences (25).

In all of the preceding mouse models, VHR conferred greater diversity or functionality upon the B cell repertoire (i.e., there was a selective pressure that favored VHR). In contrast, when VHR was compared with conventional rearrangement, using a mouse model with an out of frame VDJ rearrangement (VDJ−) that was knocked into the heavy chain locus, conventional rearrangement on the other heavy chain allele occurred far more frequently (26). Similarly, in the 56R anti-dsDNA heavy chain knock-in mouse, receptor editing was far more efficient in B cells that were heterozygous rather than homozygous for 56R (27). One caveat to the 56R study was that cells that had undergone VHR on one allele but were still left with a functional copy of the DNA-reactive 56R heavy chain on the other allele could be counter-selected.

VH REPLACEMENT IN BONE MARROW B CELLS

To gain further insight into the mechanism of VHR, studies were performed in mice to determine its developmental timing. Several studies suggest that VHR occurs at or near the time of conventional IgH gene rearrangement. The junctions of IgH sequences with evidence of VHR in IgH knock-in mice usually contain N-additions (13). Terminal deoxynucleotidyl transferase (TdT), the enzyme that carries out N-addition, is typically expressed at highest levels during H chain rearrangement in pro-B and large cycling pre-B cells (28). Therefore, the presence of N-additions provides indirect evidence that VHR occurred at the time when TdT was active and therefore probably took place in pro-B or early pre-B cells. Further evidence in support of VHR in early stage B cells includes ligation-mediated PCR to measure DNA breaks at the heavy chain locus, which occurred at the highest levels in pro-B cells (29). These studies suggest that VHR is either occurring in cells where IgH rearrangement has not yet shut down (failed allelic exclusion) or is driven by pre-BCR rather than BCR signaling, since only the former receptor is expressed at the pre-B cell stage of development.

With respect to pre-BCR signaling [reviewed in Ref. (30)], it is noteworthy that surrogate light chain knock-out mice have autoreactive antibodies with long CDR3 sequences (31). One potential explanation for this result is that, in the absence of surrogate light chain, the pre-BCR does not assemble and turn off heavy chain rearrangement. Without a heavy chain rearrangement stop signal, there may be higher frequencies of VHR, leading to CDR3 elongation. However, an alternative possibility is that peripheral selection of B cells with long CDR3 sequences is relaxed in the lymphopenic setting that arises due to inefficient primary B cell production in surrogate light chain knock-out mice. It is known that in the absence of normal numbers of peripheral B cells, the level of the B cell survival factor BLyS (also known as BAFF) increases, since B cells are the primary consumers of BLyS. It is also known that the stringency of B cell selection can be reduced when BLyS levels are increased (32, 33).

VH REPLACEMENT IN PERIPHERAL B CELLS

Some studies suggest that VHR could occur in more mature B cell subsets. For example, there are data implicating BCR signaling in VHR in the EU12 human B cell line, which phenotypically resembles IgM+, CD10+, CD24^{high} cells. In these cells, BCR crosslinking promotes VHR and, conversely, Syk and Src kinase inhibitors inhibit VHR (34). While some of the kinase inhibition experiments could also be influencing mechanisms that operate at earlier stages of B cell development, the BCR crosslinking experiment suggests that BCR signaling could promote VHR in more mature B cells. Furthermore, ligation-mediated PCR experiments documented double-stranded DNA breaks at VH3 cRSS sites in human immature (IgM+, CD27−, CD10+) and mature naïve (IgM+, CD27−, CD10−) circulating B cells, also suggesting that VHR may not be limited to immature B cells (34).

Chronic graft versus host disease (GVHD) is one of the most intriguing examples in which VHR could be occurring in more mature B cells (35). B6 mice injected with I–A incompatible T cells from bm12 mice develop chronic GVHD and produce a spectrum of autoantibodies that resembles those found in systemic autoimmune conditions such as systemic lupus erythematosus (SLE) (36). When anti-dsDNA heavy chain knock-in mice such as 3H9 and 56R are used, GVHD occurs and the production of anti-nuclear antibodies is enhanced (35). But the remarkable finding is that among IgG antibodies, a large fraction does not use the knocked in heavy chain (35). Although this unexpected skewing away from the 56R H chain could be the result of selective pressures on the minority population of H chain edited B cells that emerge from the bone marrow, it is not at all obvious how this selection could operate to disfavor the transgene, and why its effects would be largely confined to IgG and not IgM. It is possible that the transgene was revised (by further gene rearrangement) in the periphery, either because it was inactivated by somatic mutation (37), or because the stimulus afforded by cGVH re-induced the rearrangement machinery. An alternative explanation is that the 56R transgene bearing cells are disfavored during primary B cell maturation because they recognize DNA and this self-reactivity causes them to be anergized (this would predict that 56R+ cells would be over-represented amongst IgM rather than IgG B cells). Consistent with the possibility of anergy, most B cells expressing the IgM allotype of the 56R transgene have low levels of IgM (38–40).

WHAT ARE THE CONSEQUENCES AND POTENTIAL FUNCTIONS OF VH REPLACEMENT?

VH replacement allows a B cell with an inadequate pre-BCR or an autoreactive BCR to swap out the existing heavy chain and replace it with a different heavy chain. But why would this be useful? One possibility is that VHR increases the odds of generating a functional antibody. Producing a functional antibody is rather difficult (41): many rearrangements are out of frame, VH pseudogenes outnumber functional VH genes, many newly generated antibodies are autoreactive (42), some combinations may be sequestered inside the cells (38) and some H and L chain combinations may not pair well with each other. VHR may also facilitate the use of a wider array of upstream VH genes. By giving cells with defective antibody rearrangements a chance at revising those antibodies, perhaps the efficiency of primary B cell generation is greatly improved.

On the other hand, a seemingly diametrically opposed consequence of VHR is the potential generation of multireactive antibodies. VHR can sometimes result in the retention of a "footprint" that is comprised of DNA sequences downstream of the cryptic heptamer of the invaded VH gene (**Figure 1**). Because the cRSS is typically positioned further from the WGXG motif in the JH segment than the 5′RSS of a DH segment, VHR is likely to produce longer CDR3 segments than primary rearrangements. Not surprisingly, longer CDR3 sequences have a higher

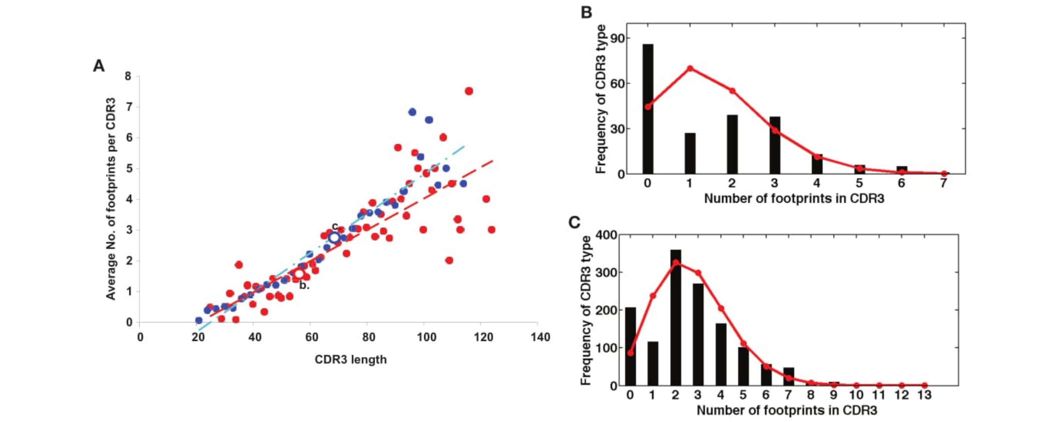


FIGURE 2 | (A) Longer CDR3 sequences have more footprints. Plotted are the average numbers of footprint 5-mers per sequence. Sequences are averaged at each CDR3 size by the number of sequences that have CDR3s of that particular length. Blue dots are in-frame (IF) rearrangements and red dots are out of frame (OF) rearrangements. We find a positive linear correlation between the length of the CDR3 and the number of footprints ($r^2 = 0.89$ for IF, $r^2 = 0.7$ for OF). The line describing this relationship has a slope of 0.06 ± 0.008 for OF (red line) and 0.05 ± 0.008 for IF rearrangements (blue line). **(B,C)** are two examples of the entire distribution of footprint numbers at two positions – position 56 (red circle OF) and position 69 (blue circle IF). Black stems indicate the numbers of footprints and red lines represent the fit of a Poisson distribution ($\lambda = 1.57$ and 2.74, respectively).

proportion of footprints (**Figure 2**), but this does not guarantee that all long CDR3 are the product of VHR. Seventy-eight percent of the potential footprint regions in functional human VH genes contain an arginine codon, so footprint-containing sequences often also harbor a larger number of charged residues. Longer CDR3s have been associated with greater multireactivity, and such multireactive B cells are normally counter-selected as B cells mature during normal B cell development (42). RA patients have antibodies with unusual CDR3 sequences in their synovium (17) and we have seen CDR3 sequences in patients with SLE that have regions of sequence homology that could arise due to VHR. For example, **Figure 1C** shows a rearrangement from an expanded B cell clone in a patient with SLE that appears to contain two footprint sequences (highlighted in yellow). Autoimmune-prone strains of mice have elongated CDR3s, although many of these may arise through mechanisms other than VHR, such as D–D fusion (43, 44). All of these findings beg the question of whether such "multireactivity" serves a useful function. Is multireactivity protective, particularly in the context of an innate immune response? Or could multireactive antibodies be useful in clearing debris that might be inflammatory if left to accumulate? It is intriguing in this regard that some multireactive IgM antibodies such as the famous T15 idiotype, which binds phosphorylcholine (45), also have anti-inflammatory properties (46).

It is possible that there is no simple single answer to the function of VHR, if it has one at all. It would certainly seem that the biological consequences of VHR depend upon the developmental context in which the rearrangement occurs. If replacement occurs centrally, as is likely to occur in wild type strains of mice such as B6 (40, 47), it could serve as a tolerance mechanism (receptor editing) or as means of increasing the efficiency of primary B cell generation. It might also generate a portion of the primary antibody repertoire that has special functional properties such as multireactivity. Conversely, if it occurs peripherally, as might arise in dysregulated states of immune activation such as GVHD (48), perhaps autoimmunity results.

VH REPLACEMENT IN pre-B CELL ALL

Given the abundance of findings linking VHR to pro- or pre-B cell development discussed above, it is not surprising that the initial demonstrations of VHR occurred in transformed pre-B cell lines. More recently, VHR has been demonstrated to be a major contributor to clonal evolution in precursor B cell acute lymphoblastic leukemia (B-ALL) (49, 50). In B-ALL, there is presumably a large clone of cells "frozen" in the pre-B cell stage. The recombinase machinery remains active in at least some of these cells and can drive VHR. It is instructive to review the early work in the murine pre-B cell line NFS5, in which VHR was found to alter not only the productive but also the non-productively rearranged allele (12). Thus assays where one attempts to define a clone based upon its predicted "conservation" of other immunoglobulin gene rearrangements (such as the other H chain allele) within the same cell are not necessarily reliable or easy to interpret. The potential for VHR to contribute to intraclonal diversification is highly relevant to the design and interpretation of assays for minimal residual disease monitoring that employ quantitative PCR with probes or primers for clone-specific junctional sequences (51) or, more recently, high throughput sequencing of heavy chain CDR3 (52). Such studies must take VHR and other forms of intraclonal diversification into account.

Table 1 | Footprint sequences in the 3′ end of human germline VH genes and alleles.

Footprint (5-mer variants)	VH gene allele(s)	
CGAGAGA (CGAGA, GAGAG, AGAGA)	CGAGAGA	VH1-18, VH1-2*1, VH1-2*2, VH1-2*3, VH1-2*5, VH1-3, VH1-46*1, VH1-46*2, VH1-69*1, VH1-69*4, VH1-69*6, VH1-69*8, VH1-69*9, VH1-69*10, VH1-69*11, VH1-69*12, VH1-69*13, VH1/OR15-1*2, VH1/OR15-1*3, VH1/OR15-1*4, VH3-11*1, VH3-11*4, VH3-11*5, VH3-21, VH3-30*1, VH3-30*3, VH3-30*4, VH3-30*5, VH3-30*6, VH3-30*7, VH3-30*9, VH3-30*10, VH3-30*11, VH3-30*12, VH3-30*13, VH3-30*14, VH3-30*15, VH3-30*16, VH3-30*17, VH3-30*18, VH3-30*19, VH3-33*1, VH3-33*2, VH3-33*4, VH3-33*5, VH3-48, VH3-53*1, VH3-53*4, VH3-64*1, VH3-64*2, VH3-64*4, VH3-66*1, VH3-66*3, VH3-7*1, VH3-7*3, VH4-28*3, VH4-30-2*4, VH4-31*1, VH4-31*2, VH4-31*3, VH4-31*10, VH4-34*9, VH4-39*2, VH4-39*6, VH4-39*7, VH4-4*2, VH4-4*6, VH4-4*7, VH4-59*1, VH4-59*2, VH4-61*1, VH4-61*2, VH4-61*3, VH4-61*8, VH4/OR15-8, VH7-4-1*2, VH7-4-1*4, VH7-4-1*5
	CGAGA	VH1-2*4, VH1-69*2, VH1-69*5, VH1/OR15-1*1, VH3-11*3, VH3-30*8, VH3-30-3*1, VH3-53*2, VH3-66*2, VH3-7*2, VH4-28*4, VH4-34*12, VH4-59*7, VH4-61*5, VH4-b, VH5-51*3, VH5-51*4, VH5-a, VH7-4-1*1
CGAGAGG (CGAGA, AGAGG)	VH1-8, VH4-34*1, VH4-34*2, VH4-34*4, VH4-34*5, VH4-34*13, VH4-59*9	
CGAGACA (CGAGA, GAGAC, AGACA)	VH3-66*4, VH4-30-2*3, VH4-39*1, VH4-59*8, VH4-61*7, VH5-51*1, VH5-51*2	
CGAGATA (CGAGA, GAGAT, AGATA)	VH4-34*10, VH4-59*10, VH7-81	
CGAGAAA (CGAGA, GAGAA, AGAAA)	VH4-28*1, VH4-28*2, VH4-28*5, VH4-28*6	
CAAGA*N*A (CAAGA, *AAGAN, AGANA*)	*CAAGANA*	*VH1-45*1*
	CAAGATA	VH1-45*2
	CAAGAGA	VH3-13*1, VH3-13*2, VH3-13*4, VH3-74*1, VH3-74*3, VH3/OR16-10*3, VH6-1
	CAAGA	VH1-45*3, VH3-13*3, VH3-74*2, VH3/OR16-10*1, VH3/OR16-10*2, VH3/OR16-12
CAACAGA	CAACAGA	VH1-24
	CAACA	VH1-f*1
CTAGAGA (CTAGA, TAGAG, AGAGA)	CTAGAGA	VH1-46*3, VH3-72*1, VH3/OR15-7*5
	CTAGA	VH3/OR15-7*1, VH3/OR15-7*2, VH3/OR15-7*3
CTAGGGA (CTAGG, TAGGG, AGGGA)	VH3-53*3	
CGAAAGA (CGAAA, GAAAG, AAAGA)	CGAAAGA	VH3-23*1, VH3-23*2, VH3-23*4, VH3-30*2, VH3-30-3*2, VH3-33*3, VH3-33*6, VH3-NL1
	CGAAA	VH3-23*3, VH3-23*5
	CGNNN	*VH4-30-2*2, VH4-31*4, VH4-34*8, VH4-39*5, VH4-59*3, VH4-59*4, VH4-59*5, VH4-59*6,*
	CG	*VH4-31*5*
CCAGATATA (CCAGA, CAGAT, AGATA, GATAT, ATATA)	VH3-38	
CCAGAGA (CCAGA, CAGAG, AGAGA)	VH4-30-2*1, VH4-30-2*5, VH4-30-4*1, VH4-30-4*2, VH4-30-4*5, VH4-30-4*6, VH4-61*6	
TGAAACA (TGAAA, GAAAC, AAACA)	TGAAA	VH3/OR16-8*1, VH3/OR16-9
	TGAAACA	VH3/OR16-8*2
TGAGA		
TGAGAGA (TGAGA, GAGAG, AGAGA)	TGAGA	VH1/OR15-5
	TGAGAGA	VH1/OR15-9, VH1/OR21-1
TGAGAAA (TGAGA, GAGAA, AGAAA)	VH3-16, VH3-35	
TGAAAGA (TGAAA, GAAAG, AAAGA)	VH3-64*3, VH3-64*5	
CGGCAGA (CGGCA, GGCAG, GCAGA)	VH1-58	
CACGGATAC (CACGG, ACGGA, CGGAT, GGATA, GATAC)	VH2-26, VH2-70*1, VH2-70*10, VH2-70*11	

(Continued)

Table 1 | Continued

Footprint (5-mer variants)	VH gene allele(s)	
CATGGAGAG (CATGG, ATGGA, TGGAG, GGAGA, GAGAG)	VH2/OR16-5	
TACGG	VH2-5*4, VH2-70*9	
TANNN	VH2-5*7	
CACGG	VH2-5*10	
CACACAGACC (CACAC, ACACA, CACAG, ACAGA, CAGAC, AGACC)	CACACAGACC CACACAGAC CACACAGA	VH2-5*1 VH2-5*5, VH2-5*8, VH2-5*9, VH2-70*12 VH2-5*6
CAAAAGATA (CAAAA, AAAAG, AAAGA, AAGAT, AGATA)	VH3-43, VH3-9	

*Two hundred and seventy-three functional VH genes, including alleles and sequences designated as open reading frames, were downloaded from the IMGT database (54) and manually scanned for footprints. A footprint is defined by the nucleotide sequence following the cryptic recombination signal sequence (cRSS), TACTGTG, at the 3′ end of each VH, and is listed in the left column of the table. The footprint 5-mer variants that were used to scan the sequences are included in parentheses. The right column lists the VH genes and alleles. An asterisk in the VH name refers to a specific allele. If all alleles of a particular VH have the same footprint, allele names are omitted. Overlapping footprints are listed for some footprints in the sub-column on the right. Footprints in red font are also found in germline DH and JH gene segments. Sequences with ambiguous nucleotide designations (N-nucleotides) are indicated by italic font. The following VH gene alleles do not have a cRSS: VH1-18*2, VH1-69*3, VH1-69*7, VH1-c, VH1-f*2, VH2-5*2, VH2-5*3, VH2-70*2, VH2-70*3, VH2-70*4, VH2-70*5, VH2-70*6, VH2-70*7, VH2-70*8, VH2-70*13, VH3-15, VH3-20, VH3-25*4, VH3-49, VH3-72*2, VH3-73*1, VH3-73*2, VH3-d, VH3/OR16-13, VH3/OR16-6*2, VH4-30-4*3, VH4-30-4*4, VH4-31*6, VH4-31*7, VH4-31*8, VH4-31*9, VH4-34*3, VH4-34*6, VH4-34*7, VH4-34*11, VH4-39*3, VH4-39*4, VH4-4*1, VH4-4*3, VH4-4*4, VH4-4*5, VH4-61*4, VH5-51*5, VH7-4-1*3.*

ANALYSIS OF VH REPLACEMENT FOOTPRINTS

The most convincing demonstrations of VHR are those in which a precursor–product relationship can be documented. For example, if the precursor VH gene is known and then additional B cells can be found to share most of the 3′ side of the CDR3 (the same DH–JH junction), but have a different VH gene, this can be compelling, as in B-ALL or in mouse models with heavy chain knock-ins. In contrast, the analysis of VHR in a physiologic and fully diversified immune repertoire has by necessity focused on indirect evidence, namely the enumeration of footprints, which are potential traces of previous VDJ rearrangements in IgH sequence data. In mice, footprints are readily observed in constrained immune repertoires [for example, Ref. (13, 15, 23)]. Footprints are also observed in humans (53). However, a fundamental issue with footprint analysis in humans is one of specificity of attribution: does the footprint arise due to VHR or is it due to some other form of junctional diversification or skewing in the rearrangement process? Or does it occur by chance?

To investigate the hypothesis that footprint sequences are due to the process of VHR, we sequenced IgH rearrangements from peripheral blood B cells of a healthy human adult subject, following an IRB-approved protocol. We identified 42,221 unique sequences from this sample, which we analyzed for VH footprints using a sliding window method (see Supplementary Material for further details). All of the potential footprints arising from sequences at the 3′ ends of the germline VH genes are listed in **Table 1**. In accordance with their conventional description in the literature (41), we required the footprint to be least five nucleotides long (we hereafter refer to these sequences as footprint 5-mers). If footprint 5-mers are due to VHR, they will have specific characteristics in antibody repertoire data, indicated by the tests described below.

TEST 1: VH REPLACEMENT FOOTPRINTS SHOULD BE LOCATED IN THE 5′ END OF THE CDR3 SEQUENCE

One way to distinguish bona fide footprints from other sources of sequence variation is to compare the number of footprints in the junction between VH and DH (referred to as N1) to the number in the junction between DH and JH (referred to as N2). Footprints arising via VHR should occur in N1 rather than N2 because the cryptic heptamer is located in N1. However, as shown in **Figure 3A**, there is a roughly bimodal distribution of footprint 5-mers. Even though we excluded the most common footprints that were found in the germline DH gene segments (**Table 2**), there were still plenty of footprint sequences in N1, DH, and N2. In **Figure 3B** we took the analysis one step further and removed some more of the common footprint sequences that are found not only in the germline DH gene segments but also in JH6. This resulted in more skewing toward N1, but a large proportion of the footprints were still outside of N1. In fact, not only were footprint 5-mers found in DH and JH, but they were also found in other parts of the VH gene. **Table 3** lists the positions of all of the footprint 5-mers found amongst the germline VH alleles listed in the IMGT database.

TEST 2: VH REPLACEMENT FOOTPRINTS SHOULD BE MORE FREQUENT IN UPSTREAM VH GENES THAN DOWNSTREAM VH GENES AND ABSENT FROM VH6-1

Another requirement for a footprint to be consistent with VHR is that the invading VH must be upstream of the VH that donated the footprint. Unfortunately, the recipient VH is often difficult to define because many VHs have the same or very similar footprints

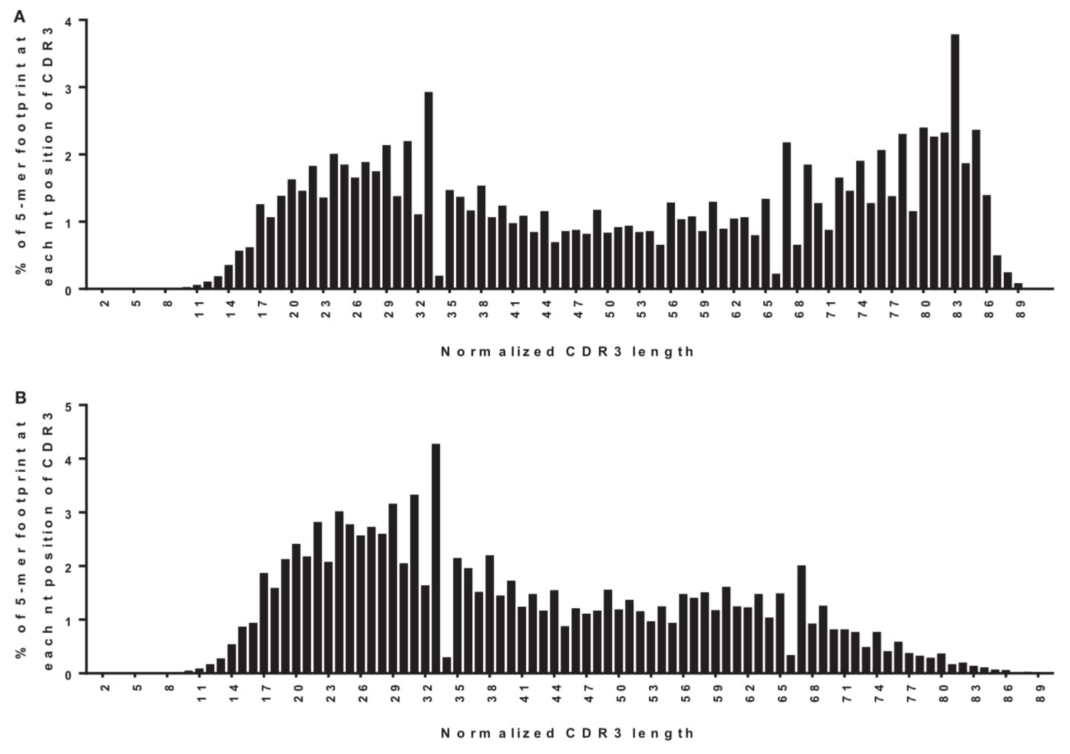


FIGURE 3 | Positions of footprint 5-mers. **(A)** Frequency distribution of all footprint 5-mers out of total unique sequences ($n = 42,221$) plotted against the normalized CDR3 position. The CDR3 is herein defined to begin at the conserved CAR amino acid sequence (TGT GCG AGA nucleotide sequence) within the 3′ end of VH and end at the conserved W (TGG nucleotide sequence) that is immediately upstream of the first conserved glycine, GGC nucleotide sequence) within the JH. "TGGAG" is excluded as it is found in many alleles of all DH genes. The position of the footprint is defined by where the footprint starts within the CDR3. For example, if a footprint occupies nucleotides 12–16 of the CDR3, it will be plotted at position 12. CDR3 lengths were normalized to a scale of 1–100 using $p1 = p/(L \times 100)$; where p is the position of the 5-mer in the real CDR3 sequence, L is the length of the CDR3 sequence, and p1 is the normalized position. Normalized positions are rounded to the nearest integer. **(B)** Frequency distribution of all footprint 5-mers plotted against the normalized CDR3 position, corrected for footprint 5-mers in the germline JH6 gene. "ATGGA," "TACGG," and "CATGG" were excluded when found in the 3′ end of the CDR3 within the JH6 gene, as these 5-mers are found in the germline JH6 sequence. Footprint 5-mers found in other JHs (see **Table 2**) are not counted in either **(A)** or **(B)** because they are all located outside CDR3 region of JH.

(see **Table 1**). However, a more straightforward test of whether a footprint 5-mer represents the product of VHR is to evaluate the frequency of footprints in different VH rearrangements. In particular, the 3′ most VH gene (VH6-1, in humans), when rearranged, should not exhibit VHR footprints as there is no downstream VH that it can invade. Conversely, VH genes that are situated in the 5′ end of the locus should have higher frequencies of footprints than 3′ VH genes, if VHR is frequent. Yet the overall frequency of footprint 5-mers was similar amongst unique sequences in all of the most commonly used VHs, including VH6-1 (**Figure 4**). The frequency of footprints was also not significantly higher in out of frame (unselected) versus in-frame rearrangements (**Figure 4A**).

We also performed this analysis using immunoglobulin analysis tool (IgAT) software (42) and observed that the frequency of footprints was not reduced in VH6-1 when compared to other VHs (**Figure 4B**). Lower VH footprint frequencies were observed overall because footprints in the 3′ end of the CDR3 are excluded by the IgAT program (42). One intriguing feature of the IgAT data was that, unlike our footprint analysis that captured 5-mers at both N1 and N2, when only N1 was analyzed, some VHs, including VH6-1, had higher footprint frequencies than others. Since VH6-1 cannot have any footprints due to VHR, we conclude that many footprint 5-mers that are found in the CDR3 do not arise by VHR.

The simplest explanation is that the great majority of 5-mer sequences found throughout the CDR3 resemble footprints by chance. The frequency of footprint 5-mers in the entire CDR3 was highly correlated with the length of the CDR3 (**Figure 2**). The ability to generate a replacement footprint by chance may be

Table 2 | Footprint sequences in DH and JH alleles.

DH gene	Sequence (footprint(s) in red font)
D1-1*01	GGTACAACTGGAACGAC
D1-14*01	GGTATAACCGGAACCAC
D1-20*01	GGTATAACTGGAACGAC
D1-26*01	GGTATAGTGGGAGCTACTAC
D1-7*01	GGTATAACTGGAACTAC
D2-15*01	AGGATATTGTAGTGGTGGTAGCTGCTACTCC
D2-2*01	AGGATATTGTAGTAGTACCAGCTGCTATGCC
D2-2*02	AGGATATTGTAGTAGTACCAGCTGCTATACC
D2-2*03	TGGATATTGTAGTAGTACCAGCTGCTATGCC
D2-21*01	AGCATATTGTGGTGGTGATTGCTATTCC
D2-21*02	AGCATATTGTGGTGGTGACTGCTATTCC
D2-8*01	AGGATATTGTACTAATGGTGTATGCTATACC
D2-8*02	AGGATATTGTACTGGTGGTGTATGCTATACC
D3-10*01	GTATTACTATGGTTCGGGGAGTTATTATAAC
D3-10*02	GTATTACTATGTTCGGGGAGTTATTATAAC
D3-16*01	GTATTATGATTACGTTTGGGGGAGTTATGCTTATACC
D3-16*02	GTATTATGATTACGTTTGGGGGAGTTATCGTTATACC
D3-22*01	GTATTACTATGATAGTAGTGGTTATTACTAC
D3-3*01	GTATTACGATTTTTGGAGTGGTTATTATACC
D3-3*02	GTATTAGCATTTTTGGAGTGGTTATTATACC
D3-9*01	GTATTACGATATTTTGACTGGTTATTATAAC
D4-11*01	TGACTACAGTAACTAC
D4-17*01	TGACTACGGTGACTAC
D4-23*01	TGACTACGGTGGTAACTCC
D4-4*01	TGACTACAGTAACTAC
D5-12*01	GTGGATATAGTGGCTACGATTAC
D5-18*01	GTGGATACAGCTATGGTTAC
D5-24*01	GTAGAGATGGCTACAATTAC
D5-5*01	GTGGATACAGCTATGGTTAC
D6-13*01	GGGTATAGCAGCAGCTGGTAC
D6-19*01	GGGTATAGCAGTGGCTGGTAC
D6-25*01	GGGTATAGCAGCGGCTAC
D6-6*01	GAGTATAGCAGCTCGTCC
D7-27*01	CTAACTGGGGA

JH gene	Sequence (footprint(s) in red font)
J1*01	GCTGAATACTTCCAGCACTGGGGCCAGGGCACCCTGGTCACCGTCTCCTCAG
J2*01	CTACTGGTACTTCGATCTCTGGGGCCGTGGCACCCTGGTCACTGTCTCCTCAG
J3*01	TGATGCTTTTGATGTCTGGGGCCAAGGGACAATGGTCACCGTCTCTTCAG
J3*02	TGATGCTTTTGATATCTGGGGCCAAGGGACAATGGTCACCGTCTCTTCAG
J4*01	ACTACTTTGACTACTGGGGCCAAGGAACCCTGGTCACCGTCTCCTCAG
J4*02	ACTACTTTGACTACTGGGGCCAGGGAACCCTGGTCACCGTCTCCTCAG
J4*03	GCTACTTTGACTACTGGGGCCAAGGGACCCTGGTCACCGTCTCCTCAG
J5*01	ACAACTGGTTCGACTCCTGGGGCCAAGGAACCCTGGTCACCGTCTCCTCAG
J5*02	ACAACTGGTTCGACCCCTGGGGCCAGGGAACCCTGGTCACCGTCTCCTCAG
J6*01	ATTACTACTACTACTACGGTATGGACGTCTGGGGGCAAGGGACCACGGTCACCGTCTCCTCAG
J6*02	ATTACTACTACTACTACGGTATGGACGTCTGGGGCCAAGGGACCACGGTCACCGTCTCCTCAN
J6*03	ATTACTACTACTACTACTACATGGACGTCTGGGGCAAAGGGACCACGGTCACCGTCTCCTCAN
J6*04	ATTACTACTACTACTACGGTATGGACGTCTGGGGCAAAGGGACCACGGTCACCGTCTCCTCAG

(Continued)

Thirty-four germline functional human DH alleles and 13 JH alleles were downloaded from the IMGT database (54). Each allele was scanned for all of the possible five nucleotide footprint motifs listed in ***Table 1****. Sequences containing five nucleotide footprints are given in red font. In some cases the region matches more than one possible footprint (for example, D5-12*01 contains three different footprints: GGATA, GATAT, and ATATA).*

under-appreciated. In a completely random DNA sequence with equal proportions of A, T, G, and C bases, the chance of finding a specific 5-mer sequence is 1/1,024 (or ~0.001). However, there are at least 50 different footprint-derived 5-mer sequences amongst human VH genes (**Table 3**), increasing the odds to 50/1,024 (~5%). But this calculation ignores the number of different positions along the VDJ rearrangement where the footprint might be detected and on how many variants of the footprint are permitted. If the 5′ end of a CDR3 sequence is 30 nucleotides long, that means that there are 6 completely non-overlapping sequences that have a length of five nucleotides, bringing the minimum likelihood of detection of at least a single footprint in that sequence up to 26% $[1-(1-0.05)^6]$ or 1 − Pr(not getting any 5-mers in the 30 bp sequence). If the base composition of the DNA is non-uniform or the entire CDR3 sequence is surveyed or if sequences with mutations are permitted (for example those matching in 4 out of 5 bp), the chances of detecting a footprint increase even further.

We also wondered why some VH genes had higher footprint frequencies in N1 than others (**Figure 4B**), as this finding is not similar to what one would expect by random chance. We wondered if the real VHR events were hiding somewhere in a large pile of non-VHR footprints. A high "false positive" rate of footprint 5-mers could come about because of sequencing errors. Alternatively or in addition, it may be easy to create false VHR 5-mer sequences in primary VDJ rearrangements through a combination of N-addition, nibbling (or sequencing deletion) and the 3′ sequence of the VH. For example AAAGA could become AAGA or AGA.

It may be worthwhile to develop a better computational approach for detecting VHR footprints with greater specificity for VHR. The IgAT software already eliminated footprints that match the germline VH sequence exactly, but this is insufficient, give the

Table 3 | The number of footprints found in various regions of human VH genes.

Footprint	Sequences	FR	CDR	FR1	FR2	FR3	CDR1	CDR2	CDR3
TGAGA	131	212	12	94		118	3	3	6
CTAGAGA	8		8						8
CGAAAGA	9		9						9
CTAGA	28	17	16		17		1	1	14
TACGG	9	5	4	1		4	1	1	2
CAAAA	45	41	13			41		9	4
CTAGGGA	2	1	1		1				1
CAACA	55	33	27	1	5	27	2	23	2
CAAGA	150	149	18		4	145		3	15
CACACAGA	20	20	6	20					6
CGAGACA	7		7						7
CGAGAAA	3		3						3
TGAAAGA	2		2						2
TGAAACA	3		3					2	1
CGAGAGG	8	2	6		2				6
CGAGAGA	86		86						86
CACGG	179	178	6			178			6
CGAGATA	3		3						3
CGAAA	11		11					1	10
TGAAA	61	79	7	33		46	1	3	3
CACACAGAC	20	20	5	20					5
CGAGA	130	3	127	1	2				127
CACACAGACC	20	20	1	20					1
TGAGAAA	6		6					3	3
CAAAAGATA	3		3						3
CGGCAGA	2		2						2
TGAGAGA	3	1	2			1			2
CCAGATATA	2		2						2
CAAGATA	1		1						1
CCAGAGA	84	80	4			80			4
CACGGATAC	5		5						5
CAAGAGA	27	19	8			19			8

Two hundred and thirty-four functional germline human VH alleles were downloaded from the IMGT database (54). For each footprint, the number of times that footprint was found in the VH alleles was recorded. The first column lists the footprint, the second column lists the number of alleles in which the footprint is found. Some alleles contain more than one footprint. The remaining columns indicate how many footprints were found in each of the corresponding regions of the V region. FR, framework; CDR, complementarity determining region. The columns named FR and CDR provide the total number of times that a particular footprint is found in the FR or the CDR.

high frequency of footprints in N1 of VH6-1. Further specificity might be achievable if one were to limit the detection of footprint 5-mers to sequences that are *unlikely* to arise through a single nucleotide change (for example, those that arise by deletion that converts a non-matching 6-mer to a matching 5-mer, or mutation of a non-matching 5-mer to a footprint 5-mer or N-addition of a nucleotide adjoining a 4-mer to create a footprint 5-mer). An alternative approach is to require that there be two footprint-like sequences in tandem. Either or both of these methods might increase specificity, but could also reduce sensitivity of footprint detection. The validity of either approach would need to be tested further using validated data sets in which VHR events are known to have or have not occurred.

We also considered the possibility that footprint 5-mers may frequently arise through some mechanism other than VHR. We considered two potential alternative mechanisms – (1) microhomology-mediated joining and (2) cleavage, nibbling, and rejoining at the cryptic heptamer.

ALTERNATIVE THEORY 1: MICROHOMOLOGY-MEDIATED JOINING

We considered the possibility that footprints at N1 were arising primarily due to microhomology-mediated joining of similar sequences between the VH and the DH segments. If microhomology-mediated joining were common, one might expect that VHs that share the same 5-mers with DHs are more likely to rearrange, but as shown in **Figure 5**, this is probably not usually the case. DH5-12 (open bars in **Figure 5**), which has three footprint 5-mers, does not appear to be used more frequently in rearrangements involving VHs that contain the same 5-mers such as VH2-26 (red arrow). Rather than being

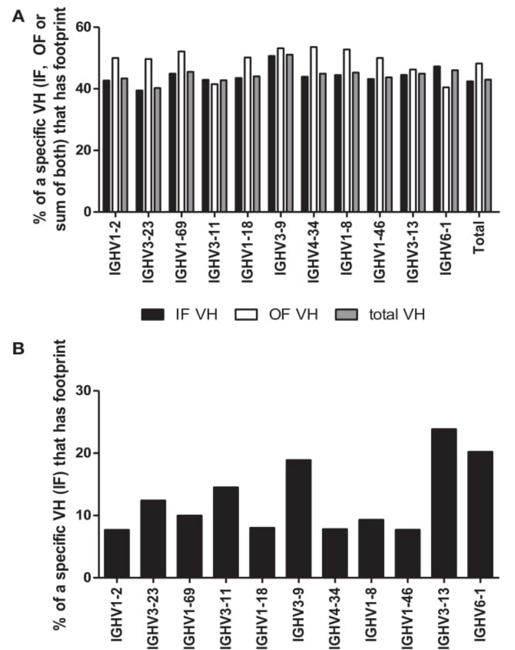


FIGURE 4 | (A) Percentage of rearrangements with footprints for the 10 most frequent VH rearrangements and VH6-1. The percentages of rearrangements that contain at least one footprint 5-mer in either (or both) the 5′ end or the 3′ end of the rearrangement are shown for unique rearrangements for each of the 10 most common VH genes and for VH6-1. CDR3 sequences are defined as described in the legend to **Figure 3** and normalized to an arbitrary length of 100. The 5′ end (which almost always contains N1) is defined as the first 20% of the sequence and the 3′ end is defined as the last 20% of the sequence (which almost always contains N2). VH6-1 is the most 3′ VH gene and cannot contain footprints that are due to VH replacement. Black bars denote unique rearrangements that are in-frame (IF VH), white bars denote out of frame rearrangements (OF VH), and gray bars indicate total unique rearrangements (total VH). **(B)** Footprint Frequency of in-Frame VH rearrangements in N1 using IgAT Software. The same unique in-frame (IF) VH rearrangements, as shown in **(A)**, were analyzed for the presence of one or more footprints using IgAT software (2). Plotted are the frequencies of unique IF VH rearrangements that have one or more footprints in the N1 region.

skewed toward particular VHs with matching or similar footprints, the frequency of rearrangements of different VHs to DH5-12 rearrangements resembled overall VH usage (**Figure 5**, closed bars). While this analysis is very preliminary and only focused on a single DH, it suggests that microhomology-mediated joining, based upon shared sequences between VH and DH, is not a frequent mechanism for generating footprint 5-mers.

ALTERNATIVE THEORY 2: CLEAVAGE, NIBBLING, AND REJOINING AT THE CRYPTIC HEPTAMER IN VH

We wondered if there could be cleavage at the cryptic heptamer, followed by exonucleolytic nibbling and re-sealing at the site of the break, without full-blown rearrangement (Figure S1 in Supplementary Material illustrates this idea for a VH6-1 rearrangement). Note that this type of rearrangement product would not involve VHR, but would have the result of diversifying the 3′ end of the VH in the primary rearrangement product, altering the primary amino acid sequence and/or the reading frame of the rearrangement. This hypothesis makes predictions regarding the sequence characteristics that would be more or less amenable to this type of atypical open-and-shut joint (2). For example, one would expect that if most footprint 5-mers at N1 arise by this mechanism, that the frequency of footprint 5-mers would be very low in VH genes that lack cryptic heptamers. Furthermore, one would expect that the 5′ footprint 5-mer seen in most rearrangements would resemble the 3′ end of the germline sequence of the same VH gene that is present in the rearrangement.

PRELIMINARY CONCLUSIONS AND CAVEATS FROM FOOTPRINT ANALYSIS

Taken together, these data suggest that many if not most footprint sequences arise by some mechanism(s) other than VHR. But there are some caveats to this analysis. First, these data were only obtained on one healthy adult. It is possible that footprints may differ in other individuals or in a minority of individuals. In addition, different findings might occur in individuals with immunologic disorders such as SLE or neoplastic conditions such as B-ALL. Furthermore, only B cells from the peripheral blood were analyzed. It is conceivable that B cell populations with extensive VHR reside elsewhere in the body, particularly within the bone marrow. Finally, as discussed above, it is possible that some of the VHR footprints that were identified are due to sequencing errors. We tried to protect against this artifact by selectively analyzing unique sequences that were present in at least two copies. But even with this precaution, there are still likely to be many sequencing errors.

CONCLUSION

VH replacement exchanges the VH within a pre-existing VDJ rearrangement with an upstream VH gene, while preserving most of the original CDR3 sequence. It also sometimes results in the retention of a footprint sequence in the VH gene that was invaded. The result of VHR is an alteration in the specificity or functional status of the antibody. But the mechanistic consequence of that alteration is unclear. Is it to diversify the repertoire once a good CDR3 sequence has been found? Or is it to reduce autoreactivity or generate some form of protective multireactivity? Or is it simply a means by which B cells with non-productive rearrangements on one or both alleles have another shot at creating a productive rearrangement? In humans, the analysis of VHR is confounded by not having a means of definitively identifying the precursor rearrangement. Rather, the analysis of VHR in humans is accomplished indirectly through footprint analysis, but as demonstrated herein, footprints may arise for reasons other than VHR. Thus, while VHR certainly occurs, footprint analysis is a poor measure of its frequency because of the high rate of false positives and an unknown rate of false negatives. Nevertheless, it is possible that footprints may provide other insights into the mechanisms of V(D)J recombination and its potentially aberrant regulation in disease states. With the advent of high throughput sequencing

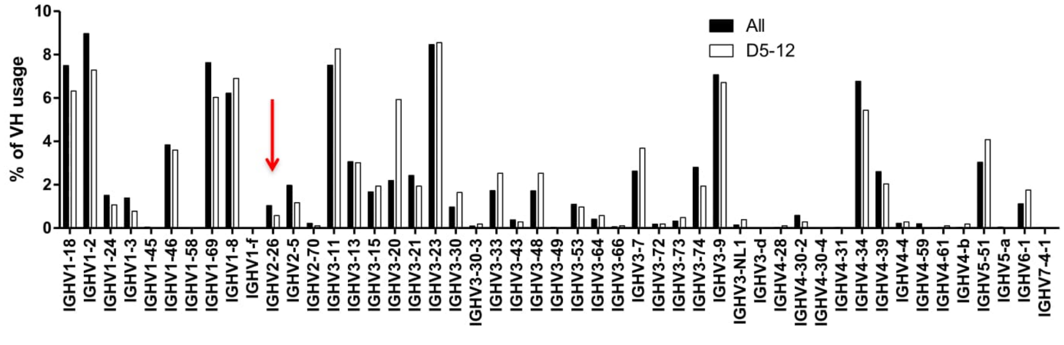


FIGURE 5 | VH usage is similar in all unique rearrangements and in rearrangements with DH5-12. VH usage is shown for total unique sequences (closed bars, $n = 42{,}221$) and for those with DH5-12 (open bars, $n = 1{,}029$). Sequences containing the DH5-12 gene segment in the CDR3 region were recognized using IMGT high V quest analysis, and further analysis on VH usage was performed in-house (see text). The red arrow points to VH2-26, which contains some of the same footprint 5-mers as DH5-12, but does not appear to be used more frequently in rearrangements that include DH5-12.

studies, further analysis of IgH gene rearrangements for VHR and other mechanisms of CDR3 diversification promise to be illuminating.

ACKNOWLEDGMENTS

The authors acknowledge the Arthritis Foundation (Wenzhao Meng), the Lupus Research Institute (Eline T. Luning Prak), the Hematology Translational Center of Excellence at the Perelman School of Medicine of the University of Pennsylvania, NIH T32 CA009615 (Alfred L. Garfall), the Institute for Translational Medicine and Therapeutics at the University of Pennsylvania (Alfred L. Garfall) and NIH R25-DK066028 (Sahana Jayaraman) for funding this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fimmu.2014.00010/abstract

Figure S1 | Model for atypical open-and-shut joints. Footprint 5-mers can be generated by cleaving at the cryptic heptamer, followed by preferential trimming by exonucleolytic nibbling at the 3′ end of the double strand break. Shown is the generation of a footprint sequence in a VH6-1 rearrangement. The cryptic heptamer is indicated by a dashed triangle, colored squares indicate the VH, DH, and JH gene segments (not drawn to scale). According to this model, there is cleavage at the cryptic heptamer, followed by nibbling at the 3′ end of the break (red wavy line), leading to selective loss of the A residue. The third line of the figure shows the repaired rearrangement, in which the A residue is missing from the rearrangement.

REFERENCES

1. Jung D, Giallourakis C, Mostoslavsky R, Alt FW. Mechanism and control of V(D)J recombination at the immunoglobulin heavy chain locus. *Annu Rev Immunol* (2006) **24**:541–70. doi:10.1146/annurev.immunol.23.021704.115830
2. Gellert M. V(D)J recombination: RAG proteins, repair factors, and regulation. *Annu Rev Biochem* (2002) **71**:101–32. doi:10.1146/annurev.biochem.71.090501.150203
3. Vinocur JM, Fesnak AD, Liu Y, Charan D, Luning Prak ET. Violations of the 12/23 rule at the mouse immunoglobulin kappa locus, including V kappa-V kappa rearrangement. *Mol Immunol* (2009) **46**:2183–9. doi:10.1016/j.molimm.2009.04.021
4. Bassing CH, Alt FW, Hughes MM, D'Auteuil M, Wehrly TD, Woodman BB, et al. Recombination signal sequences restrict chromosomal V(D)J recombination beyond the 12/23 rule. *Nature* (2000) **405**:583–6. doi:10.1038/35014635
5. Briney BS, Willis JR, Hicar MD, Thomas JW II, Crowe JE Jr. Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire. *Immunology* (2012) **137**:56–64. doi:10.1111/j.1365-2567.2012.03605.x
6. Luning Prak ET, Monestier M, Eisenberg RA. B cell receptor editing in tolerance and autoimmunity. *Ann N Y Acad Sci* (2011) **1217**:96–121. doi:10.1111/j.1749-6632.2010.05877.x
7. 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**:553–63. doi:10.1016/S0092-8674(00)00078-7
8. Reynaud CA, Garcia C, Hein WR, Weill JC. Hypermutation generating the sheep immunoglobulin repertoire is an antigen-independent process. *Cell* (1995) **80**:115–25. doi:10.1016/0092-8674(95)90456-5
9. Darlow JM, Stott DI. Gene conversion in human rearranged immunoglobulin genes. *Immunogenetics* (2006) **58**:511–22. doi:10.1007/s00251-006-0113-6
10. Duvvuri B, Wu GE. Gene conversion-like events in the diversification of human rearranged IGHV3-23*01 gene sequences. *Front Immunol* (2012) **3**:158. doi:10.3389/fimmu.2012.00158
11. Darlow JM, Stott DI. V(H) replacement in rearranged immunoglobulin genes. *Immunology* (2005) **114**:155–65. doi:10.1111/j.1365-2567.2004.02084.x
12. Kleinfield RW, Weigert MG. Analysis of VH gene replacement events in a B cell lymphoma. *J Immunol* (1989) **142**:4475–82.
13. Chen C, Nagy Z, Luning Park ET, Weigert M. Immunoglobulin heavy chain gene replacement: a mechanism of receptor editing. *Immunity* (1995) **3**:747–55. doi:10.1016/1074-7613(95)90064-0
14. Golub R, Martin D, Bertrand FE, Cascalho M, Wabl M, Wu GE. VH gene replacement in thymocytes. *J Immunol* (2001) **166**:855–60.
15. Taki S, Schwenk F, Rajewsky K. Rearrangement of upstream DH and VH genes to a rearranged immunoglobulin variable region gene inserted into the DQ52-JH region of the immunoglobulin heavy chain locus. *Eur J Immunol* (1995) **25**:1888–96.
16. Wilson PC, Wilson K, Liu YJ, Banchereau J, Pascual V, Capra JD. Receptor revision of immunoglobulin heavy chain variable region genes in normal human B lymphocytes. *J Exp Med* (2000) **191**:1881–94. doi:10.1084/jem.191.11.1881
17. Itoh K, Meffre E, Albesiano E, Farber A, Dines D, Stein P, et al. Immunoglobulin heavy chain variable region gene replacement As a mechanism for receptor revision in rheumatoid arthritis synovial tissue B lymphocytes. *J Exp Med* (2000) **192**:1151–64. doi:10.1084/jem.192.8.1151
18. Lenze D, Greiner A, Knörr C, Anagnostopoulos I, Stein H, Hummel M. Receptor revision of immunoglobulin heavy chain genes in human MALT lymphomas. *Mol Pathol* (2003) **56**:249–55. doi:10.1136/mp.56.5.249

19. Ouled-Haddou H, Ghamlouch H, Regnier A, Trudel S, Herent D, Lefranc M, et al. Characterization of a new V gene replacement in the absence of activation-induced cytidine deaminase and its contribution to human BCR diversity. *Immunology* (2014) **141**:268–75. doi:10.1111/imm.12192
20. Zhang Z, Wang YH, Zemlin M, Findley HW, Bridges SL, Burrows PD, et al. Molecular mechanism of serial VH gene replacement. *Ann N Y Acad Sci* (2003) **987**:270–3. doi:10.1111/j.1749-6632.2003.tb06060.x
21. Kowal C, Weinstein A, Diamond B. Molecular mimicry between bacterial and self antigen in a patient with systemic lupus erythematosus. *Eur J Immunol* (1999) **29**:1901–11. doi:10.1002/(SICI)1521-4141(199906)29:06<1901::AID-IMMU1901>3.0.CO;2-L
22. Reth M, Gehrmann P, Petrac E, Wiese P. A novel VH to VHDJH joining mechanism in heavy-chain-negative (null) pre-B cells results in heavy-chain production. *Nature* (1986) **322**:840–2. doi:10.1038/322840a0
23. Cascalho M, Wong J, Wabl M. VH gene replacement in hyperselected B cells of the quasimonoclonal mouse. *J Immunol* (1997) **159**:5795–801.
24. Lutz J, Muller W, Jack HM. VH replacement rescues progenitor B cells with two nonproductive VDJ alleles. *J Immunol* (2006) **177**:7007–14.
25. Rahman NS, Godderz LJ, Stray SJ, Capra JD, Rodgers KK. DNA cleavage of a cryptic recombination signal sequence by RAG1 and RAG2. Implications for partial V(H) gene replacement. *J Biol Chem* (2006) **281**:12370–80. doi:10.1074/jbc.M507906200
26. Koralov SB, Novobrantseva TI, Konigsmann J, Ehlich A, Rajewsky K. Antibody repertoires generated by VH replacement and direct VH to JH joining. *Immunity* (2006) **25**:43–53. doi:10.1016/j.immuni.2006.04.016
27. Liu Y, Li L, Kumar KR, Xie C, Lightfoot S, Zhou XJ, et al. Lupus susceptibility genes may breach tolerance to DNA by impairing receptor editing of nuclear antigen-reactive B cells. *J Immunol* (2007) **179**:1340–52.
28. Park YH, Osmond DG. Dynamics of early B lymphocyte precursor cells in mouse bone marrow: proliferation of cells containing terminal deoxynucleotidyl transferase. *Eur J Immunol* (1989) **19**:2139–44. doi:10.1002/eji.1830191125
29. Davila M, Liu F, Cowell LG, Lieberman AE, Heikamp E, Patel A, et al. Multiple, conserved cryptic recombination signals in VH gene segments: detection of cleavage products only in pro B cells. *J Exp Med* (2007) **204**:3195–208. doi:10.1084/jem.20071224
30. Vettermann C, Herrmann K, Jack HM. Powered by pairing: the surrogate light chain amplifies immunoglobulin heavy chain signaling and pre-selects the antibody repertoire. *Semin Immunol* (2006) **18**:44–55. doi:10.1016/j.smim.2006.01.001
31. Keenan RA, De Riva A, Corleis B, Hepburn L, Licence S, Winkler TH, et al. Censoring of autoreactive B cell development by the pre-B cell receptor. *Science* (2008) **321**:696–9. doi:10.1126/science.1157533
32. Ota M, Duong BH, Torkamani A, Doyle CM, Gavin AL, Ota T, et al. Regulation of the B cell receptor repertoire and self-reactivity by BAFF. *J Immunol* (2010) **185**:4128–36. doi:10.4049/jimmunol.1002176
33. Hondowicz BD, Alexander ST, Quinn WJ III, Pagán AJ, Metzgar MH, Cancro MP, et al. The role of BLyS/BLyS receptors in anti-chromatin B cell regulation. *Int Immunol* (2007) **19**:465–75. doi:10.1093/intimm/dxm011
34. Liu J, Lange MD, Hong SY, Xie W, Xu K, Huang L, et al. Regulation of VH replacement by B cell receptor-mediated signaling in human immature B cells. *J Immunol* (2013) **190**:5559–66. doi:10.4049/jimmunol.1102503
35. Sekiguchi DR, Eisenberg RA, Weigert M. Secondary heavy chain rearrangement: a mechanism for generating anti-double-stranded DNA B cells. *J Exp Med* (2003) **197**:27–39. doi:10.1084/jem.20020737
36. Eisenberg R. The chronic graft-versus-host model of systemic autoimmunity. *Curr Dir Autoimmun* (2003) **6**:228–44. doi:10.1159/000066864
37. Brard F, Shannon M, Luning Prak ET, Litwin S, Weigert M. Somatic mutation and light chain rearrangement generate autoimmunity in anti-single-stranded DNA transgenic MRL/lpr mice. *J Exp Med* (1999) **190**:691–704. doi:10.1084/jem.190.5.691
38. Khan SN, Witsch EJ, Goodman NG, Panigrahi AK, Chen C, Jiang Y, et al. Editing and escape from editing in anti-DNA B cells. *Proc Natl Acad Sci U S A* (2008) **105**:3861–6. doi:10.1073/pnas.0800025105
39. Sekiguchi DR, Yunk L, Gary D, Charan D, Srivastava B, Allman D, et al. Development and selection of edited B cells in B6.56R mice. *J Immunol* (2006) **176**:6879–87.
40. Yunk L, Meng W, Cohen PL, Eisenberg RA, Luning Prak ET. Antibodies in a heavy chain knock-in mouse exhibit characteristics of early heavy chain rearrangement. *J Immunol* (2009) **183**:452–61. doi:10.4049/jimmunol.0804060
41. Coleclough C, Perry RP, Karjalainen K, Weigert M. Aberrant rearrangements contribute significantly to the allelic exclusion of immunoglobulin gene expression. *Nature* (1981) **290**:372–8. doi:10.1038/290372a0
42. Wardemann H, Yurasov S, Schaefer A, Young JW, Meffre E, Nussenzweig MC. Predominant autoantibody production by early human B cell precursors. *Science* (2003) **301**:1374–7. doi:10.1126/science.1086907
43. Klonowski KD, Primiano LL, Monestier M. Atypical VH-D-JH rearrangements in newborn autoimmune MRL mice. *J Immunol* (1999) **162**:1566–72.
44. Klonowski KD, Monestier M. Heavy chain revision in MRL mice: a potential mechanism for the development of autoreactive B cell precursors. *J Immunol* (2000) **165**:4487–93.
45. Chen C, Bruderer U, Rittenberg MB. The developmental patterns of B cell precursors distinguishing between environmental and nonenvironmental forms of phosphocholine. *Cell Immunol* (1992) **143**:378–88. doi:10.1016/0008-8749(92)90034-M
46. Binder CJ, Hörkkö S, Dewan A, Chang MK, Kieu EP, Goodyear CS, et al. Pneumococcal vaccination decreases atherosclerotic lesion formation: molecular mimicry between *Streptococcus pneumoniae* and oxidized LDL. *Nat Med* (2003) **9**:736–43. doi:10.1038/nm876
47. Watson LC, Moffatt-Blue CS, McDonald RZ, Kompfner E, Ait-Azzouzene D, Nemazee D, et al. Paucity of V-D-D-J rearrangements and VH replacement events in lupus prone and non-autoimmune TdT−/− and TdT+/+ mice. *J Immunol* (2006) **177**:1120–8.
48. Eisenberg RA. Secondary receptor editing in the generation of autoimmunity. *Autoimmun Rev* (2012) **11**:787–9. doi:10.1016/j.autrev.2012.02.004
49. Choi Y, Greenberg SJ, Du TL, Ward PM, Overturf PM, Brecher ML, et al. Clonal evolution in B-lineage acute lymphoblastic leukemia by contemporaneous VH-VH gene replacements and VH-DJH gene rearrangements. *Blood* (1996) **87**:2506–12.
50. Gawad C, Pepin F, Carlton VE, Klinger M, Logan AC, Miklos DB, et al. Massive evolution of the immunoglobulin heavy chain locus in children with B precursor acute lymphoblastic leukemia. *Blood* (2012) **120**:4407–17. doi:10.1182/blood-2012-05-429811
51. van derVelden VH, Hochhaus A, Cazzaniga G, Szczepanski T, Gabert J, van Dongen JJ. Detection of minimal residual disease in hematologic malignancies by real-time quantitative PCR: principles, approaches, and laboratory aspects. *Leukemia* (2003) **17**:1013–34. doi:10.1038/sj.leu.2402922
52. Faham M, Zheng J, Moorhead M, Carlton VE, Stow P, Coustan-Smith E, et al. Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia. *Blood* (2012) **120**:5173–80. doi:10.1182/blood-2012-07-444042
53. Zhang Z, Zemlin M, Wang YH, Munfus D, Huye LE, Findley HW, et al. Contribution of Vh gene replacement to the primary B cell repertoire. *Immunity* (2003) **19**:21–31. doi:10.1016/S1074-7613(03)00170-5
54. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, et al. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res* (2009) **37**:D1006–12. doi:10.1093/nar/gkn838

Conflict of Interest Statement: Dr. Christopher S. Carlson is affiliated with Adaptive Biotechnologies, a for-profit high throughput sequencing company. The high throughput sequencing data that were analyzed for this paper were, however, generated at the University of Pennsylvania, not at Adaptive Biotechnologies. The other co-authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 October 2013; accepted: 07 January 2014; published online: 30 January 2014.

Citation: Meng W, Jayaraman S, Zhang B, Schwartz GW, Daber RD, Hershberg U, Garfall AL, Carlson CS and Luning Prak ET (2014) Trials and tribulations with VH replacement. Front. Immunol. 5:10. doi: 10.3389/fimmu.2014.00010

This article was submitted to B Cell Biology, a section of the journal Frontiers in Immunology.

Article 3. Persistence and selection of an expanded B-cell clone in the setting of rituximab therapy for Sjögren's syndrome

Hershberg *et al. Arthritis Research & Therapy* 2014, **16**:R51
http://arthritis-research.com/content/16/1/R51

arthritis research&therapy

RESEARCH ARTICLE **Open Access**

Persistence and selection of an expanded B-cell clone in the setting of rituximab therapy for Sjögren's syndrome

Uri Hershberg[1,2], Wenzhao Meng[3], Bochao Zhang[1], Nancy Haff[3], E William St. Clair[4], Philip L Cohen[5], Patrice D McNair[4], Ling Li[6], Marc C Levesque[6] and Eline T Luning Prak[3*]

Abstract

Introduction: Subjects with primary Sjögren's syndrome (SjS) have an increased risk of developing B-cell lymphoma and may harbor monoclonal B-cell expansions in the peripheral blood. Expanded B-cell clones could be pathogenic, and their persistence could exacerbate disease or predispose toward the development of lymphoma. Therapy with anti-CD20 (rituximab) has the potential to eliminate expanded B-cell clones and thereby potentially ameliorate disease. This study was undertaken to identify and track expanded B-cell clones in the blood of subjects with primary SjS who were treated with rituximab.

Methods: To determine whether circulating B-cell clones in subjects with primary SjS emerge or remain after B cell-depleting therapy with rituximab, we studied the antibody heavy-chain repertoire. We performed single-memory B-cell and plasmablast sorting and antibody heavy-chain sequencing in six rituximab-treated SjS subjects over the course of a 1-year follow-up period.

Results: Expanded B-cell clones were identified in four out of the six rituximab-treated SjS subjects, based upon the independent amplification of sequences with identical or highly similar VH, DH, and JH gene segments. We identified one SjS subject with a large expanded B-cell clone that was present prior to therapy and persisted after therapy. Somatic mutations in the clone were numerous but did not increase in frequency over the course of the 1-year follow-up, suggesting that the clone had been present for a long period of time. Intriguingly, a majority of the somatic mutations in the clone were silent, suggesting that the clone was under chronic negative selection.

Conclusions: For some subjects with primary SjS, these data show that (a) expanded B-cell clones are readily identified in the peripheral blood, (b) some clones are not eliminated by rituximab, and (c) persistent clones may be under chronic negative selection or may not be antigen-driven. The analysis of sequence variation among members of an expanded clone may provide a novel means of measuring the chronicity and selection of expanded B-cell populations in humans.

Introduction

It has been estimated that up to 3 million adults in the US suffer from primary Sjögren's syndrome (SjS) [1]. Primary SjS is an autoimmune disorder characterized by chronic inflammation of the salivary and lacrimal glands and the presence of antinuclear antibodies, most often of the anti-SSA(Ro) and anti-SSB (La) specificities. Patients are often middle-aged females who present with sicca symptoms, such as dry eyes and dry mouth, fatigue, and joint pain, as well as other extraglandular manifestations, including lung disease and neuropathy. In primary SjS, it is believed that both T cells and B cells contribute to disease pathogenesis. Both cell types infiltrate the salivary and other exocrine glands and show evidence of clonal expansion in the affected tissues as well as the circulation [2]. Notably, there is an increased risk of lymphoma in patients with primary SjS [3].

Why patients with primary SjS are at increased risk for lymphoma is unclear and has been the subject of several

* Correspondence: luning@mail.med.upenn.edu
[3]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, 405B Stellar Chance Labs, 422 Curie Boulevard, Philadelphia, PA 19104, USA
Full list of author information is available at the end of the article

studies (reviewed in [4]). One theory is that B-cell hyperactivity in primary SjS results in the abnormal activation of autoreactive B-cells and contributes to their clonal expansion [5]. Autoreactive B-cell clones, such as the recently described CD21^{dim} population in SjS, may remain chronically activated instead of anergic in the presence of self-antigens [6]. Autoreactive B-cell clones may also have increased resistance to apoptosis in primary SjS by virtue of elevated levels of the B-cell survival factor, BAFF (B-cell activating factor) [7]. Another theory is that B cells in primary SjS accumulate and persist due to abnormal or inadequate regulation by other cells of the immune system. Borrowing an example from the field of tumor immunology, B cells that are transformed by Epstein-Barr virus are efficiently killed by T cells and natural killer cells [8]. T cells in SjS not only may be derelict in their duties to constrain or kill transformed B cells but may have joined forces with the enemy: SjS T cells are the predominant inflammatory cell population in the exocrine gland lesions, appear to respond to autoantigens on apoptotic cells, secrete pro-inflammatory cytokines, and stimulate B cells (reviewed in [2]).

Clonal expansions that are evident from abnormalities in the blood may reveal underlying processes that can evolve into a malignant B-cell neoplasia. For example, some patients with monoclonal gammopathy of uncertain significance (MGUS), a condition characterized by the presence of a monoclonal immunoglobulin protein present in the serum, can progress to multiple myeloma [9]. Similarly, a subset of patients with monoclonal B-cell lymphocytosis (MBL), a condition characterized by the presence of a clonal B-cell expansion in the peripheral blood, can progress to chronic lymphocytic leukemia [10].

We wondered whether a similar spectrum of conditions occurs in primary SjS, with some patients with primary SjS having polyclonal B-cell expansions, others having monoclonal expansions, and still others having transformed monoclonal expansions [4]. Since primary SjS is a systemic autoimmune disease, expanded B-cell clones in the circulation could both be pathogenic (autoreactive) and be linked to an increased risk of lymphoma. However, in primary SjS-associated lymphoma, the B-cell neoplasm is often of the mucosal associated lymphoid tissue (MALT) type [11]. It is not clear whether B-cell populations in the peripheral blood overlap with those in the glands and other tissues, although one intriguing report documents the presence of shared expanded clones in the lymph node tissue and peripheral blood of a patient with SjS who also had marginal zone lymphoma [12].

The results from several studies suggest that therapy with the B cell-depleting antibody, rituximab (anti-CD20), leads to symptomatic improvement in patients with primary SjS [13-15]. Furthermore, rituximab is known to modulate the antibody repertoire, particularly in antigen-experienced B cells [16,17]. We therefore reasoned that rituximab therapy might also alter the circulating repertoire of expanded B-cell clones in patients with primary SjS who did not have lymphoma or overt evidence of MGUS or MBL. If CD20 antibody depletion of B cells eliminated expanded B-cell clones and re-set the B-cell repertoire, this approach might favorably modify the disease course in primary SjS by eliminating pathogenic B-cell clones that could be contributing to autoimmunity or predisposing to lymphoma. Conversely, the finding of persistent B-cell clones despite CD20 B-cell depletion therapy with rituximab implies that the restoration of the B-cell repertoire to a pre-disease state is incomplete at best and that the underlying mechanisms responsible for driving the disease process remain unchecked by B-cell depletion. Furthermore, the analysis of diversification by somatic hypermutation of the antibody gene rearrangements within the expanded clone might reveal mechanistic insights into how the clone is being selected (either by self antigen or by negative regulation from the immune system). Therefore, we analyzed the B-cell repertoire of SjS subjects at baseline and after treatment with rituximab by single B-cell sorting and immunoglobulin heavy-chain gene sequencing to identify and molecularly characterize expanded B-cell clones in the blood.

Methods

Study subjects

Subjects with primary SjS, as defined by the revised American-European Consensus Group criteria, were enrolled in an open-label Autoimmunity Centers of Excellence-sponsored clinical trial of the B cell-depleting antibody, rituximab (anti-CD20). Clinical features of the study subjects have been described previously [15]. Subjects received two 1-gram doses of rituximab and 100 mg of methylprednisolone on days 0 and 15. For this study, the peripheral blood of six out of the 12 subjects with SjS were studied by flow cytometry, and antibody heavy-chain cloning was performed from single-sorted cells at baseline and at weeks 8, 14, 26, 36, and 52 after the second dose of rituximab [15]. Subjects provided informed consent to participate in this study, which was carried out in accordance with a study protocol that was approved by the institutional review boards of both the Perelman School of Medicine at the University of Pennsylvania and Duke University Medical Center.

Single-cell cloning of immunoglobulin heavy-chain gene rearrangements

Memory and germinal center B-cell subsets (shown in Additional file 1: Figure S1) were recovered from the peripheral blood as described previously [18,19]. Memory and plasmablast (PB) B-cell subsets were sorted into single wells of 96-well polymerase chain reaction (PCR) plates, and rearranged immunoglobulin heavy-chain (IgH) genes were cloned and sequenced as described previously [19].

Cloning of the germline *VH1-69* immunoglobulin heavy-chain gene rearrangement

Genomic DNA from subject 2 (SjS2) was subjected to PCR amplification with primers that flanked the germline *VH1-69* sequence. The primers used were *VH1-69* germline forward: 5′-GTG CCC TGA GAG CAT CAC ATA ACA-3′ and VH1-69 germline reverse: 5′-TTC TCC CTC AGG GTT TCT GAC ACT-3′. The cycling conditions were 10 minutes at 94°C, followed by 40 cycles of 94°C for 30 seconds, 58°C for 30 seconds, and 72°C for 30 seconds, followed by 20 minutes at 72°C. Amplicons were TA-cloned (TOPO TA kit; Life Technologies, Grand Island, NY, USA) and sequenced.

Identification of VH, DH, and JH immunoglobulin heavy-chain gene sequences

For each sequenced immunoglobulin gene rearrangement, the most closely matching germline VH, DH, and JH alleles were identified, along with the CDR3 length, functional status (in frame, out of frame, termination codon), and predicted (translated) amino acid sequence by local alignment comparison with the ImMunoGeneTics (IMGT) database [20]. Alignments were verified by using the high-throughput sequencing analysis algorithms from IMGT (V-QUEST version 1.1.2 [21]), and individual sequences were spot-checked with IgBLAST (National Library of Medicine). To avoid confusing mutations with sequencing imperfections, we considered only positions 10 to 381 of the sequences (by IMGT numbering). The numbers of identifiable IgH variable, diversity and joining (VDJ) rearrangements for each subject at each time point are summarized in Additional file 2: Table S1.

Clonal immunoglobulin gene tree construction

FASTA sequences for all of the clonal variant sequences were analyzed and graphically displayed by using neighbor joining with ClustalX, version 2.1 [22] and by using the default parameters.

Detecting selection from immunoglobulin gene sequence mutation patterns

We used the focused selection test, which is explained in detail in [23], to detect the forces of selection in the complementarity-determining regions (CDRs) and in the framework regions (FWRs) of the B-cell clones. This test has been shown to be the only one to give reliable and accurate results inferring selection from replacement-to-silent (R/S) ratios [24] and takes into account micro-sequence specificity [25,26] and transition bias [27] in somatic hypermutation. Taking these two factors into account, we calculated for any given set of nucleotides their expected mutability (that is, their relative likelihood to mutate). To estimate the expected level of selection, we count the number of silent mutations in the entire sequence and then use a micro-sequence-based targeting model of somatic mutation as described in [28] and the actual sequence composition of the mutant's source germline sequence to estimate the expected number of R mutations [29,30]. We then used the mutability of the sequence as a background to calculate, under a condition of no selection, the expected rate of R mutations in the area of interest (CDR or FWR) and the S mutations throughout the sequence. In this way, we avoided mixing potentially conflicting forces of selection in FWR and CDR. An overabundance of R mutations was considered an indication of positive selection, and a lack of R mutations was considered an indication of negative selection. When a set of clonally related sequences was analyzed, all unique mutations were grouped together, since mutations that occurred multiple times were probably an indication of common ancestry rather than being mutations that occurred more than once. In this manner, we raised the sensitivity above that obtained by analyzing a single sequence without spuriously counting mutations twice.

We further quantified selection strength by using BASELINe [30]. This algorithm extrapolates the distribution of potential models of selection pressures that could result in the observed patterns of mutation. To accomplish this task, BASELINe considers all mutations from the same condition and considers their origin (clone/sequence/CDR/FWR). In this way, we compared clones with unequal sequence numbers. BASELINe outputs a selection strength (positive or negative) whose certainty is expressed by the width of the distribution curve. All calculations presented here were done with BASELINe's web version [31].

Results

We have previously described the results of our open-label pilot trial of rituximab therapy for primary SjS [15]. In this companion study, our goal was to determine whether therapy with rituximab altered or re-set the repertoire in patients with primary SjS, using expanded B-cell clones as a read-out.

B-cell clones are frequent in peripheral blood memory B cells and plasmablasts of rituximab-treated Sjögren's syndrome subjects

To explore the effects of CD20 B-cell depletion on the B-cell repertoire, antibody heavy-chain gene rearrangements were cloned and sequenced from sorted PBs and memory B cells (see Methods and Additional file 1: Figure S1) of six subjects with SjS before and at various time points after therapy with rituximab. Since B-cell clonal expansions have previously been described in patients with primary SjS [32-34], we reasoned that it would be possible to find and track an expanded B-cell clone in one or more of the subjects over the different study time points.

Here we define clonally related heavy-chain sequences as those that share the same VH (variable), DH (diversity), and JH (joining) gene segments; have the same CDR3 (third complementarity-determining sequence) length; and have highly related CDR3 sequences (identical or differ by only one amino acid or up to three nucleotides). Among the estimated tens of billions of different CDR3 sequences [35], rearrangements that share the same VH, DH, and JH; have the same CDR3 length; and have highly similar CDR3 sequence overall are very likely to be derived from clonally related B cells. Because the sequences were recovered from single cells, finding the same sequence in two separate cells (which were separately amplified), even if they were obtained at the same time point, was considered to be indicative of clonal expansion.

A remarkable feature of the antibody heavy-chain sequencing data is that clones were identified despite sequencing relatively few B cells: out of 303 sequences, 12 independently expanded clones were identified. Members of these clones comprised 80 out of the 303 sequences (26%). The distribution of clonally related sequences and the number of sequences sampled at each time point are shown in Figure 1 and in Additional file 2: Table S1. Owing to the low numbers of circulating B cells, very few sequences were obtained between weeks 8 and 26, but a high fraction of the sequences that were obtained at these time points (mostly from SjS2) were clonally related.

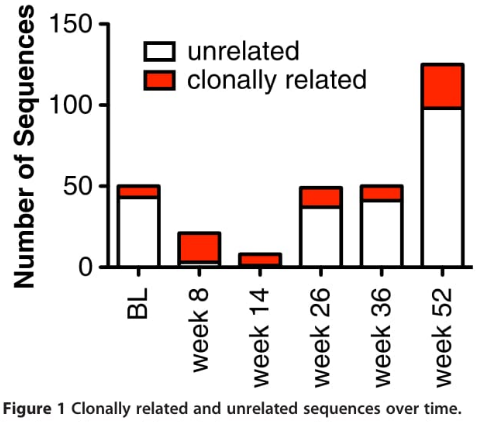


Figure 1 Clonally related and unrelated sequences over time. The stacked bar graph indicates the numbers of immunoglobulin heavy-chain (IgH) sequences with identifiable rearrangements that were obtained by single-cell sequencing in all six subjects with Sjögren's syndrome (SjS). White bars indicate unrelated IgH rearrangements. Red bars indicate IgH rearrangements that are present in at least two independent sequences and are therefore deemed to be clonally related. The numbers of unique and clonally related sequences obtained for each subject at each time point are given in Additional file 2: Table S1. BL, baseline.

Analysis of the heavy-chain sequences revealed that the B-cell clones from different subjects (and in the same subjects) did not share any obvious sequence similarities (Table 1). Additional file 2: Table S1 summarizes the clonal expansions that were identified in the subjects with SjS. In SjS1, we identified two expanded clones among only 24 total sequences. In SjS3, we identified two clones in 30 sequences, and in SjS4 we identified two clones in 65 sequences. But the most interesting subject was SjS2, who had a large expanded clone that was present during all of the time points. Members of this clone were present in the circulation in spite of rituximab therapy and consisted of over 50 independently amplified sequences (FASTA files are provided in Additional file 3: Table S2). Also in SjS2, four clones (including the large clone) were present at baseline. These data indicate that clones are also detectable in the memory B-cell repertoire in at least some subjects prior to the administration of rituximab.

The large clone from SjS2 was identified as an in-frame rearrangement of VH1-69, DH7-27, DH4-23, and JH4-02. Additional file 4: Figure S2 shows the alignment of the most frequently recovered member of the clone with the corresponding germline VH, DH, and JH sequences. We could not account for 10 nucleotides in the CDR3 by their presence in the germline VH, DH, and JH genes. Since these 10 nucleotides were shared in nearly all of the sequences, we inferred that these nucleotides most likely arose by junctional diversification (n- or p-addition) rather than by somatic hypermutation.

Members of the large expanded clone are found in circulating plasmablast and memory B-cell pools

The distribution of the members of this large expanded clone with respect to time point and B-cell subset is shown graphically, using neighbor joining in Figure 2. The lengths of the branches are fairly long (distant) when compared with the germline (GL) sequence, consistent with mutations in most of the clone members. The shape of the clone alignment is more like a "bush" than a "tree", showing no clear single direction of mutation. Rather, there are several small branches with sequence variants. The sequence variants do not display a clear-cut progression with time, at least during the time points that clonal variants were surveyed. This suggests that the major changes in the clone sequence likely occurred prior to this analysis of the clone. Of note, members of the clone were present in both CD38^{++} (PB) and CD38$^{+/-}$ class-switched (IgD^{-}) memory (CD27^{+}) B cells, indicating potential clonal overlap between these two

Table 1 CDR3 sequence features of expanded B-cell clones

Clone	VH	DH	JH	CDR3 (amino acid)
1-1	VH3-30-03*01	D3-3*01	JH6*03	CASPYYDFWSGYYMDYYYYYMDVW
1-2	VH3-30*03	DH6-19*01	JH5*01	CAKEAGSSGRAGWFDPW
2-1	VH1-69*01	DH4-23*01, DH4*02	JH4*02	CARGTGDHTTVVTPFDYW
2-2	VH3-23*04	DH6-13*01	JH3*01	CAKAVAPVGSAYDVW
2-3	VH1-2*04	DH3-10*01	JH3*02	CARDSSGGGNDAFDMG
2-4	VH4-59*01	DH3-22*01	JH6*02	CARGMKVVAGYYYYGMDVW
2-5	VH3-23*04	DH6-19*01	JH3*02	CAKAAAVGSAYDIW
2-6	VH4-59*01	DH4-17*01	JH2*01	CAREDYGDYVRW
3-1	VH4-31*03	DH3-10*01	JH6*02	CAREGNTFIRGVIGWDPKPMDVW
3-2	VH1-46*01	DH3-10*01	JH4*02	CARDGSHYDFDYW
4-1	VH3-30*02	DH1-1*01	JH6*03	CARDSRGATGTSYYYYYMDVW
4-2	VH1-2*02	DH6-19*01	JH4*02	CARDAGSAGNYDTAVAGGGFVDYW

Shown are the best matches for the corresponding VH, DH, and JH gene alleles and the translated amino acid junction (CDR3) based upon the ImMunoGeneTics (IMGT) server output for all of the expanded B-cell clones. The clone numbers indicate the subject number followed by a unique identifier for the clone. For example, 2-1 refers to first (and largest) expanded clone in Sjögren's syndrome subject 2 (SjS2). Expanded clones are molecularly defined as described in the text. DH, diversity gene segment; JH, joining gene segment; VH, variable gene segment.

pools of CD27$^+$ peripheral B-cell pools. Small clusters of sequence variants that coincide with the memory B-cell subset can be seen, further suggesting that selection for the clone members could be distinctive within the PB and memory B-cell subsets.

Somatic mutations are frequent in the large expanded clone

To gain further insight into how the expanded clone was selected, we analyzed its pattern of mutation. Although it is possible that some of the intra-clonal variation is

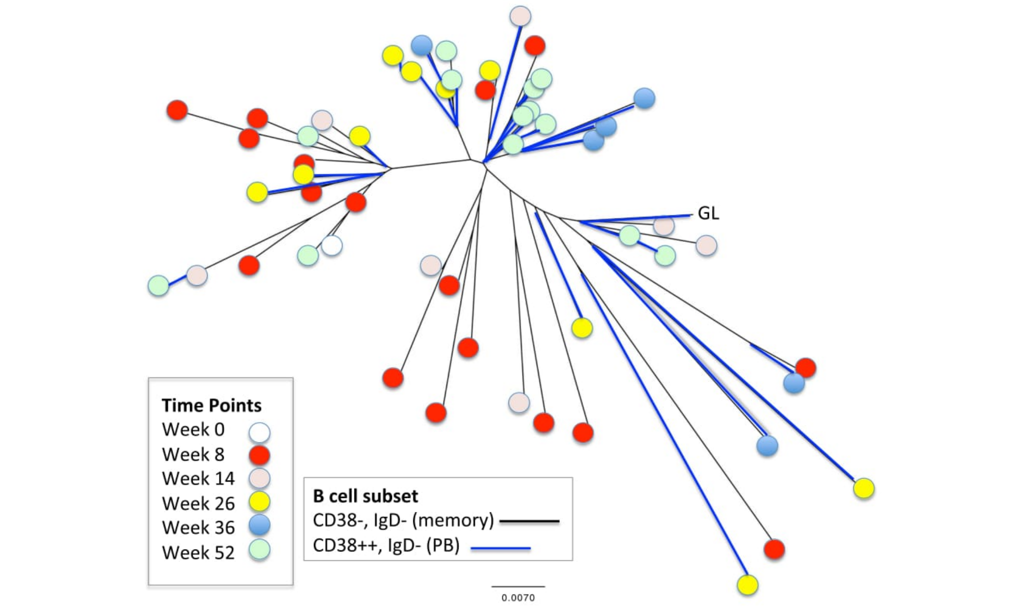


Figure 2 Time point and B-cell subset distribution of the large expanded clone. The graphical ClustalX alignment of all of the expanded clone members from Sjögren's syndrome subject 2 (SjS2) is based upon their degree of nucleotide sequence similarity. The circles at the end of each branch indicate the time point, and the colors of the branches show the memory B-cell subset from which the clone member was recovered. Blue indicates CD38^{++} (plasmablasts, PB), and black indicates CD38$^-$ (memory cells). GL, inferred germline sequence.

due to PCR error, we do not think that PCR error contributes substantially to the sequence diversity of this clone based on the large number of shared mutations between sequences and the ability to independently amplify identical sequences from different B cells.

We also considered the possibility that some of the mutations represented variations in the GL sequence due to the hypothetical presence of an unusual *VH1-69* allele in subject SjS2. We therefore cloned the subject's GL VH1-69 gene by using primers that were present in intronic sequences surrounding the unrearranged *VH1-69* gene (see Methods). We obtained only a single *VH1-69* sequence out of several independent PCRs (data not shown). Upon alignment (see Methods), the VH sequence matched VH1-69*01 (100% identity, Additional file 4: Figure S2). We therefore conclude that SjS2 is most likely homozygous for the VH1-69*01 allele and that the majority of the sequence differences among different members of the expanded clone are due to somatic hypermutation. Additional file 5: Table S3 summarizes the unique somatic mutations in the members of the expanded clone, relative to the GL sequences for VH1-69*01, JH4-02, and the CDR3 (comprised of DH7-27, DH4-23, and consensus in the areas with presumed n and p nucleotide insertions at the junctions).

Many of the somatic hypermutations in the expanded clone are silent

When the nucleotide sequences of the clone members are compared with the corresponding GL sequence, approximately half of the mutations, on average, are silent (Figure 3a). If the clone were positively selected, one would instead expect to find an increased frequency of R mutations compared with S mutations. A high frequency of S mutations is also not what one would expect to find at random. Owing to redundancies in the genetic code, the R/S ratio without any selection is generally approximately 3 to 1 in favor of R mutations. In contrast, the trends in this clone are all in the opposite direction: we find an overabundance of S mutations. However, the analysis in Figure 3a ignores the fact that many of the mutations among different members of the clone are shared and therefore are unlikely to be independent events. Therefore, we performed a more rigorous computational analysis of the somatic hypermutation pattern that took the shared mutations as well as their locations in the V region into account.

Not surprisingly, the computational analysis of the somatic hypermutation pattern was also consistent with negative selection in both the CDRs and the FWRs. Using the focused test of selection, replacement mutations are significantly under-represented among the mutations that were identified (P <0.05 at baseline and P <0.001 at other time points and overall) in all regions and overall (Table 2). To identify more clearly the strength of the negative selection and not just its presence, we used BASELINe on members of the clone (Methods), and we see quite clearly that there are high levels of negative selection (Figure 3b). Thus, there is negative selection in both the CDRs and the FWRs of the members of this clone. Furthermore, the other expanded clones identified in SjS2 and some of the other subjects exhibit somatic mutations that are consistent with negative selection (Additional file 6: Table S4). Thus, the mutation pattern found in the big clone in SjS2 is not unique to this particular clone and may be a more general feature of expanded clones in SjS.

Two additional features of this large B-cell clonal expansion become apparent when examining the distribution of mutations of the clonal variants at the six different time points (Figures 2 and 4 and Table 2). First, there is very little detectable accumulation of mutations over time, as most mutations are already observed at week 8. Second, the common mutations (that is, those found in many sequences across all time points; Figure 4) are not enriched for R mutations when compared with the rare mutations (that is, those found in only a few sequences and in one time point). We therefore considered it possible that none of these mutations is important for the formation of this lineage (that is, they may not confer selective advantage compared with the non-mutants). The common mutations could simply be derived from the same lineage that was present in the subject prior to the study.

Discussion

Clonal expansion is a fundamental property of B cells that participate in an adaptive immune response. In patients with SjS, the identification, characterization, and tracking of expanded B-cell clones over time can provide insights into how clones are selected and whether they are affected by therapy. In this study, we show that expanded B-cell clones are present in the peripheral blood of patients with primary SjS undergoing B-cell depletion therapy with rituximab. There were three main findings: (1) a large expanded clone persisted in the blood of a patient with primary SjS (subject SjS2) despite B-cell depletion with rituximab; this finding indicates that rituximab does not fully re-set the B-cell repertoire in at least some patients with primary SjS; (2) the relative proportion of expanded B-cell clones among sequences recovered from memory and PB cells was highest during the period of maximal B-cell lymphopenia; this finding suggests that the sensitivity for detecting clonal expansions is greatest during periods of iatrogenic or disease-induced lymphopenia; (3) the expanded clones harbored numerous silent mutations; this finding indicates that expanded clones are under negative selection. The frequency and kinds of somatic mutations have potential implications for monitoring the longevity and pathogenic potential of expanded clones in patients with SjS.

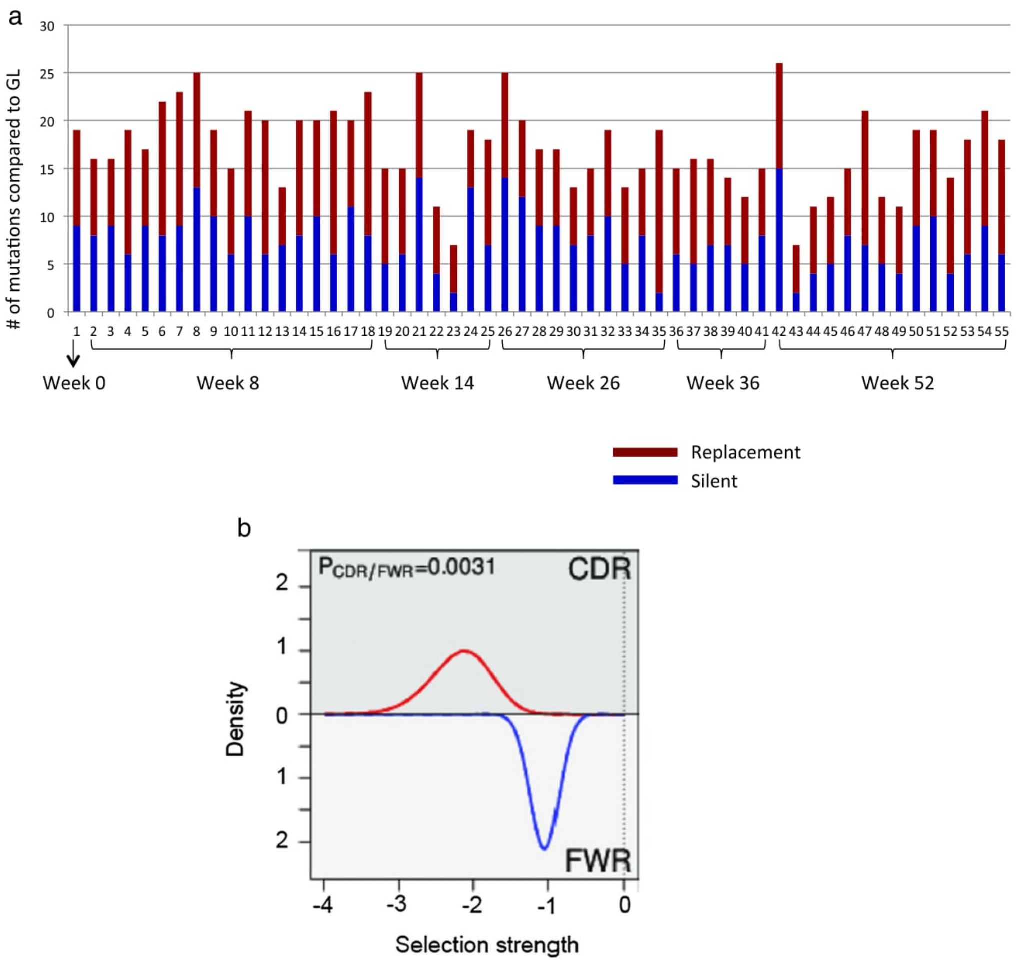


Figure 3 Evidence for negative selection in the large expanded clone. (a) High frequency of silent mutations in members of the large expanded clone. Each sequence variant (member) of the clone was compared with the germline sequence to determine the numbers and kinds of mutations it contained. The heights of the bars represent the number of mutations, and the proportions of the bars that correspond to replacement (R, red) or silent (S, blue) mutations are shown. Sequence variant numbers are given on the x-axis, as are their corresponding time points. Nucleotide positions 10-381 (ImMunoGeneTics, or IMGT) were used for the alignment. **(b)** Graphical output of BASELINe analysis [30] (see Methods). Negative values indicate negative selection, and positive values indicate positive selection. Confidence in the magnitude of selection is depicted by the width of the curve; the narrower the curve, the more confident we are of the results. Both the complementarity-determining region (CDR) and the framework region (FWR) show strong negative selection. The CDR shows evidence of significantly more negative selection than the FWR (*P* ≤0.0031).

Expanded clones are readily detected following rituximab therapy

The presence of expanded B-cell clones and their persistence after rituximab therapy are consistent with the recently reported finding of large persistent B-cell clones in the glands of SjS patients treated with rituximab [36] and extend the concept of clonal persistence in SjS to the peripheral blood. The expanded B-cell clone in SjS2 persists over the entire 1-year period of the study. It is worth noting that the large clone in SjS2 comprised the highest fraction of the circulating memory B cells during the period of maximal B-cell lymphopenia. This apparent enrichment for the clone during lymphopenia could indicate resistance to rituximab. It is possible that B cells

Table 2 Observed and expected numbers of unique R and S mutations per time point and overall in the large clone of Sjögren's syndrome subject 2

Time	N	CDR R obs	CDR S obs	FWR R obs	FWR S obs	CDR R exp	CDR S exp	FWR R exp	FWR S exp
All	55	6	24	46	47	27.18	9.59	62.36	23.86
Week 0	1	1	1	6	9	3.76	1.33	8.59	3.32
Week 8	17	5	18	30	36	19.67	6.94	45.12	17.27
Week 14	7	1	11	17	22	11.27	3.98	25.86	9.89
Week 26	10	1	14	22	25	13.70	4.84	31.43	12.03
Week 36	6	2	5	18	22	10.39	3.67	23.83	9.12
Week 52	14	1	5	10	22	8.40	2.96	19.27	7.37

Shown are the observed (obs) and expected (exp) numbers of replacement (R) and silent (S) mutations within each unique clonally related variant at each time point for members of the big clone in Sjögren's syndrome subject 2. Both the complementarity-determining region (CDR) and the framework region (FWR) exhibit significant negative selection at all time points and overall (*P* <0.05 by the focused selection test; see Methods). N, number.

in SjS are intrinsically more resistant to depletion. To this point, autoimmune mice are more resistant to anti-CD20 depletion than wild-type mice [37], but the exact mechanism is unknown. Memory B cells and PBs (the latter expressing lower levels of CD20 and dominating during the B-cell nadir) may be more resistant to anti-CD20 depletion (reviewed in [38]). Yet another non-mutually exclusive possibility is that rituximab-resistant B cells reside in protective niches in the secondary lymphoid organs or the exocrine glands and are released with time from these sites during depletion [37,39]. Another possibility is that the expanded clone undergoes more

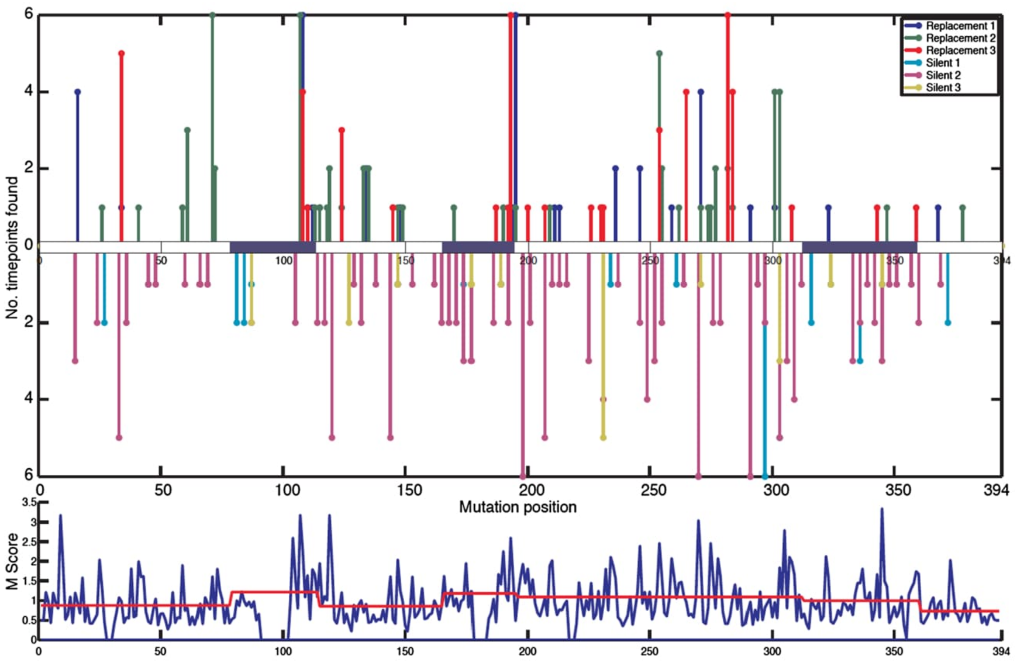


Figure 4 Time points that include specific mutations relative to the imputed germline sequence. (Top) number of time points with specific R mutations (line going up) and S mutations (line going down) at a specific position in the V region sequence. Some positions allow multiple different nucleotide substitutions. For this reason, each R or S mutation can be a combination of up to three lines, each of which represents a different nucleotide substitution at that position. The complementarity-determining regions (CDRs) are marked in blue shaded boxes. (Bottom) mutability score (see Methods) by position. A mutability score above 1 indicates a relatively increased tendency to mutate. The average mutability per region—CDR or framework region (FWR)—is indicated by the red line.

exuberant homeostatic proliferation during the period of B-cell lymphopenia than polyclonal B cells. The B-lymphopenic state brought about by rituximab therapy is accompanied by increased levels of the B-cell survival factor, BAFF [40], which, in turn, could relax selection stringency and potentially allow self-reactive or multireactive B-cell clones to flourish. In primary SjS, this speculation is particularly apt, as BAFF levels are known to be elevated [41] and to increase after rituximab therapy in this disease [42,43], including the current subject cohort and in SjS2 in particular [15]. Furthermore, high BAFF levels appear to correlate with B-cell clonal expansion in the salivary glands and are associated with an increased risk of lymphoproliferative disorders [32].

Irrespective of the mechanism of clonal resistance, the facile detection of expanded B-cell clones after rituximab therapy points to a possible way to evaluate the risk of lymphoma or the effects of B-cell targeted therapy in patients with primary SjS. It may be easiest to identify expanded clones during the period of B-cell lymphopenia, when there may be a greater opportunity for oligoclonal expansions to be recognized. Selectively surveying the memory B-cell compartment also may have increased the sensitivity of detection of expanded clones, as expanded clones are readily recovered from circulating PBs in vaccinated individuals with intact immune systems [44].

Molecular features of the expanded clone immunoglobulin heavy-chain rearrangements

The expanded clones recovered from the different subjects with SjS used a variety of VH genes. However, the big clone from SjS2 was interesting because it expressed a VH1-69/JH4 rearrangement with a D-D fusion that could be relevant to the loss of tolerance and disease pathogenesis. The *VH1-69* heavy chain includes hydrophobic amino acids in CDR2 that can adopt an unusual conformation and bind a variety of antigens [45,46]. *VH1-69* is frequently used in polyreactive and crossreactive antibody responses to HIV [47] and in the antibody responses to influenza [46] and hepatitis C virus (HCV) [48]. Moreover, molecular analysis of antibody gene rearrangements in salivary gland biopsies from patients with SjS reveals increased usage of *VH1-69*, and even specific H + L chain combinations (VH1-69 + Vκ3-20), that have also been recovered from HCV-associated non-Hodgkin lymphoma cases [49,50]. IgM+κ+B cells with rheumatoid factor activity are expanded in blood from patients with HCV-associated cryoglobulinemia that express VH1-69/JH4 and Vκ3-20 [51], suggestive of a common antigenic drive. Vκ3-20 has also been recovered repeatedly in anti-Ro 60 antibodies from separate individual patients with SjS [52]. Of note, SjS2 did have serologic evidence of rheumatoid factor (data not shown) but was HCV-negative and did not have clinical evidence of lymphoma at the time of the rituximab trial.

The D-D fusion found in the SjS2 big clone is fairly uncommon in the normal antibody repertoire, reportedly occurring in approximately 1/800 rearrangements from the blood B cells of healthy individuals [53]); however, D-D fusions may be more common in autoimmune strains of mice [54]. D-D fusions tend to lengthen the CDR3, although this particular CDR3 sequence is not exceptionally long. The translated CDR3 sequence, which for most members is CARGTGDHTTVVTPFDYW, is highly conserved among the sequence variants of this expanded clone. This sequence, like many other CDR3 sequences, has several hydrophilic residues at the VH end and more hydrophobic residues on the JH end. Whether such a mildly amphipathic sequence with extra length afforded by the second D gene segment can confer greater multireactivity is unclear and may depend critically upon the light chain, which is a major determinant of rheumatoid factor activity [55,56].

Silent mutations in the expanded clones

The most striking feature is the overall predominance of silent mutations in the heavy-chain sequences among members of the expanded clone, best exemplified in the big clone of subject SjS2. This pattern of mutations is indicative of strong negative selection and may occur because the clone has evolved to the point where it can no longer improve its affinity through mutation; any further replacement mutation thereby might lower the antibody affinity for its antigen. It seems quite clear that whatever characteristic is being 'protected' by the negative selection is related to antigen binding given that the negative selection is more pronounced in the CDRs than in the FWRs (Figure 3b, $P \leq 0.0031$). As such, the clone could be highly dependent on a specific antigen. A second possibility is that it is not selection for binding to one antigen but selection for binding to several different antigens that results in a survival advantage. In this scenario, mutations of the CDRs would limit the receptor's potential interactors, by making it less multireactive. It is intriguing in this respect that some of the commonly encountered autoantigens in systemic autoimmune disease, including the Ro and La antigens that are commonly targeted in SjS, appear to have high degrees of molecular disorder or entropy [57]. Antibody binding to a high-entropy molecule would require a greater activation energy than binding to a lower-entropy molecule and may thereby introduce structural or functional constraints in either the FWR or CDRs. Of course, if the antibody were a rheumatoid factor, it would be able to interact via a wide range of antigens by forming complexes with antibodies bound to other antigens. A third possibility is that the antibody is being selected for some property other than antigen binding, such as the ability to multimerize or self-associate [58].

In a B-cell that has lost negative regulatory signals from tonic antigen receptor crosslinking, chronic autoantibody crosslinking via persistent self or foreign antigen (model 1), multireactivity (model 2) or antibody self-association (model 3) could promote survival. It is also possible that B cells expressing certain variants of the antibody are kept in check by the immune system. A more detailed analysis of expanded clones and their functional characterization will be needed to distinguish between these intriguing potential modes of negative selection. The existence of negative selection in the CDR serves to remind us that the divisions of CDR and FWR and the impact of mutation are not set in stone. At times, a beneficial mutation may improve antibody affinity by changing an amino acid in the FWR and another mutation may degrade affinity and receptor function by changing an amino acid in the CDR. More generally, high affinity and receptor specificity are not in and of themselves always the goals of somatic selection. As in larger-scale processes of selection that underlie evolution, what determines the function under selection is the specific environment and its constraints.

Conclusions

This study documents the presence of expanded B-cell clones among memory cells and PBs in the circulation of subjects with SjS being treated with rituximab. Expanded clones were readily detected among circulating memory and PB B-cell subsets during periods of B-cell lymphopenia. One SjS subject had an expanded clone that was present prior to and at various times after B-cell depletion therapy. A detailed analysis of the nucleotide sequences of 55 members of this clone revealed a high proportion of silent mutations, suggesting that the clone was under chronic negative selection. Some of the other expanded clones isolated from other rituximab-treated SjS subjects also had frequent silent mutations. The frequency of silent mutations in an expanded clone and its B-cell subset distribution may provide a means of measuring the chronicity and selection of potentially pathogenic B cells in humans.

Additional files

Additional file 1: Figure S1. Overview of sorting and single-cell polymerase chain reaction (PCR) workflow. Peripheral blood mononuclear cells were stained with antibodies to CD3, CD14, CD16, CD19, IgD, and CD38. CD19^{+}, CD3^{-}, CD14^{-}, and CD16^{-} lymphocytes were analyzed for IgD and CD38 expression, and memory cells (CD38^{+}, IgD^{-}) and plasmablast (PB) phenotype cells (CD38^{++}, IgD^{-}) were sorted into 96-well plates for single-cell amplification and cloning as described in Methods.

Additional file 2: Table S1. Numbers of sequences and clonally related sequences from the six subjects with Sjögren's syndrome (SjS). Shown are the six subjects (SjS1 to SjS6) and the number of sequences with identifiable (VDJ) rearrangements recovered from plasmablast (PB) or memory B cells at each time point. Also shown are the numbers of sequences that are members of the 12 expanded clones (defined as sequences that share the same VH, DH, and JH and have a very similar CDR3 sequence, within one amino acid and within three nucleotides) and the percentage of total sequences that are members of expanded clones (these data are graphically shown in Figure 1). The final column lists each clone and the number of times (in parenthesis) that the clone was identified in each subject. For example, in SjS2, one clone was found in all six time points, indicated as 1(6), one clone was found at two time points, indicated as 1(2), and four clones at one time point, 4(1). BL, baseline; DH, diversity gene segment; JH, joining gene segment; VH, variable gene segment.

Additional file 3: Table S2. List of clonally related sequences of the large expanded clone from Sjögren's syndrome subject 2 (SjS2) that were analyzed for their mutation pattern. Shown are the sequences (in FASTA format) and their corresponding time points. Sequence names include a unique identifier that details the CD38 status of the B-cell subset from which the sequence was cloned (CDR38^{++} for plasmablasts or CD38$^{+/-}$ for memory cells).

Additional file 4: Figure S2. Alignment of the most common clone sequence with germline sequences. Shown is an alignment of the most common clone sequence (upper case) to the most closely matching germline sequences in the ImMunoGeneTics (IMGT) database (lower case). Probable regions of junctional diversification are indicated (n for n-addition and p for p-addition), although somatic hypermutations in the CDR3 sequence cannot be ruled out, since the germline version of the CDR3 sequence is not available for comparison.

Additional file 5: Table S3. List of unique mutations in the large expanded clone from Sjögren's syndrome subject 2 (SjS2). List of unique mutations found in the 55 sequences in Table S2, organized by position. For each mutation, we list the germline codon it mutated from, its position, the mutant codon the mutated nucleotide is part of, whether that change by itself would be an S or an R mutation, and the number of sequences in which it is found. Note that some codons are mutated in more than one position and so will appear more than once.

Additional file 6: Table S4. Analysis of somatic hypermutation selection in other expanded clones. Following the same format as Table 2, this table summarizes the selection in all of the clones identified in our experiments. Overall they exhibit negative selection, much like the large VH1-69 clone in Sjögren's syndrome subject 2 (SjS2). However, their small numbers weaken the strength of detection.

Abbreviations

BAFF: B-cell activating factor; CDR: complementarity-determining region; DH: diversity gene segment; FWR: framework region; GL: germline; HCV: hepatitis C virus; IgH: immunoglobulin heavy-chain; IMGT: ImMunoGeneTics; JH: joining gene segment; MBL: monoclonal B-cell lymphocytosis; MGUS: monoclonal gammopathy of uncertain significance; PB: plasmablast; PCR: polymerase chain reaction; SjS: Sjögren's syndrome; VH: variable gene segment.

Competing interests

ML and EWS received funding from Genentech (South San Francisco, CA, USA) for mechanistic assays that accompanied the original clinical trial [15], but Genentech did not fund the studies in this report. ML currently serves as a consultant for Genentech and has research funding from Genentech. The other authors declare that they have no competing interests.

Authors' contributions

ML contributed to the conception and design of the study. ELP contributed to the conception and design of the study and helped to analyze the sequence data and to draft the manuscript. PC and EWS contributed to the conception and design of the study and helped to analyze the clinical data. PM and LL helped to perform the cell sorting and sequencing experiments. ML helped to perform the cell sorting and sequencing experiments and to draft the manuscript. UH helped to analyze the sequence data and to draft the manuscript. WM, BZ, and NH helped to analyze the sequence data. All authors read and approved the final manuscript.

Acknowledgments

This study was supported by the Autoimmunity Centers of Excellence (U19 AI-056363), a consortium funded by the National Institute of Allergy and Infectious Disease. Additional support was provided by Genentech, Inc., and the Pennsylvania Department of Health. ELP is supported by a grant from the Lupus Research Institute, and WM is the recipient of a postdoctoral fellowship award from the Arthritis Foundation. BZ and UH are supported by a joint Drexel/Hebrew University of Jerusalem, Israel (HUJI) Coulter award. We thank the Autoimmunity Centers of Excellence publications committee for helpful feedback.

Author details

[1]School of Biomedical Engineering, Science and Health Systems, 711 Bossone Building, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA. [2]Department of Microbiology and Immunology, College of Medicine, 2900 Queen Lane, Philadelphia, PA 19129, USA. [3]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, 405B Stellar Chance Labs, 422 Curie Boulevard, Philadelphia, PA 19104, USA. [4]Duke University Medical Center, 3874 200 Trent Drive, Durham, NC 27710, USA. [5]Section of Rheumatology and Temple Autoimmunity Center, Temple University School of Medicine, 3322 North Broad Street, Philadelphia, PA 19140, USA. [6]Division of Rheumatology and Clinical Immunology, University of Pittsburgh School of Medicine, 3500 Terrace Street, BST S709, Pittsburgh, PA 15261, USA.

Received: 28 August 2013 Accepted: 13 January 2014
Published: 11 February 2014

References

1. Helmick CG, Felson DT, Lawrence RC, Gabriel S, Hirsch R, Kwoh CK, Liang MH, Kremers HM, Mayes MD, Merkel PA, Pillemer SR, Reveille JD, Stone JH, National Arthritis Data Workgroup: **Estimates of the prevalence of arthritis and other rheumatic conditions in the United States.** *Part I. Arthritis Rheum* 2008, **58:**15–25.
2. Singh N, Cohen PL: **The T cell in Sjogren's syndrome: force majeure, not spectateur.** *J Autoimmun* 2012, **39:**229–233.
3. Shiboski SC, Shiboski CH, Criswell L, Baer A, Challacombe S, Lanfranchi H, Schiødt M, Umehara H, Vivino F, Zhao Y, Dong Y, Greenspan D, Heidenreich AM, Helin P, Kirkham B, Kitagawa K, Larkin G, Li M, Lietman T, Lindegaard J, McNamara N, Sack K, Shirlaw P, Sugai S, Vollenweider C, Whitcher J, Wu A, Zhang S, Zhang W, Greenspan J *et al*: **American college of rheumatology classification criteria for Sjogren's syndrome: a data-driven, expert consensus approach in the Sjogren's International collaborative clinical alliance cohort.** *Arthritis Care Res* 2012, **64:**475–487.
4. Youinou P, Devauchelle-Pensec V, Pers JO: **Significance of B cells and B cell clonality in Sjogren's syndrome.** *Arthritis Rheum* 2010, **62:**2605–2610.
5. Smith AJ, Gordon TP, Macardle PJ: **Increased expression of the B-cell-regulatory molecule CD72 in primary Sjogren's syndrome.** *Tissue Antigens* 2004, **63:**255–259.
6. Saadoun D, Terrier B, Bannock J, Vazquez T, Massad C, Kang I, Joly F, Rosenzwajg M, Sene D, Benech P, Musset L, Klatzmann D, Meffre E, Cacoub P: **Expansion of autoreactive unresponsive CD21-/low B cells in Sjogren's syndrome-associated lymphoproliferation.** *Arthritis Rheum* 2013, **65:**1085–1096.
7. Pers JO, D'Arbonneau F, Devauchelle-Pensec V, Saraux A, Pennec YL, Youinou P: **Is periodontal disease mediated by salivary BAFF in Sjogren's syndrome?** *Arthritis Rheum* 2005, **52:**2411–2414.
8. Hislop AD, Taylor GS, Sauce D, Rickinson AB: **Cellular responses to viral infection in humans: lessons from Epstein-Barr virus.** *Annu Rev Immunol* 2007, **25:**587–617.
9. Weiss BM, Abadie J, Verma P, Howard RS, Kuehl WM: **A monoclonal gammopathy precedes multiple myeloma in most patients.** *Blood* 2009, **113:**5418–5422.
10. Landgren O, Albitar M, Ma W, Abbasi F, Hayes RB, Ghia P, Marti GE, Caporaso NE: **B-cell clones as early markers for chronic lymphocytic leukemia.** *N Engl J Med* 2009, **360:**659–667.
11. Voulgarelis M, Ziakas PD, Papageorgiou A, Baimpa E, Tzioufas AG, Moutsopoulos HM: **Prognosis and outcome of non-Hodgkin lymphoma in primary Sjogren syndrome.** *Med* 2012, **91:**1–9.
12. Hansen A, Reiter K, Pruss A, Loddenkemper C, Kaufmann O, Jacobi AM, Scholze J, Lipsky PE, Dorner T: **Dissemination of a Sjogren's syndrome-associated extranodal marginal-zone B cell lymphoma: circulating lymphoma cells and invariant mutation pattern of nodal Ig heavy- and light-chain variable-region gene rearrangements.** *Arthritis Rheum* 2006, **54:**127–137.
13. Pijpe J, van Imhoff GW, Spijkervet FK, Roodenburg JL, Wolbink GJ, Mansour K, Vissink A, Kallenberg CG, Bootsma H: **Rituximab treatment in patients with primary Sjogren's syndrome: an open-label phase II study.** *Arthritis Rheum* 2005, **52:**2740–2750.
14. Gottenberg JE, Cinquetti G, Larroche C, Combe B, Hachulla E, Meyer O, Pertuiset E, Kaplanski G, Chiche L, Berthelot JM, Gombert B, Goupille P, Marcelli C, Feuillet S, Leone J, Sibilia J, Zarnitsky C, Carli P, Rist S, Gaudin P, Salliot C, Piperno M, Deplas A, Breban M, Lequerre T, Richette P, Ghiringhelli C, Hamidou M, Ravaud P, Mariette X *et al*: **Efficacy of rituximab in systemic manifestations of primary Sjogren's syndrome: results in 78 patients of the autoImmune and rituximab registry.** *Ann Rheum Dis* 2013, **72:**1026–1031.
15. St Clair EW, Levesque MC, Luning Prak ET, Vivino FB, Alappatt CJ, Spychala ME, Wedgwood J, McNamara J, Moser Sivils KL, Fisher L, Cohen P: **Rituximab therapy for primary sjogren's syndrome: an open-label clinical trial and mechanistic analysis.** *Arthritis Rheum* 2013, **65:**1097–1106.
16. Rouziere AS, Kneitz C, Palanichamy A, Dorner T, Tony HP: **Regeneration of the immunoglobulin heavy-chain repertoire after transient B-cell depletion with an anti-CD20 antibody.** *Arthritis Res Ther* 2005, **7:**R714–R724.
17. Palanichamy A, Roll P, Theiss R, Dorner T, Tony HP: **Modulation of molecular imprints in the antigen-experienced B cell repertoire by rituximab.** *Arthritis Rheum* 2008, **58:**3665–3674.
18. Levesque MC, Moody MA, Hwang KK, Marshall DJ, Whitesides JF, Amos JD, Gurley TC, Allgood S, Haynes BB, Vandergrift NA, Plonk S, Parker DC, Cohen MS, Tomaras GD, Goepfert PA, Shaw GM, Schmitz JE, Eron JJ, Shaheen NJ, Hicks CB, Liao HX, Markowitz M, Kelsoe G, Margolis DM, Haynes BF: **Polyclonal B cell differentiation and loss of gastrointestinal tract germinal centers in the earliest stages of HIV-1 infection.** *PLoS Med* 2009, **6:**e1000107.
19. Liao HX, Levesque MC, Nagel A, Dixon A, Zhang R, Walter E, Parks R, Whitesides J, Marshall DJ, Hwang KK, Yang Y, Chen X, Gao F, Munshaw S, Kepler TB, Denny T, Moody MA, Haynes BF: **High-throughput isolation of immunoglobulin genes from single human B cells and expression as monoclonal antibodies.** *J Virol Methods* 2009, **158:**171–179.
20. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, Wu Y, Gemrot E, Brochet X, Lane J, Regnier L, Ehrenmann F, Lefranc G, Duroux P: **IMGT, the international ImMunoGeneTics information system.** *Nucleic Acids Res* 2009, **37:**D1006–D1012.
21. Alamyar E, Giudicelli V, Li S, Duroux P, Lefranc MP: **IMGT/HighV-QUEST: the IMGT(R) web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing.** *Immunome Res* 2012, **8:**26.
22. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23:**2947–2948.
23. Uduman M, Yaari G, Hershberg U, Stern JA, Shlomchik MJ, Kleinstein SH: **Detecting selection in immunoglobulin sequences.** *Nucleic Acids Res* 2011, **39:**W499–W504.
24. MacDonald CM, Boursier L, D'Cruz DP, Dunn-Walters DK, Spencer J: **Mathematical analysis of antigen selection in somatically mutated immunoglobulin genes associated with autoimmunity.** *Lupus* 2010, **19:**1161–1170.
25. Shapiro GS, Aviszus K, Murphy J, Wysocki LJ: **Evolution of Ig DNA sequence to target specific base positions within codons for somatic hypermutation.** *J Immunol* 2002, **168:**2302–2306.
26. Cowell LG, Kepler TB: **The nucleotide-replacement spectrum under somatic hypermutation exhibits microsequence dependence that is strand-symmetric and distinct from that under germline mutation.** *J Immunol* 2000, **164:**1971–1976.
27. Hershberg U, Shlomchik MJ: **Differences in potential for amino acid change after mutation reveals distinct strategies for kappa and lambda light-chain variation.** *Proc Natl Acad Sci U S A* 2006, **103:**15963–15968.
28. Smith DS, Creadon G, Jena PK, Portanova JP, Kotzin BL, Wysocki LJ: **Di- and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive B cells.** *J Immunol* 1996, **156:**2642–2652.

29. Hershberg U, Uduman M, Shlomchik MJ, Kleinstein SH: **Improved methods for detecting selection by mutation analysis of Ig V region sequences.** *Int Immunol* 2008, **20:**683–694.
30. Yaari G, Uduman M, Kleinstein SH: **Quantifying selection in high-throughput Immunoglobulin sequencing data sets.** *Nucleic Acids Res* 2012, **40:**e134.
31. **The BASELINe program can be obtained from the Kleinstein Laboratory.** [http://medicine.yale.edu/labs/kleinstein/www/software.html]
32. Quartuccio L, Salvin S, Fabris M, Maset M, Pontarini E, Isola M, De Vita S: **BLyS upregulation in Sjogren's syndrome associated with lymphoproliferative disorders, higher ESSDAI score and B-cell clonal expansion in the salivary glands.** *Rheumatology (Oxford)* 2013, **52:**276–281.
33. Guzman LM, Castillo D, Aguilera SO: **Polymerase chain reaction (PCR) detection of B cell clonality in Sjogren's syndrome patients: a diagnostic tool of clonal expansion.** *Clin Exp Immunol* 2010, **161:**57–64.
34. Vitali C, Bombardieri S, Jonsson R, Moutsopoulos HM, Alexander EL, Carsons SE, Daniels TE, Fox PC, Fox RI, Kassan SS, Pillemer SR, Talal N, Weisman MH, European Study Group on Classification Criteria for Sjögren's Syndrome: **Classification criteria for Sjogren's syndrome: a revised version of the European criteria proposed by the American-European consensus group.** *Ann Rheum Dis* 2002, **61:**554–558.
35. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, Ni I, Mei L, Sundar PD, Day GM, Cox D, Rajpal A, Pons J: **Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire.** *Proc Natl Acad Sci USA* 2009, **106:**20216–20221.
36. Hamza N, Bootsma H, Yuvaraj S, Spijkervet FK, Haacke EA, Pollard RP, Visser A, Vissink A, Kallenberg CG, Kroese FG, Bos NA: **Persistence of immunoglobulin-producing cells in parotid salivary glands of patients with primary Sjogren's syndrome after B cell depletion therapy.** *Ann Rheum Dis* 2012, **71:**1881–1887.
37. Ahuja A, Shupe J, Dunn R, Kashgarian M, Kehry MR, Shlomchik MJ: **Depletion of B cells in murine lupus: efficacy and resistance.** *J Immunol* 2007, **179:**3351–3361.
38. Leandro MJ: **B-cell subpopulations in humans and their differential susceptibility to depletion with anti-CD20 monoclonal antibodies.** *Arthritis Res Ther* 2013, **15:**S3.
39. Chan AC, Carter PJ: **Therapeutic antibodies for autoimmunity and inflammation.** *Nat Rev Immunol* 2010, **10:**301–316.
40. Miller JP, Stadanlick JE, Cancro MP: **Space, selection, and surveillance: setting boundaries with BLyS.** *J Immunol* 2006, **176:**6405–6410.
41. Mariette X, Roux S, Zhang J, Bengoufa D, Lavie F, Zhou T, Kimberly R: **The level of BLyS (BAFF) correlates with the titre of autoantibodies in human Sjogren's syndrome.** *Ann Rheum Dis* 2003, **62:**168–171.
42. Lavie F, Miceli-Richard C, Ittah M, Sellam J, Gottenberg JE, Mariette X: **Increase of B cell-activating factor of the TNF family (BAFF) after rituximab treatment: insights into a new regulating system of BAFF production.** *Ann Rheum Dis* 2007, **66:**700–703.
43. Pollard RP, Abdulahad WH, Vissink A, Hamza N, Burgerhof JG, Meijer JM, Visser A, Huitema MG, Spijkervet FK, Kallenberg CG, Bootsma H, Kroese FG: **Serum levels of BAFF, but not APRIL, are increased after rituximab treatment in patients with primary Sjogren's syndrome: data from a placebo-controlled clinical trial.** *Ann Rheum Dis* 2013, **72:**146–148.
44. Wrammert J, Smith K, Miller J, Langley WA, Kokko K, Larsen C, Zheng NY, Mays I, Garman L, Helms C, James J, Air GM, Capra JD, Ahmed R, Wilson PC: **Rapid cloning of high-affinity human monoclonal antibodies against influenza virus.** *Nature* 2008, **453:**667–671.
45. Chothia C, Lesk AM, Gherardi E, Tomlinson IM, Walter G, Marks JD, Llewelyn MB, Winter G: **Structural repertoire of the human VH segments.** *J Mol Biol* 1992, **227:**799–817.
46. Sui J, Hwang WC, Perez S, Wei G, Aird D, Chen LM, Santelli E, Stec B, Cadwell G, Ali M, Wan H, Murakami A, Yammanuru A, Han T, Cox NJ, Bankston LA, Donis RO, Liddington RC, Marasco WA: **Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses.** *Nat Struct Mol Biol* 2009, **16:**265–273.
47. Stewart A, Harrison JS, Regula LK, Lai JR: **Side chain requirements for affinity and specificity in D5, an HIV-1 antibody derived from the VH1-69 germline segment.** *BMC Biochem* 2013, **14:**9.
48. Perotti M, Ghidoli N, Altara R, Diotti RA, Clementi N, De Marco D, Sassi M, Clementi M, Burioni R, Mancini N: **Hepatitis C virus (HCV)-driven stimulation of subfamily-restricted natural IgM antibodies in mixed cryoglobulinemia.** *Autoimmun Rev* 2008, **7:**468–472.
49. De Re V, De Vita S, Gasparotto D, Marzotto A, Carbone A, Ferraccioli G, Boiocchi M: **Salivary gland B cell lymphoproliferative disorders in Sjogren's syndrome present a restricted use of antigen receptor gene segments similar to those used by hepatitis C virus-associated non-Hodgkins's lymphomas.** *Eur J Immunol* 2002, **32:**903–910.
50. Miklos JA, Swerdlow SH, Bahler DW: **Salivary gland mucosa-associated lymphoid tissue lymphoma immunoglobulin V(H) genes show frequent use of V1-69 with distinctive CDR3 features.** *Blood* 2000, **95:**3878–3884.
51. Carbonari M, Caprini E, Tedesco T, Mazzetta F, Tocco V, Casato M, Russo G, Fiorilli M: **Hepatitis C virus drives the unconstrained monoclonal expansion of VH1-69-expressing memory B cells in type II cryoglobulinemia: a model of infection-driven lymphomagenesis.** *J Immunol* 2005, **174:**6532–6539.
52. Lindop R, Arentz G, Chataway TK, Thurgood LA, Jackson MW, Reed JH, McCluskey J, Gordon TP: **Molecular signature of a public clonotypic autoantibody in primary Sjogren's syndrome: a "forbidden" clone in systemic autoimmunity.** *Arthritis Rheum* 2011, **63:**3477–3486.
53. Briney BS, Willis JR, Hicar MD, Thomas JW 2nd, Crowe JE Jr: **Frequency and genetic characterization of V(DD)J recombinants in the human peripheral blood antibody repertoire.** *Immunology* 2012, **137:**56–64.
54. Klonowski KD, Primiano LL, Monestier M: **Atypical VH-D-JH rearrangements in newborn autoimmune MRL mice.** *J Immunol* 1999, **162:**1566–1572.
55. Silverman GJ, Goldfien RD, Chen P, Mageed RA, Jefferis R, Goni F, Frangione B, Fong S, Carson DA: **Idiotypic and subgroup analysis of human monoclonal rheumatoid factors. Implications for structural and genetic basis of autoantibodies in humans.** *J Clin Invest* 1988, **82:**469–475.
56. Shlomchik M, Nemazee D, van Snick J, Weigert M: **Variable region sequences of murine IgM anti-IgG monoclonal autoantibodies (rheumatoid factors). II. Comparison of hybridomas derived by lipopolysaccharide stimulation and secondary protein immunization.** *J Exp Med* 1987, **165:**970–987.
57. Carl PL, Temple BR, Cohen PL: **Most nuclear systemic autoantigens are extremely disordered proteins: implications for the etiology of systemic autoimmunity.** *Arthritis Res Ther* 2005, **7:**R1360–R1374.
58. Kang CY, Cheng HL, Rudikoff S, Kohler H: **Idiotypic self binding of a dominant germline idiotype (T15). Autobody activity is affected by antibody valency.** *J Exp Med* 1987, **165:**1332–1343.

doi:10.1186/ar4481
Cite this article as: Hershberg *et al.*: Persistence and selection of an expanded B-cell clone in the setting of rituximab therapy for Sjögren's syndrome. *Arthritis Research & Therapy* 2014 **16**:R51.

Article 4. An atlas of B-cell clonal distribution in the human body

RESOURCE

nature
biotechnology

An atlas of B-cell clonal distribution in the human body

Wenzhao Meng[1,8], Bochao Zhang[2,8], Gregory W Schwartz[2], Aaron M Rosenfeld[2], Daqiu Ren[1], Joseph J C Thome[3], Dustin J Carpenter[3], Nobuhide Matsuoka[3], Harvey Lerner[4], Amy L Friedman[4], Tomer Granot[3], Donna L Farber[3,5], Mark J Shlomchik[6], Uri Hershberg[2,7] & Eline T Luning Prak[1]

B-cell responses result in clonal expansion, and can occur in a variety of tissues. To define how B-cell clones are distributed in the body, we sequenced 933,427 B-cell clonal lineages and mapped them to eight different anatomic compartments in six human organ donors. We show that large B-cell clones partition into two broad networks—one spans the blood, bone marrow, spleen and lung, while the other is restricted to tissues within the gastrointestinal (GI) tract (jejunum, ileum and colon). Notably, GI tract clones display extensive sharing of sequence variants among different portions of the tract and have higher frequencies of somatic hypermutation, suggesting extensive and serial rounds of clonal expansion and selection. Our findings provide an anatomic atlas of B-cell clonal lineages, their properties and tissue connections. This resource serves as a foundation for studies of tissue-based immunity, including vaccine responses, infections, autoimmunity and cancer.

B cells are key players in the generation of protective immunity[1]. During an immune response, B cells recognize antigen through their B-cell receptors (antibodies) and can receive T-cell help in specialized tissue-based structures termed germinal centers[2]. The antibody genes in activated B cells can undergo somatic hypermutation (SHM), generating antibody sequence variants within lineages of clonally related B cells[3,4]. Activated B cells can become memory B cells or differentiate to become antibody-secreting plasma cells[5]. Secreted antibodies contribute to the humoral immune response by neutralizing viruses and toxins, interacting with other immune cells via their constant regions and forming immune complexes that are processed by the reticuloendothelial system[6].

B cells can combat infection locally, activating antigen-specific T cells and elaborating cytokines that influence nearby immune cells. The tissue distribution and trafficking of B-cell clones influences how infections are controlled throughout the body. Animal studies indicate that tissue localization of B cells and plasma cells is important for protective immunity and homeostasis of bacterial microflora[7–9]. However, unlike laboratory mice, humans are outbred, and live for decades in diverse environments with exposures to many different antigens and pathogens. Humans and mice also differ in the microanatomy of their tissues and in how their B-cell subsets are defined[10,11]. Tissue-based B-cell subsets are not well understood in humans. Furthermore, most studies of human B cells have sampled the blood or tonsils. Consequently, how clones are localized to specific regions or tissues in the human body—as has been described for tissue-resident T cells[12,13]—is not known for B cells.

To understand how B-cell clones are distributed in the human body, we performed next-generation sequencing of antibody heavy-chain gene rearrangements directly from the tissues. Because clonal lineages are somatically generated, the definition of clonal networks required the sampling and comparison of several different tissues from the same individual. Hence, VH rearrangements in seven different tissues and blood were analyzed from six different human organ donors. After extensive sampling within the tissues, we identified the largest B cell clones and studied their distribution in different tissues. From these tissue distribution patterns, we created an atlas of B-cell clonal networks.

RESULTS

Sequencing pipeline and clone size thresholding

Using our resource of human tissues obtained from organ donors[12–14], we extracted DNA from blood, bone marrow, spleen, lung, mesenteric lymph node (MLN), jejunum, ileum and colon. Donor information is provided in **Table 1**. Samples were amplified and sequenced at high depth from two donors (D207 and D181) and at lower depth from four additional donors (D145, D149, D168 and D182) for confirmatory analyses. As different B-cell subsets differ in their antibody RNA transcript levels, are not fully defined in human tissues and vary in their ease of recovery in single-cell suspensions from tissues, we extracted DNA from whole tissue samples[15]. The analysis of DNA permitted efficient (one template per cell), large-scale and agnostic sampling of all B cells. Antibody heavy-chain gene rearrangements were amplified and sequenced (**Supplementary Tables 1** and **2**). Rearranged heavy-chain VH regions were used to distinguish clonally related B cells from each other by virtue of the highly diverse junctions between the V (variable), D (diversity) and J (joining) gene sequences, which comprise the third complementarity determining region (CDR3)[16,17].

[1]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [2]School of Biomedical Engineering, Science and Health Systems, Drexel University, Philadelphia, Pennsylvania, USA. [3]Columbia Center for Translational Immunology, Columbia University Medical Center, New York, New York, USA. [4]LiveOnNY, New York, New York, USA. [5]Department of Surgery and Department of Microbiology and Immunology, Columbia University School of Medicine, New York, New York, USA. [6]Department of Immunology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. [7]Department of Microbiology and Immunology, Drexel College of Medicine, Drexel University, Philadelphia, Pennsylvania, USA. [8]These authors have contributed equally to this work. Correspondence should be addressed to E.T.L.P. (luning@mail.med.upenn.edu) or U.H. (uh25@drexel.edu).

Received 20 December 2016; accepted 13 July 2017; published online 21 August 2017; doi:10.1038/nbt.3942

Table 1 Demographic characteristics of the organ donors

Donor	Age (years)	Sex	Race	Cause of death	WBC final	HCV	CMV	EBV
D145	58	M	White	CVA	15.8	0	1	1
D149	55	M	White	Anoxia	12.7	0	0	0
D168	56	F	Hispanic	CVA	2.6	0	1	1
D181	46	M	Black	CVA	10.3	0	0	0
D182	46	M	Hispanic	CVA	11.2	0	0	1
D207	23	M	Hispanic	Head trauma	15.7	0	1	1

Donor numbers are assigned by the Farber Lab. Cause of death is classified as cerebrovascular accident (CVA), head trauma or anoxia. WBC, white blood cell count in thousands per microliter. Serologic status (IgG) for hepatitis C virus (HCV), cytomegalovirus (CMV) and Epstein Barr virus (EBV). 1, positive; 0, negative.

Table 2 Sequencing metadata

Donor	Library	Total copies	Unique sequences	Clones	C_{20} clones
D145	48	2,439,338	143,573	67,342	400
D149	45	1,456,188	79,933	12,183	501
D168	47	1,224,202	68,537	23,810	356
D181	111	8,077,742	567,444	225,950	1,074
D182	51	1,302,469	80,741	24,810	375
D207	257	23,583,180	1,418,182	579,332	5,214

Library indicates the number of sequencing libraries generated per donor. Total copies refers to the total number of valid immunoglobulin VH region sequences. Unique sequences refers to the total number of unique in-frame sequences without a stop codon (productive rearrangements). Clones refers to the number of clonally related sequences, defined as having the same VH gene, the same CDR3 length and at least 85% sequence identity in the CDR3. C_{20} clones have at least 20 unique sequence instances.

We defined clonally related B cells as those sharing the same VH and JH gene groups[18], having the same CDR3 length and exhibiting at least 85% CDR3 amino acid identity[19]. This sequence similarity threshold was low enough to group together clonally related sequences with somatic mutations, while being high enough to limit too many unrelated sequences from being incorrectly combined into the same lineage[19,20]. The definition used for clone-size thresholding was the sum of the number of unique sequence variants weighted by the number of instances of each unique sequence in independent PCR amplifications. This hybrid definition ('unique sequence instances') affords some correction for differences in sequencing depth (which can influence unique sequence numbers), while not relying exclusively on resampling to establish clone size cut-offs. To assess the power of detection of clonal overlap, we performed rarefaction analysis[21] on replicate sequencing libraries from each tissue (**Supplementary Fig. 1a**). Clones larger than 20 unique sequence instances (hereafter called C_{20} clones) had at least a 75% chance of being resampled within most tissues (**Supplementary Fig. 1b**). D207 had 579,332 clones, of which 5,214 were C_{20} clones (**Table 2**).

Clonal networks within and between tissues

Tissues differed in the diversity of their C_{20} clones (**Fig. 1a**). Of the tissues analyzed, the spleen had the highest level of internal similarity and the blood had the lowest internal similarity in replicate sequencing libraries from each tissue (**Fig. 1b**). These findings are consistent with blood having the highest sampled diversity. The high diversity observed in the blood may be due to the fact that the blood undergoes extensive mixing, unlike the tissues. The internal similarity tissue map in D181 differed from D207, but became more similar to D207's map when the largest clones were removed (**Supplementary Fig. 2a**). D181 was noteworthy for having a very large clone that comprised 15% of total sequence copies in the lung that was also detected in the spleen (~5%), bone marrow (~2.5%) and blood (~1%). This massive clone, (ID 183,264 in the ImmuneDB database; http://immunedb.com/tissue-atlas), had 218,000 sampled copies in the entire body, including 4,265 unique sequence variants. In the blood, a clone of this size could be a worrisome sign of malignancy or a pre-malignant condition such as monoclonal B-cell lymphocytosis[22]. However, in the tissues, the levels of clonal expansion are not fully defined and may be influenced by regional localization of clones, including tissue-resident cells, in the sampled tissue fragments.

Analysis of the distribution of clones in the tissues revealed two prominent networks of overlapping clones in both of the deeply sequenced donors (**Fig. 1c**). One network comprised the blood, bone marrow, lung and spleen, and the other network consisted of the jejunum, ileum and colon. Clones in the MLN spanned both the blood-rich sites and the GI tract, but exhibited greater overlap with the GI tract. When clones with sequences found in the blood were computationally removed, the network of overlapping clones within the blood-rich sites was more diminished than the GI tract network (**Supplementary Fig. 2b**). Regardless of the definition of clone size used, similar networks in the blood and GI tract were observed (**Supplementary Fig. 3**). These data support the existence of two major networks of expanded B-cell clones, one in blood-rich tissues and a separate network in the GI tract.

The preceding analysis focused on pairwise tissue comparisons. To evaluate the distribution of C_{20} clones across all of the tissues, C_{20} clones from D207 and D181 were classified into different tissue representation categories: global (found in six to eight tissues), regional (three to five tissues), two-tissue and single-tissue clones (**Fig. 2**, **Supplementary Tables 3** and **4**). Globally distributed clones tended to be the largest (**Supplementary Fig. 4**). Both the regional and two-tissue clones echoed the patterns of overlap observed in the paired tissue analysis; the clones were usually present in either the GI tract or the blood-rich tissues (**Fig. 2**). C_{20} clonal distribution patterns in D207 and D181 confirmed the existence of two distinct networks inferred from the pairwise comparisons (**Supplementary Figs. 5** and **6**). Blood and GI tract networks across all of the tissues were observed at different clone-size cut-offs (**Supplementary Fig. 7a**) and at two different stringencies of clone collapsing (**Supplementary Fig. 7b**). Furthermore, the analysis of networks revealed similar blood and GI tract networks in all six donors (**Supplementary Fig. 8**). As with the two-tissue comparisons, computational removal of the clones with peripheral blood sequences had a more profound influence on the blood-rich than the GI-tract-tissue clonal networks (**Supplementary Figs. 9** and **10**).

Clonal lineage analysis

To gain further insight into how clones interconnected in different tissues, we performed clonal lineage analysis[23,24]. Clonal lineage tree structures are inferred by neighbor-joining based upon the different somatic mutations found in their unique sequence variants[23]. Lineage trees can have trunks (consisting of somatic mutations that are shared among sequences) and branches and leaves (the latter consisting of sequences with mutations that are restricted to successively smaller subsets of the sequences in the tree). Clonally diversifying lineages within the blood-rich tissues had a branch structure in which the sequences from single tissues tended to all be located on single branches (**Fig. 3a**, left and center trees). In contrast, similar lineages in the GI tract had multiple tissues represented in each branch (**Fig. 3a**, right tree). Furthermore, the lineage tree nodes (which represent unique sequences) in such GI tract clones had contributions from multiple tissues (11 of 20 nodes in the right tree vs. 1 of 45 populated nodes in blood-rich trees, left and center; **Fig. 3a**). In order to quantify

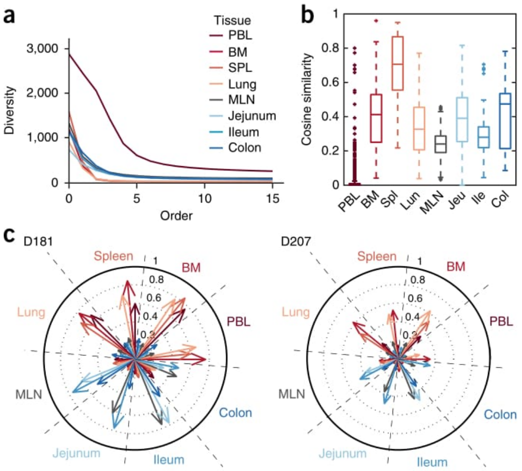


Figure 1 Diversity, similarity and networks of large clones. (**a**) Peripheral blood clones exhibit the highest sampled diversity. Diversity of clones with at least 20 unique sequence instances (C_{20} clones) is plotted at different orders (Hill numbers) in different tissues from D207. At an order of 0, the diversity is the number of different clones. At orders >1, diversity is influenced more by the most abundant clones. (**b**) Tissues exhibit higher internal similarity than blood. Box plots represent the distribution of cosine similarity between all pairs of sequencing libraries within a tissue. Similarity is assessed for C_{20} clones in D207. Boxes represent the first and third quartiles bisected by the median. Whiskers represent the most extreme data excluding outliers, where outliers (dots) are data beyond the third or first quartile by a distance exceeding 1.5 times the interquartile interval. Higher cosine values correspond to greater sharing of large clones between replicate libraries from the same tissue. (**c**) Large clones form two major networks—one in blood-rich compartments (red tones) and one in the GI tract (blue tones). Shown are the cosine similarities of C_{20} clones between tissue pairs in D181 and D207. Each wedge within the circle represents a tissue. Each arrow represents the level of overlap (cosine similarity, 0–0.8) in clones from other tissues to the clones in that wedge. Longer arrows indicate more overlap between the tissues. PBL, peripheral blood; BM, bone marrow; SPL, spleen; MLN, mesenteric lymph node.

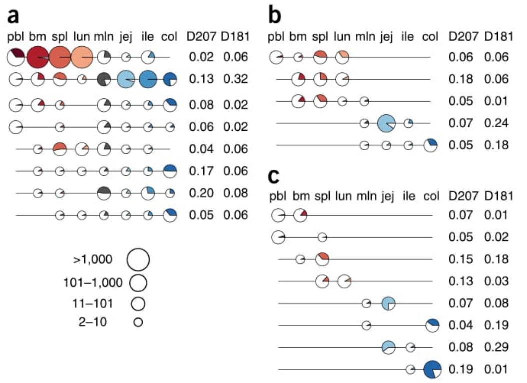


Figure 2 Tissue distributions of large clones. (**a**–**c**) Global (found in six to eight tissues) (**a**), regional (three to five tissues) (**b**) and two-tissue C_{20} clones (**c**). Each line is a clone. Each circle denotes membership of the clone in a particular tissue. The size of a circle represents the total number of sequence instances the clones have in each tissue (depicted in key). The portion of the circle that is colored represents the fraction of sequencing libraries from that tissue that contain at least one sequence of the clone (with at least two copies). The frequencies of each distribution type are indicated to the right of each clone line. Only the most frequent tissue distribution types (those that are present in at least 5% of a given tissue category in at least one of the two donors—D181 or D207) are shown. Tissues are colored as in **Figure 1**. lun, lung; jej, jejunum; col, colon; other abbreviations are as in **Figure 1**.

the extent of dispersion of mutated clonal lineages throughout different tissues, we calculated for each lineage with a specific combination of tissues, the extent of sequence sharing in every pairwise combination of tissues[25]. The level of sequence sharing we observed between different GI tract tissues was higher than that between blood-rich tissues (**Fig. 3** and **Supplementary Fig. 11**).

The presence of identical sequence variants within clones that reside in different parts of the GI tract (**Fig. 3**) implies local proliferation after somatic hypermutation[26]. Furthermore, in the GI tract, the branching structures of clonal trees show participation of multiple tissue sites in the ongoing mutation and selection process. The fact that identically mutated sequences can often be found at multiple GI sites can be explained if mutated clones disseminate throughout the GI tract, yet also undergo serial rounds of mutation and selection. B cells in these clones are most likely taken up by regional lymphatics, enter the thoracic duct and eventually reseed the GI tract via the blood[27,28].

Sequence variation within lineages also provides information about antigen experience and the extent of clonal expansion[1,29,30]. Clones with somatic VH region mutations were more frequent in the tissues than in the blood or bone marrow, suggesting that there were more memory B cells in the tissues (**Fig. 4a**), consistent with previous studies[31–33]. Of note, we found that the jejunum had the highest fraction of somatically mutated clones in every donor (**Fig. 4a**). The jejunum also has many memory T cells[13] and is a location where memory T cells accumulate in early life[12]. Across all donors, GI tract B-cell clones contained higher fractions of mutated clones and higher numbers of mutations per clone than B cells sampled in the other sites (**Fig. 4b**). GI tract clones also exhibited the highest accumulation of synonymous mutations from the nearest germline sequence, consistent with them having undergone the greatest number of divisions while the mutation process was engaged (**Fig. 4c**)[34,35].

DISCUSSION

Here we used deep immune repertoire profiling to construct an atlas that describes how expanded B-cell clones distribute and develop within the human body. This resource data set is unique in two respects. First, the generation of these data could only be accomplished through the analysis of multiple samples from multiple tissues of multiple organ donors. Second, the computational analysis of clonal lineages, which relied upon a data set of 559 sequencing libraries with over 38 million total immunoglobulin heavy-chain gene rearrangements, was at a substantially higher scale than previous work and required the development of data analysis and visualization tools. The atlas revealed two major networks of large clones, one in the blood, bone marrow, spleen and lung, and another in the GI tract. While some clones overlapped between the mucosal sites, the restriction of expanded clones in the GI tract was striking. Our analysis also revealed that the evolution of clonal lineages appeared to be distinctive in the GI tract, which contained large clones with the highest levels of somatic hypermutation.

The wide dissemination of some clones within the GI tract is consistent with earlier observations in mice and in humans[8,36].

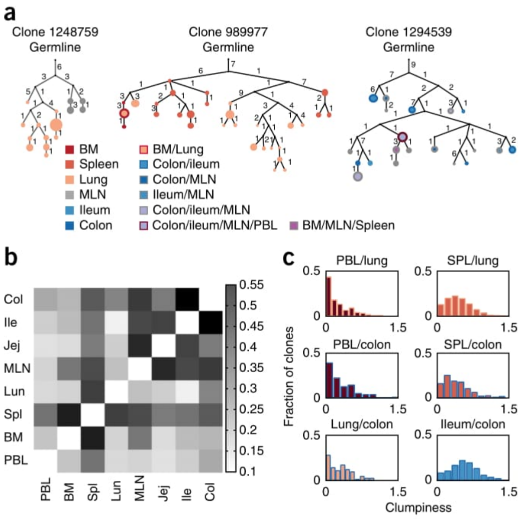


Figure 3 Analysis of sequence variants in clonal lineages. (**a**) Multi-tiered clonal lineages exhibit diversification and sharing of sequence variants within and between tissues. Trees are rooted in the closest germline VH gene allele in the IMGT database. Numbers indicate somatic mutations. Circles are colored according to the tissue distribution of the sequence variants. Circle sizes are proportional to sequence copy numbers. Black dots indicate inferred nodes. Each clone is identified by a unique identifier number in ImmuneDB (http://immunedb.com/tissue-atlas). (**b**) GI tract tissue clones exhibit extensive sharing of sequence variants. The median of the distribution of clumpiness (a metric of sequence sharing within clonal lineages, see Online Methods) is shown for all two-tissue pairs across all C_{20} clones. (**c**) Sequence sharing distributions within clonal lineages that are found in different tissue pairs. Clones with peripheral blood (PBL) and another tissue were the least mixed, followed by clones mixing blood and GI tract tissues, then blood tissue clones and, finally, GI tract clones, which were the most mixed. SPL, spleen.

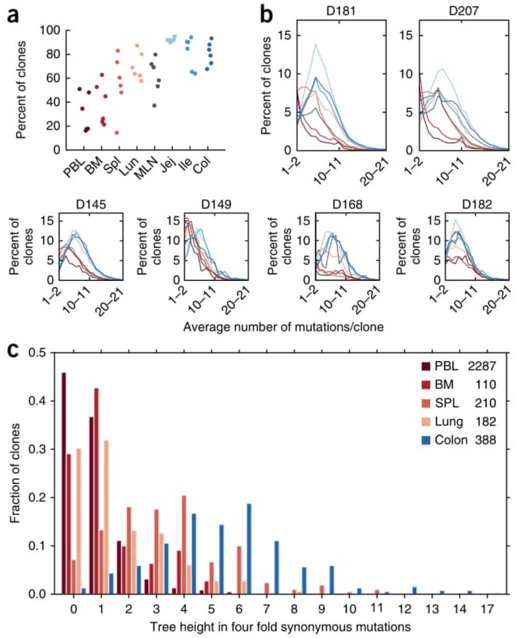


Figure 4 Somatic hypermutation in different tissues. (**a**) Clones are more mutated in the tissues. Shown are percentages of clones that have average mutation frequencies of 1% or more. In all clones, only sequences from a specific tissue are counted. Each column represents a separate tissue. Each dot represents an individual donor. (**b**) GI tract clones have right-shifted mutation frequency distributions compared to blood tissue clones in most donors. The average number of mutations per clone is plotted versus the percent of clones with that mutation level. Each line denotes a separate tissue. Segregation of mutations per clone to different tissues was accomplished as in **a**. (**c**) Lineage tree heights of C_{20} single-tissue clones in different tissues of D207. Only tissues with >100 single-tissue clones are shown. Numbers of clones are indicated to the right of the tissue names. The tree height is defined as the maximum distance of a sequence in the clone from the germline when considering only four-fold redundant synonymous mutations.

The distinctive clonal lineages in the GI tract and high levels of somatic hypermutation are consistent with antigen experience and clonal longevity, and potentially are due to chronic exposure to environmental antigens and gut microbiota. In mice, bacterial microflora have been shown to play a role in clonal generation and localization[37]; however, in humans at least some large IgA clones can persist in the colon despite antibiotic therapy[8]. There may be different classes of endogenous and environmental antigens that periodically restimulate large, tissue-resident B-cell clones in humans. The patterns of clonal expansion we have described here also imply that there is a dynamic mechanism for mutating, expanding and periodically redistributing members of the same clone to different locations within the GI tract. Clones that overlap between the blood and one or more tissues, such as the colon, may be circulating subsets of B cells that home to specific tissues, as has been described for some IgA-expressing B cells in the blood that share antigenic specificities with clones in the human GI tract[38].

The generation of this B-cell-clone tissue atlas required the development of data sharing, analysis and visualization tools. The raw sequencing data are available in GenBank and can be downloaded for further analysis as described in Online Methods. The data can also be accessed and analyzed further using our framework for B-cell repertoire analysis, ImmuneDB (http://immunedb.com/tissue-atlas)[23]. With ImmuneDB, we have created software applications for data analysis and visualization of high-throughput, immune-repertoire profiling experiments. The applications are continuously updated and are available at http://immunedb.com and at https://github.com/DrexelSystemsImmunologyLab/. This analysis pipeline permits modification of how clones are defined and how clone sizes are calculated. While these parameters and calculations are fundamental to all immune repertoire studies[19,39–41], they are often not explicitly tested when conclusions are drawn about clonal diversity or clonal overlap. Our analysis pipeline also is scalable to millions of DNA sequences, and permits clone tracing across different tissues and samples. We also provide new data visualization tools, including "line circle" plots that can be used to view tissue distributions of clones by copy number and instance number across different sampling and sequencing depths and 'clumpiness' to measure the degree of intermingling of sequence variants among different tissues within clonal lineages[25].

The tissue localization of large clones defined in this study, along with their VH and CDR3 sequence information, provide normative data for future studies of tissue-specific and response-specific motifs in antigen-driven immune responses. These data can also be used for general repertoire comparisons between health and disease, where control samples from tissues are rarely available. Even when antibody sequences are available from tissues in other data sets, they may not represent bona fide tissue-specific antibodies because the level of clonal overlap with other tissues or the blood is not known as it is with the large clones in this data set. In addition to providing data on tissue-based antibody repertoires, this atlas provides insights into which tissues have B-cell clones that are most connected with the blood.

The analysis of somatic hypermutations within individual clonal lineages may contribute to a better understanding of how and in what order B cells traffic through tissues. For example, clones that originate in one tissue may migrate and accumulate additional somatic hypermutations in a different tissue. Clonal lineages can also be analyzed for B-cell subset representation to gain further insights into ontogenic relationships between different tissue-based B-cell subsets, with more closely related subsets sharing more clonal overlap or other repertoire features. Understanding the types of B-cell subsets that reside in different tissues and how they move through the body could reveal new ways of tracking and targeting B-cell clones. Furthermore, different selective pressures may be exerted upon large clones residing in or passing through different tissues. These differences in the immune environment will need to be further defined and taken into account as we monitor and attempt to manipulate tissue-specific immunity.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the Human Immunology Core facility at the University of Pennsylvania for assistance with cell preparations and sequencing support. We thank Y. Louzoun, M. Cancro and R. Thomas Jr. for discussions. We thank the organ donor families and the LiveOnNY transplant staff and coordinators. This work was supported by NIH P01 AI106697 (D.L.F., U.H., M.J.S., W.M., B.Z., A.M.R., D.R., D.J.C., N.M., T.G. and E.T.L.P.), P30-CA016520 (E.T.L.P.) and F31AG047003 (J.J.C.T.). G.W.S. was funded by the US Department of Education Graduate Assistance in Areas of National Need (GAANN) program, CFDA Number: 84.200.

AUTHOR CONTRIBUTIONS

W.M. designed and performed experiments, developed methods, analyzed data, prepared figures and helped write the manuscript. B.Z., G.W.S. and A.M.R. developed methods for data analysis and visualization, analyzed data, prepared figures and helped revise the manuscript. D.R. performed experiments and helped revise the manuscript. J.J.C.T., D.J.C., N.M., H.L., A.L.F. and T.G. were involved in donor recruitment, organ recovery and helped revise the manuscript. D.L.F. directs the organ donor tissue resource for acquisition of tissues and helped write the manuscript. M.J.S. designed experiments, contributed ideas and helped write the manuscript. E.T.L.P. and U.H. planned the study. U.H. developed methods for data analysis and visualization, designed the data analysis and helped write the manuscript. E.T.L.P. designed experiments, contributed ideas to the data analysis, oversaw the overall study and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. McKean, D. *et al.* Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. *Proc. Natl. Acad. Sci. USA* **81**, 3180–3184 (1984).
2. Berek, C., Berger, A. & Apel, M. Maturation of the immune response in germinal centers. *Cell* **67**, 1121–1129 (1991).
3. Weigert, M.G., Cesari, I.M., Yonkovich, S.J. & Cohn, M. Variability in the lambda light chain sequences of mouse antibody. *Nature* **228**, 1045–1047 (1970).
4. Jacob, J., Kelsoe, G., Rajewsky, K. & Weiss, U. Intraclonal generation of antibody mutants in germinal centres. *Nature* **354**, 389–392 (1991).
5. Nossal, G.J. & Lederberg, J. Antibody production by single cells. *Nature* **181**, 1419–1420 (1958).
6. Schroeder, H.W. Jr. & Cavacini, L. Structure and function of immunoglobulins. *J. Allergy Clin. Immunol.* **125** (Suppl. 2), S41–S52 (2010).
7. Kroese, F.G., de Waard, R. & Bos, N.A. B-1 cells and their reactivity with the murine intestinal microflora. *Semin. Immunol.* **8**, 11–18 (1996).
8. Lindner, C. *et al.* Diversification of memory B cells drives the continuous adaptation of secretory antibodies to gut microbiota. *Nat. Immunol.* **16**, 880–888 (2015).
9. Rhee, K.J., Sethupathi, P., Driks, A., Lanning, D.K. & Knight, K.L. Role of commensal bacteria in development of gut-associated lymphoid tissues and preimmune antibody repertoire. *J. Immunol.* **172**, 1118–1124 (2004).
10. Benitez, A. *et al.* Differences in mouse and human nonmemory B cell pools. *J. Immunol.* **192**, 4610–4619 (2014).
11. Steiniger, B.S. Human spleen microanatomy: why mice do not suffice. *Immunology* **145**, 334–346 (2015).
12. Thome, J.J. *et al.* Spatial map of human T cell compartmentalization and maintenance over decades of life. *Cell* **159**, 814–828 (2014).
13. Sathaliyawala, T. *et al.* Distribution and compartmentalization of human circulating and tissue-resident memory T cell subsets. *Immunity* **38**, 187–197 (2013).
14. Thome, J.J. *et al.* Early-life compartmentalization of human T cell differentiation and regulatory function in mucosal and lymphoid tissues. *Nat. Med.* **22**, 72–77 (2016).
15. Vallangeon, B.D., Tyer, C., Williams, B. & Lagoo, A.S. Improved detection of diffuse large B-cell lymphoma by flow cytometric immunophenotyping-Effect of tissue disaggregation method. *Cytometry B Clin. Cytom.* **90**, 455–461 (2016).
16. Sakano, H., Kurosawa, Y., Weigert, M. & Tonegawa, S. Identification and nucleotide sequence of a diversity DNA segment (D) of immunoglobulin heavy-chain genes. *Nature* **290**, 562–565 (1981).
17. Glanville, J. *et al.* Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. USA* **106**, 20216–20221 (2009).
18. Zhang, B., Meng, W., Luning Prak, E.T. & Hershberg, U. Discrimination of germline V genes at different sequencing lengths and mutational burdens: A new tool for identifying and evaluating the reliability of V gene assignment. *J. Immunol. Methods* **427**, 105–116 (2015).
19. Hershberg, U. & Luning Prak, E.T. The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Phil. Trans. R. Soc. Lond. B* **370**, 20140239 (2015).
20. Yaari, G. & Kleinstein, S.H. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med.* **7**, 121 (2015).
21. Colwell, R.K. *et al.* Models and estimators linking individual-based and sample based rarefaction, extrapolation and comparison of assemblages. *J. Plant Ecol.* **5**, 3–21 (2012).
22. Goldin, L.R., McMaster, M.L. & Caporaso, N.E. Precursors to lymphoproliferative malignancies. *Cancer Epidemiol. Biomarkers Prev.* **22**, 533–539 (2013).
23. Rosenfeld, A.M., Meng, W., Luning Prak, E.T. & Hershberg, U. ImmuneDB: a system for the analysis and exploration of high-throughput adaptive immune receptor sequencing data. *Bioinformatics* **33**, 292–293 (2017).
24. Sheneman, L., Evans, J. & Foster, J.A. Clearcut: a fast implementation of relaxed neighbor joining. *Bioinformatics* **22**, 2823–2824 (2006).
25. Schwartz, G.W., Shokoufandeh, A., Ontanon, S. & Hershberg, U. Using a novel clumpiness measure to unite data with metadata: Finding common sequence patterns in immune receptor germline V genes. *Pattern Recognit. Lett.* **74**, 24–29 (2016).
26. Yuvaraj, S. *et al.* Evidence for local expansion of IgA plasma cell precursors in human ileum. *J. Immunol.* **183**, 4871–4878 (2009).
27. Gowans, J.L. & Knight, E.J. The route of re-circulation of lymphocytes in the rat. *Proc. R. Soc. Lond. B Biol. Sci.* **159**, 257–282 (1964).
28. McDermott, M.R. & Bienenstock, J. Evidence for a common mucosal immunologic system. I. Migration of B immunoblasts into intestinal, respiratory, and genital tissues. *J. Immunol.* **122**, 1892–1898 (1979).
29. Rudikoff, S., Pawlita, M., Pumphrey, J. & Heller, M. Somatic diversification of immunoglobulins. *Proc. Natl. Acad. Sci. USA* **81**, 2162–2166 (1984).
30. Sablitzky, F., Wildner, G. & Rajewsky, K. Somatic mutation and clonal expansion of B cells in an antigen-driven immune response. *EMBO J.* **4**, 345–350 (1985).
31. Briney, B.S., Willis, J.R., Finn, J.A., McKinney, B.A. & Crowe, J.E. Jr. Tissue-specific expressed antibody variable gene repertoires. *PLoS One* **9**, e100839 (2014).
32. Spencer, J., Barone, F. & Dunn-Walters, D. Generation of immunoglobulin diversity in human gut-associated lymphoid tissue. *Semin. Immunol.* **21**, 139–146 (2009).
33. Dunn-Walters, D.K., Isaacson, P.G. & Spencer, J. Sequence analysis of human IgVH genes indicates that ileal lamina propria plasma cells are derived from Peyer's patches. *Eur. J. Immunol.* **27**, 463–467 (1997).
34. Uduman, M. *et al.* Detecting selection in immunoglobulin sequences. *Nucleic Acids Res.* **39**, W499–W504 (2011).
35. Hershberg, U., Uduman, M., Shlomchik, M.J. & Kleinstein, S.H. Improved methods for detecting selection by mutation analysis of Ig V region sequences. *Int. Immunol.* **20**, 683–694 (2008).

36. Holtmeier, W., Hennemann, A. & Caspary, W.F. IgA and IgM V(H) repertoires in human colon: evidence for clonally expanded B cells that are widely disseminated. *Gastroenterology* **119**, 1253–1266 (2000).
37. Masahata, K. *et al.* Generation of colonic IgA-secreting cells in the caecal patch. *Nat. Commun.* **5**, 3704 (2014).
38. Berkowska, M.A. *et al.* Circulating human CD27-IgA+ memory B cells recognize bacteria with polyreactive Igs. *J. Immunol.* **195**, 1417–1426 (2015).
39. Khan, T.A. *et al.* Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci. Adv.* **2**, e1501371 (2016).
40. Gupta, N.T. *et al.* Hierarchical clustering can identify B cell clones with high confidence in ig repertoire sequencing data. *J. Immunol.* **198**, 2489–2499 (2017).
41. Ralph, D.K. & Matsen, F.A. IV Likelihood-based inference of B cell clonal families. *PLoS Comput. Biol.* **12**, e1005086 (2016).

ONLINE METHODS

Study subjects. Tissue acquisition was coordinated through a research protocol and material transfer agreement with LiveOnNY (formerly the New York Organ Donor Network, NYODN). All tissues were obtained from research consented deceased (brain dead) organ donors at the time of clinical procurement for life-saving transplantation. All donors were free of cancer as well as hepatitis B, hepatitis C, and were HIV negative. The study was determined to be non-human subjects research by the Columbia University Institutional Review Board, as tissue samples were obtained from deceased individuals. LiveOnNY transplant coordinators identified research-consented organ donors, and then coordinated tissue acquisition with the on-call surgeon for the project. Tissue collection occurred immediately after the donor organs were flushed with cold preservation solution and the clinical procurement process was completed. Once removed, tissue samples were placed in sterile specimen cups, submerged in cold saline, and brought to the laboratory and processed within approximately 2–4 h of organ procurement.

Sample processing and DNA extraction. Tissues from organ donors were shipped overnight to the University of Pennsylvania. Cells were liberated from tissue fragments (approximately 1 cm^3 in size) using physical methods (tissues were washed with PBS and cut into small pieces using razor blades) and placed in lysis buffer with proteinase K following the manufacturer's directions (Qiagen, Valencia, CA, Cat. No. 158667). DNA from peripheral blood and bone marrow was extracted using a Qiagen Gentra Puregene blood kit, following the manufacturer's directions (Qiagen, Valencia, CA, Cat. No. 158389). All blood and bone marrow samples were processed upon receipt. For D145 and D149, fresh tissue samples were processed when received. For donor D168, D181, D182 and D207, tissues were snap-frozen and processed later for DNA extraction. DNA quality and yield were evaluated by spectroscopy (Nanodrop, ThermoFisher Scientific, Waltham, MA).

VH rearrangement amplification. Immunoglobulin heavy-chain family-specific PCRs were performed on genomic DNA samples. The libraries for sequencing used the Illumina MiSeq platform and were prepared using a cocktail of VH1, VH2, VH3, VH4, VH5, VH6 from framework region (FR)1 forward primers, and one consensus J region reverse primer modified from the BIOMED2 primer series[42]. To capture the full-length VH region sequences, VH family leader primers were also used as described[43]. Primer sequences and locations are provided in **Supplementary Table 1** and the number of sequencing libraries prepared with the different primer mixes is given for each donor tissue combination in **Supplementary Table 2**. For the VH leader PCR, three separate amplification mixes were prepared for each sample: VH3 and VH3-21 primers (mix 1), VH4 and VH6 primers (mix 2), VH1, VH2 and VH5 primers (mix 3). In each 50 μL mix with 1 unit of AmpliTaq gold (Applied Biosystems, Foster City, CA), VH leader primers were used at a concentration of 0.6 μM, genomic DNA at 200–400 ng (except for D168 MLN (mesenteric lymph node), which was hypocellular), 0.2 mM dNTPs and 1× PCR buffer with 1.5 mM $MgCl_2$. For the FR1 PCR, the VHFR1 and 3′ JH primers were used at 0.6 μM in a reaction volume of 50 μL using the same AmpliTaq Gold system. Amplification conditions for the leader PCR were primary denaturation followed by cycling at 95 °C 30s, Ta (56 °C for mix 1, 58 °C for mix 2 and 60 °C for mix 3) for 90s, extension at 72 °C for 90s for 35 cycles, and a final extension step at 72 °C for 10 min. Amplification conditions for FR1 PCR were primary denaturation, followed by cycling at 95 °C 45 s, 60 °C for 45 s, extension at 72 °C for 90 s for 35 cycles, and a final extension step at 72 °C for 15 min.

Library preparation and sequencing. Amplicons were purified using the Agencourt AMPure XP beads system (Beckman Coulter, Inc., Indianapolis, IN) in a 1:1 ratio of beads to sample. Second-round PCRs to generate the sequencing library were carried out using 2–4 μL of the first round PCR product and 2.5 μL each of NexteraXT Index Primer S5XX and N7XX primers, 0.325 μL KAPA in a reaction volume of 25 μL. Amplifications were carried out as recommended by the manufacturer (Illumina, San Diego, CA). To confirm adequacy of amplification, aliquots of both the 1st and the 2nd round PCR products were run on agarose gels. Library quality was evaluated using Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA) and quantified by Qubit Fluorometric Quantitation (Thermo Fisher Scientific, Grand Island, NY). A sharp single band from Bioanalyzer analysis indicated a good quality library and was used for sequencing. The reading from Qubit using dsDNA HS (high sensitivity) assay kit (Cat. No. Q32851) was used to calculate the molarity of the library. Libraries were then loaded onto an Illumina MiSeq in the Human Immunology Core Facility at the University of Pennsylvania. 2 × 300 bp paired-end kits were used for all experiments (Illumina MiSeq Reagent Kit v3, 600-cycle, Illumina Inc., San Diego, Cat. No. MS-102-3003).

Filtering and germline gene association of raw sequence data. Raw sequence data (fastq files) were filtered based upon the positional Q score. A sliding window of 10 bases was used to run through all sequences. Any sequence with a 10 bp section that had an average Q score lower than 20 was removed, along with any sequence that was shorter than 100 bases. R1 and R2 sequences were paired using pRESTO version 0.4.7 (ref. 44). Sequences which were amplified by FR1 primers were trimmed after alignment to the international ImMunoGeneTics information system (IMGT) position 82, inclusive, to remove the primer sequence. In the composite R1+R2 read, any base with a Q score lower than 20 was replaced with an N and all sequences with more than 10 N's were removed. Filtered sequences were aligned to their appropriate germline gene using the Anchor method[18]. Then a sliding window of 30 nucleotides was used to flag insertion/deletions (indels). Any sequence with more than 18 mismatches in a sliding window was flagged as potentially having indels. These sequences were also removed.

Identification of unique sequences and association of sets of clonally related sequences. After V(D)J identification, duplicated sequences (those with identical nucleotide sequences) were collapsed in a two-level process. First, sequences were collapsed within each sample library into sets of unique sequences. Partial sequences were reconstructed into full length with their missing bases replaced by 'N'. All duplicated sequences were collapsed to the longest sequence. Second, the numbers of different sequencing libraries that contained the same sequence were counted. Each time a sequence appeared at least one time in another library, it was counted as a separate "instance" of that sequence. Because DNA was sequenced, these instances represented separate B cells in which the same sequence was present. For each unique sequence, its total copy number and its number of instances were tracked for each tissue and overall in each donor. As PCR amplification and DNA sequencing are prone to error[45], only sequences with more than one copy were considered for analysis. Next, unique sequences (copy >1) were parsed into clones (sets of B cells with sufficiently similar VH region rearrangements that they very likely share a common ancestry). Sequences were divided into bins with a common VH/JH gene and the same CDR3 length[18]. Then, all of the sequences in the same bin from a given donor were grouped into clones based on their CDR3 amino acid similarity. Any two sequences within the same clone have CDR3 sequences that have 85% or higher amino acid identity[19]. This process is dependent on order, so sequences with the highest copy numbers in each bin were the first to be associated with a specific clone.

Clone size thresholding and rarefaction analysis. Even at the current level of high-depth sequencing, the true repertoire is being under-sampled. Sampling is particularly important for estimations of clonal overlap. Only clones with larger sizes will be sufficiently sampled to demonstrate overlap or lack of overlap. For this analysis, we considered clone size to be the sum of the number of uniquely mutated sequences and all the different instances of the same unique sequence that are found in separate sequencing libraries. In other words, we sum across all sequencing libraries the unique members of a clone that are found in each sequencing library. Thus, for instance, if a clone C_1 were found in two sequencing libraries (L_1 and L_2) from spleen and had n unique sequences in L_1 and m unique sequences in L_2, we would say the clone size was $n+m$, even if all of the unique sequences in L_2 were identical to those of L_1. We refer to this hybrid clone size measure as "unique sequence instances."

To estimate clone sizes for which sampling was sufficient, a sample-based rarefaction analysis was performed on clones of different sizes found in exactly two samples[21]. This is a widely accepted method to estimate sufficient sampling in different fields of ecology and plant biology. The code written for diversity and rarefication analyses can be found at https://github.com/DrexelSystemsImmunologyLab/diversity and in references[46,47]. Using these methods, we determined that the repertoire of clones with at least 20 unique sequence instances in a tissue were sufficiently large that we should be observing three quarters of their real population. For this reason, unless stated otherwise, we only considered in our analysis clones that had at least 20 unique sequence instances (C_{20} clones) in at least one tissue.

Measures of clonal overlap across samples and across tissues. To measure the amount of overlap of clones between different tissues, the cosine similarity was calculated for all combinations of two-tissue samples for the most heavily sequenced donors, D207 and D181. The cosine similarity is calculated as:

$$\frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

where A_i and B_i are components of vectors A and B, respectively. Each attribute in vector A or B represents the size of the clone in sample 1 (e.g., colon from D207) and sample 2 (e.g., spleen from D207), respectively.

To quantify overlap within a tissue, the clone size was defined as the number of unique sequences in each sequencing library from that tissue. Thus the size of clone C in each sequencing library L_i from a given tissue, was calculated as $S(C_{L_i}) = U_{L_i}$, where U_{L_i} is the number of unique sequences of clone C in sequencing library L_i.

To quantify overlap between tissues, clone size was defined as the sum of the number of sequencing libraries in which at least one unique sequence variant of the clone occurred for each sequence variant within each tissue. In other words, the size of clone C in a given tissue was calculated $S(C_{tissue}) = k_{tissue}$, where k_{tissue} is the number of sequencing libraries of a given tissue in which clone C is found.

To characterize the interconnections of clones between more than two tissues, sequence alignment and clone assembly were first processed by MiXCR (version 1.7)[48]. To compare the distribution of overlapping clones across all tissues, the TrackClonotypes function of VDJtools (version 1.0.7)[49] was used with the default settings. MiXCR was used to concatenate all of the FR1 and leader sequencing libraries for each tissue within each donor. The analysis was performed with different clone size cut-offs, including 0.01% (the default), 0.005% and 0.001% of total copies within at least one tissue (**Supplementary Fig. 7a**). The clone size cut-off is expressed as a percentage of total copies in a given tissue from a single donor.

Construction of B-cell clonal lineages. Lineage trees of clonally related sequences (clonal lineages) were made based on their mutations using clearcut v1.0.9 with traditional neighbor joining and deterministic joining[24]. To reduce the contribution of sequencing errors, when calculating clumpiness (see section on Sequence sharing across tissues, Methods), we considered only mutations that occurred in at least two sequences in a donor. Lineage nodes could thus include multiple unique sequences a single mutation apart and their multiple sample instances.

Sequence sharing across tissues measured by clumpiness within clonal lineages. Clumpiness is a measure of the tendency of two label types to be close on a given hierarchical structure[25]. In the lineage tree of a clone that spans two or more tissues, a higher clumpiness value indicates that members of the clone exhibit more intermingling (mixing and overlapping of sequence variants) in the different tissues. In this case, we assessed the clumpiness of tissue types that annotated the mutant B-cell receptor lineages. The clumpiness of clonal lineages was calculated by applying the clumpiness measure to each tissue and pairs of tissues on a B-cell lineage tree. Clumpiness measures the clumping of groups of leaves, so the lineages were transformed in such a way that any intermediate vertex containing a compartment label became an unlabeled vertex with an additional edge connecting a leaf with the compartment label. For the analysis in **Figure 3b** and **Supplementary Figure 11**, we considered all C_{20} clones in D207. Some sequences/nodes are themselves observed in multiple tissues. However, in most cases the relationship was lopsided, with one tissue being highly dominant. We therefore assigned a tissue to each node by the tissue of the majority of sequence instances found at that position in the lineage. The code written for the analysis of clumpiness can be found at https://github.com/DrexelSystemsImmunologyLab/find-clumpiness and is described in ref. 25.

Mutation analysis. We calculated the average mutation frequency of each clone in each tissue. All instances of each unique sequence within a clone (across multiple sequencing libraries from a given tissue) were compared to their assigned germline VH gene. For **Figure 4a**,**b**, the mutation frequency was calculated as the number of mismatched nucleotides divided by the total number of nucleotides to the end of V gene until three mismatches were found in CDR3. The mutation frequency of a clone in a given tissue was calculated as the average of mutation frequency of all sequences in that clone that belonged to that tissue. To assess clonal division during somatic mutation, we counted the maximal number of synonymous mutations from the germline observed in each clone. Only synonymous mutations in amino acids that are encoded with four codons were considered; such 'four-fold silent' codons encode the same amino acid with any nucleotide in the third position, hence mutations in the third position of these codons should be neutral to selection.

Significance testing. In all cases non-parametric statistical tests were used to compare distributions and the necessary assumptions for these tests were met. The Kruskal–Wallis one-way analysis of variance was used to show that sets of distributions differed. The Mann–Whitney test was used to compare between individual distributions.

Data availability and accession code availability statements. Raw data, including barcoding schema and sequencing-run metadata are publicly available. QC filtered data in the form of fasta files that have undergone quality thresholding, paired read assembly and collapsing into unique sequences are available via SRA under accession number PRJNA343738. Analyzed data (e.g., VDJ assignment, clonal lineage assignment and somatic hypermutation selection analysis, etc.) are available for viewing and download at (http://immunedb.com/tissue-atlas). A description of the immune database that was constructed for viewing these data is provided in http://immunedb.com[23]. Listings of C_{20} clones in the two most highly sequenced donors are provided in **Supplementary Tables 3** (D207) and **4** (D181). All code for antibody repertoire data analysis and visualization used in this manuscript is freely available without restriction at https://github.com/DrexelSystemsImmunologyLab. A **Life Sciences Reporting Summary** is available for this paper.

42. van Dongen, J.J. *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).
43. Meng, W. *et al.* Trials and tribulations with VH replacement. *Front. Immunol.* **5**, 10 (2014).
44. Vander Heiden, J.A. *et al.* pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* **30**, 1930–1932 (2014).
45. Schirmer, M. *et al.* Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res.* **43**, e37 (2015).
46. Schwartz, G.W. & Hershberg, U. Germline amino acid diversity in B cell receptors is a good predictor of somatic selection pressures. *Front. Immunol.* **4**, 357 (2013).
47. Schwartz, G.W. & Hershberg, U. Conserved variation: identifying patterns of stability and variability in BCR and TCR V genes with different diversity and richness metrics. *Phys. Biol.* **10**, 035005 (2013).
48. Bolotin, D.A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).
49. Shugay, M. *et al.* VDJtools: unifying post-analysis of T cell receptor repertoires. *PLoS Comput. Biol.* **11**, e1004503 (2015).

doi:10.1038/nbt.3942

Leader Primer Mixes

NexteraR2-Hu-VH1-LD	5’-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCATGGACTGGACCTGGAG-3’
NexteraR2-Hu-VH2-LD	5’-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATGGACACACTTTGCTCCAC-3’
NexteraR2-Hu-VH3-LD	5’-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTANCCATGGAGTTTGGGCTGAG-3’
NexteraR2-Hu-VH3-21-LD	5’-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCATGGAACTGGGGCTC-3’
NexteraR2-Hu-VH4-LD	5’-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATGAAACACCTGTGGTTCTTCC-3’
NexteraR2-Hu-VH5-LD	5’-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTCAACCGCCATCCTCG-3’
NexteraR2-Hu-VH6-LD	5’-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTCTGTCTCCTTCCTCATCTTCC-3’
NexteraR1-Hu-JHmix1	5’-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTACGTNCTGAGGAGACGGTGACC-3’
NexteraR1-Hu-JHmix2	5’-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGCNCTGAGGAGACGGTGACCA-3’
NexteraR1-Hu-JHmix3	5’-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGNCTGAGGAGACGGTGACCAGG-3’

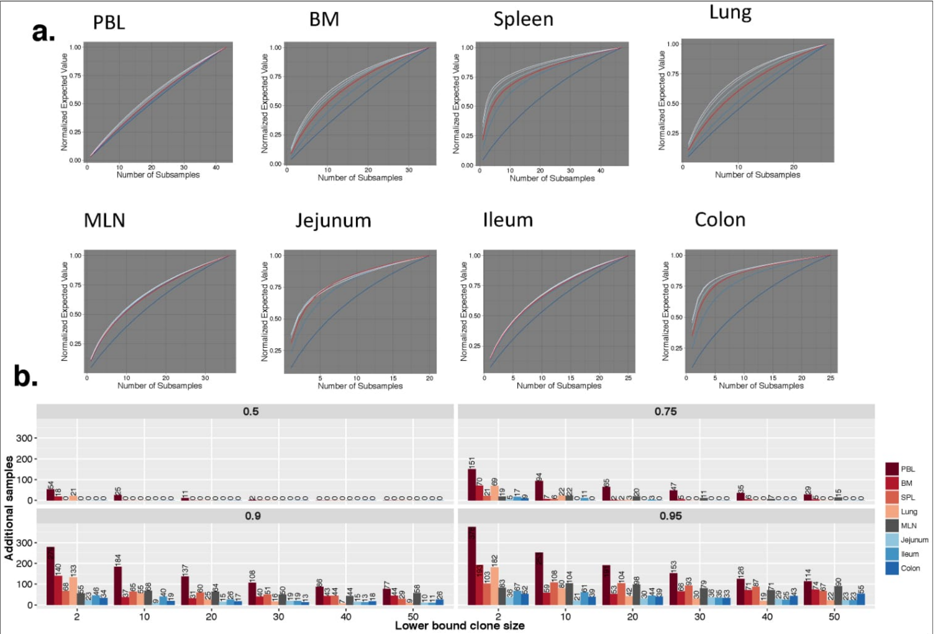
FR1 Primer Mixes

NexteraR2-Hu-VH1-FW1	5’- GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGCCTCAGTGAAGGTCTCCTGCAAG -3’
NexteraR2-Hu-VH2-FW1	5’- GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTCTGGTCCTACGCTGGTGAAACCC -3’
NexteraR2-Hu-VH3-FW1	5’- GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTGGGGGGTCCCTGAGACTCTCCTG -3’
NexteraR2-Hu-VH4-FW1	5’- GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTTCGGAGACCCTGTCCCTCACCTG -3’
NexteraR2-Hu-VH5-FW1	5’- GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGGGGAGTCTCTGAAGATCTCCTGT -3’
NexteraR2-Hu-VH6-FW1	5’- GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCGCAGACCCTCTCACTCACCTGTG -3’
NexteraR1-Hu-JHmix4	5’- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTACGTNCTTACCTGAGGAGACGGTGACC -3’
NexteraR1-Hu-JHmix5	5’- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGCNCTTACCTGAGGAGACGGTGACC -3’
NexteraR1-Hu-JHmix6	5’- TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGNCTTACCTGAGGAGACGGTGACC -3’

Supplementary Table 1: PCR Primers. Shown are the nucleotide sequences of the primers used for PCR amplifications (see **Methods** for PCR details).

Donor	Primers	PBL	BM	SPL	Lung	MLN	Jejunum	Ileum	Colon
D145	LD	6	6	6	6	6	6	6	6
D149	LD	3	6	6	6	6	6	6	6
D168	LD	6	5	6	6	6	6	6	6
D181	LD	9	9	3	6	6	6	6	6
D181	FR1	8	7	8	5	8	8	8	8
D182	LD	9	6	6	6	6	6	6	6
D207	LD	6	6	16	6	6	6	6	6
D207	FR1	37	29	31	20	30	14	19	19

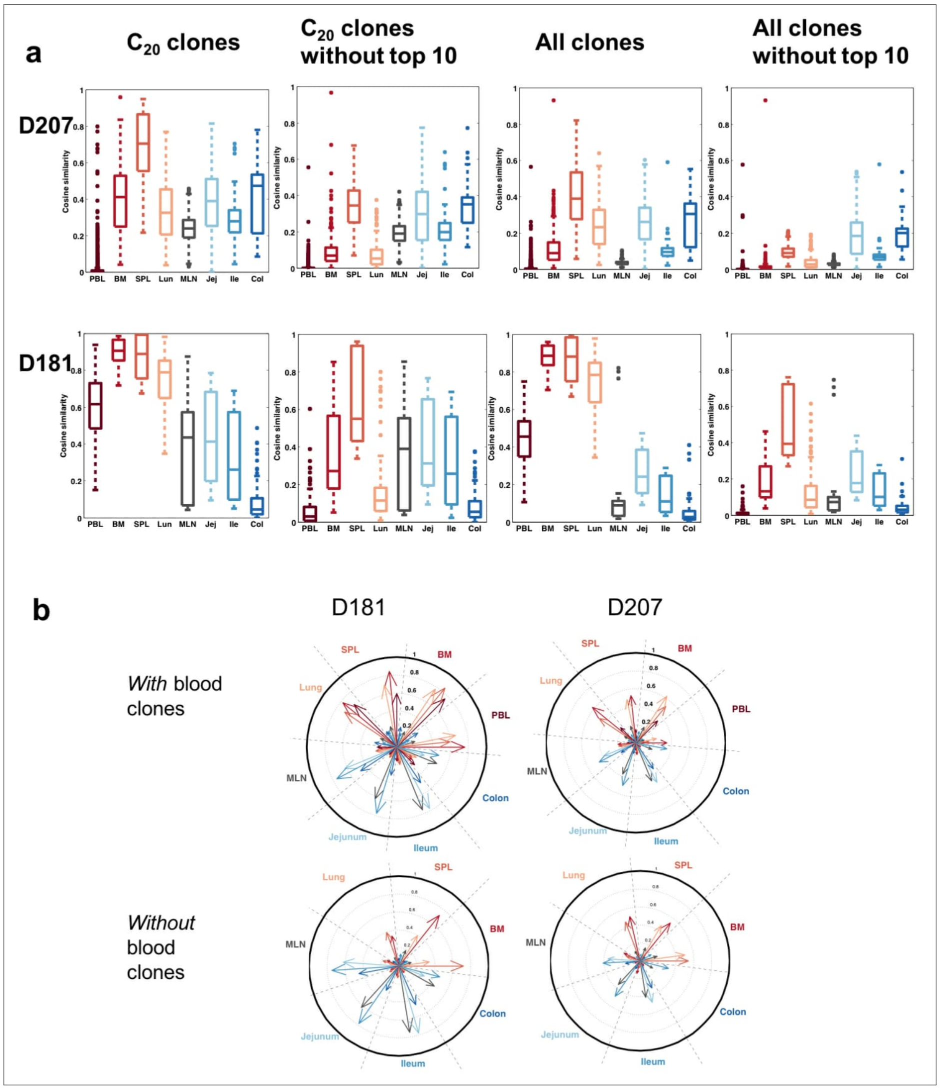
Supplementary Table 2: Donor Tissue Amplicon Libraries. Shown are the numbers of sequencing libraries for each donor tissue. LD = leader primer mixes; FR1 = framework region 1 primer mix.



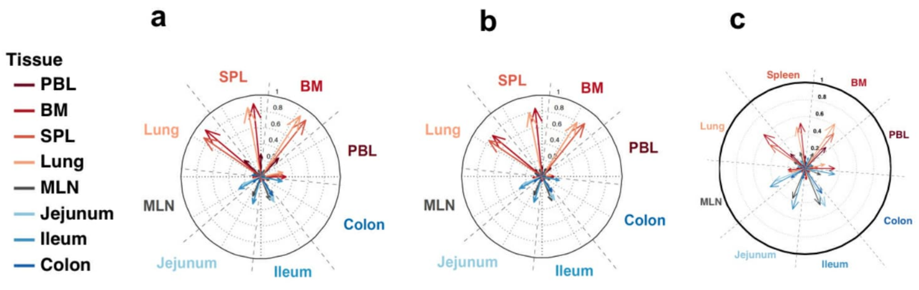
Supplementary Figure 1

Rarefaction Analysis and Sampling.

a, Rarefaction plots show the expected levels of coverage for different subsamples of sequencing libraries for repertoires of clone size of at least 2 (darkest blue line), 10 (second darkest blue), 20 (red, the cut-off chosen for the overlap analysis), 30 (light blue), 40 (lighter blue) or 50 and greater (light grey) in the repertoire of D207. Separate plots are shown for each tissue. ***b,*** Sampling needed for clones of different sizes and coverage thresholds in D207. Bars indicate the number of additional libraries needed to sample 0.5,0.75,0.9 and 0.95 (fractions) of all clones at different clone size cut-offs. Colors of the bars indicate the tissue. PBL = peripheral blood; BM = bone marrow; SPL = spleen; MLN = mesenteric lymph node.


a
C_{20} clones
C_{20} clones without top 10
All clones
All clones without top 10
D207
D181
Cosine similarity
0
0.2
0.4
0.6
0.8
1
PBL
BM
SPL
Lun
MLN
Jej
Ile
Col
b
D181
D207
With blood clones
Without blood clones
SPL
BM
PBL
Lung
MLN
Jejunum
Ileum
Colon

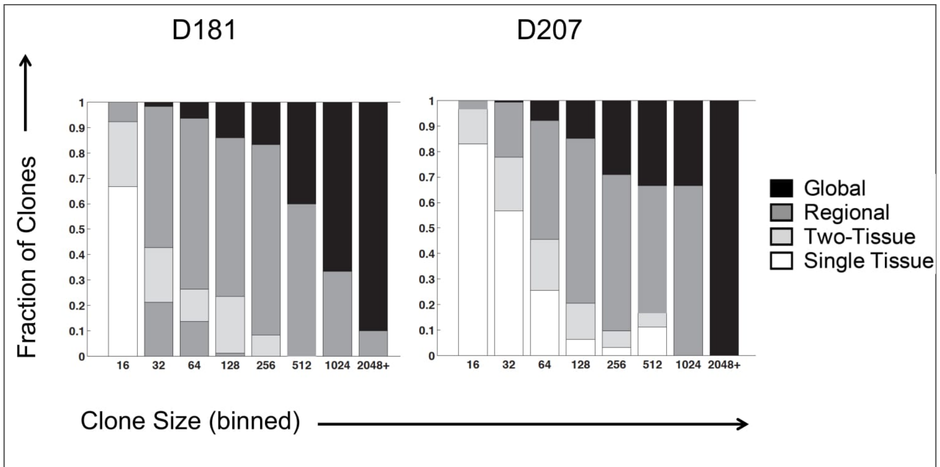
Supplementary Figure 2
Cosine Similarity Analysis.
a, Box plots of distribution of cosine similarity between all sample pairs within a tissue. Similarity is assessed for C_{20} clones, C_{20} clones without the top 10 clones in each tissue, all clones, or all clones without the top 10 copy number clones in each tissue. Boxes represent the first and third quartiles bisected by the median. Whiskers represent the most extreme data excluding outliers, where outliers (dots) are data beyond the third or first quartile by a distance exceeding 1.5 times the inter-quartile interval. The C_{20} clone panel for D207 corresponds to **Fig. 1b** in the main manuscript. ***b,*** The cosine similarity of clonal populations between tissue pairs in D181 and D207 with and without blood clones. Each section represents a tissue and each arrow within each section represents the level of overlap (cosine similarity) between that tissue and clones from other tissues. Longer arrows indicate higher value of cosine similarity indicating more overlap of clonal populations between the tissues. Top: all C_{20} clones (same as **Fig. 1c**). Bottom: C_{20} overlap without blood clones; clones that have sequences in the blood were computationally removed from all tissues.



Supplementary Figure 3

Analysis of Clonal Networks in D207 Using Different Clone Size Definitions.

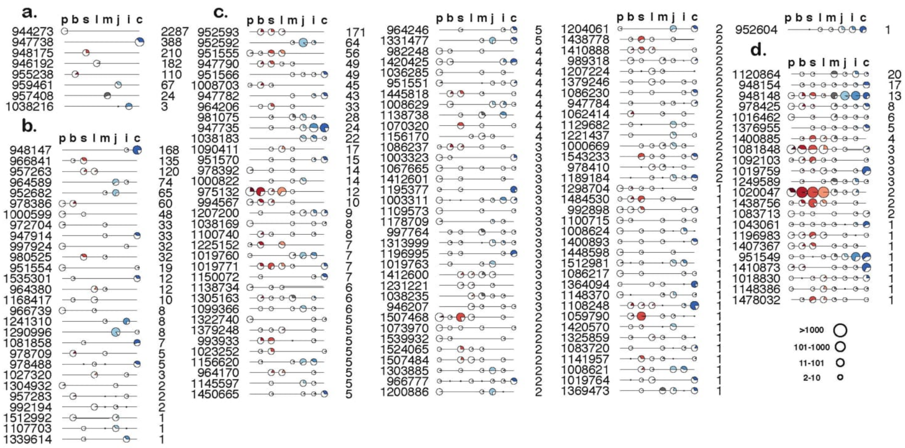
Clonal overlap was evaluated using cosine similarities between different tissue pairs, as described in the manuscript methods. Each wedge represents a tissue and each arrow represents the cosine similarity between the tissues (arrow colors are indicated in the legend, on the left). Three different definitions of clone size are used for the analysis: ***a,*** the copy number of all sequences in a clone in a given tissue; ***b*,** the total number of unique sequence instances in a clone in a given tissue; and ***c*,** the definition that we used in the manuscript (same image as **Fig. 1c**) - the number of independent sequencing library samples in which the clone appears, in a tissue.



Supplementary Figure 4

Clone Sizes Grouped by Clone Category.

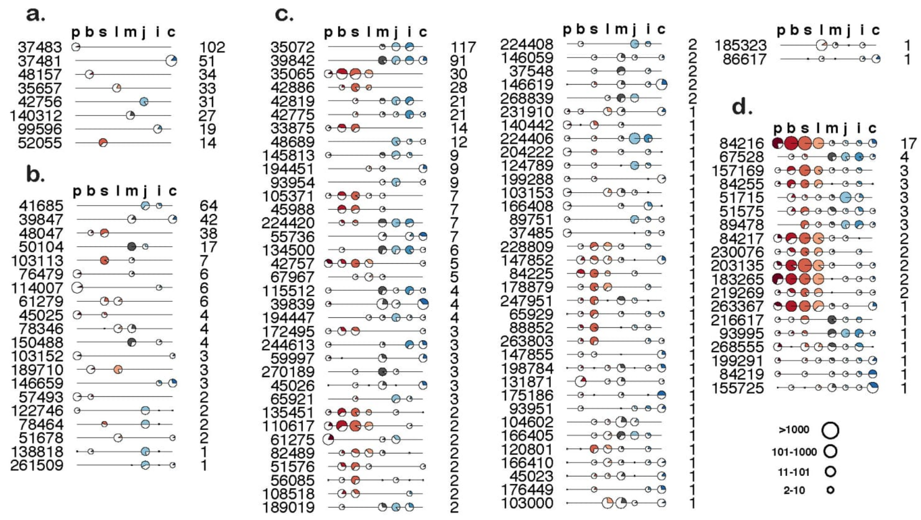
Sizes of large clones (in binned clone sizes of sequence instances, see **Methods**), grouped by tissue distribution category in D181 (left) and D207 (right).



Supplementary Figure 5

Large Clone Tissue Distribution Types in D207

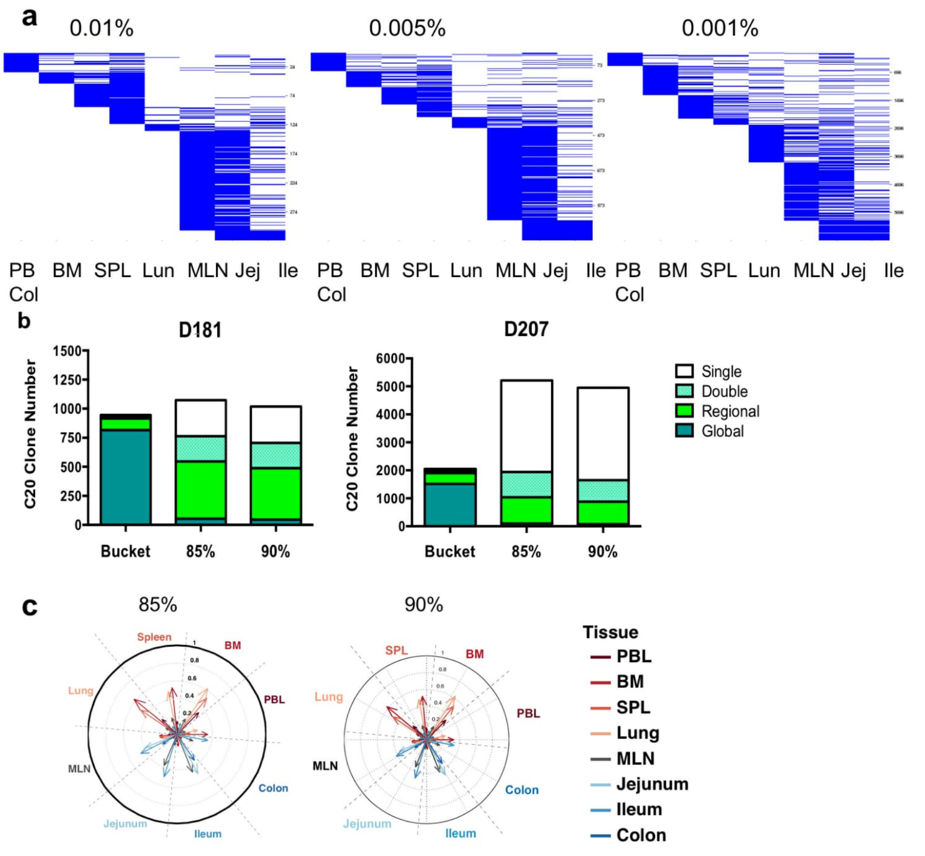
C$_{20}$ clones (including clones in the blood): ***a***, Single tissue clones; ***b***, Two-tissue clones; ***c***, Regional (found in 3-5 tissues) and ***d***, Global clones (found in 6-8 tissues). Tissues that contain only a single instance of a clone are not counted towards the distribution pattern of that clone. Each line is a clone. The position along the line represents a tissue. The size of the circles at each position represents the total number of instances the clones have in each tissue. The colored wedges represent the fraction of sequencing libraries that contain at least one instance of the clone in each tissue. A sample clone is shown for each tissue distribution pattern. The number to the left of each line is that clone's unique clone ID in http://immunedb.com/tissue-atlas. The number to the right is the number of different clones with this tissue distribution pattern. p = peripheral blood; b = bone marrow; s = spleen; l = lung; m = mesenteric lymph node; j = jejunum; i= ileum; c = colon.



Supplementary Figure 6

Large Clone Tissue Distribution Types in D181.

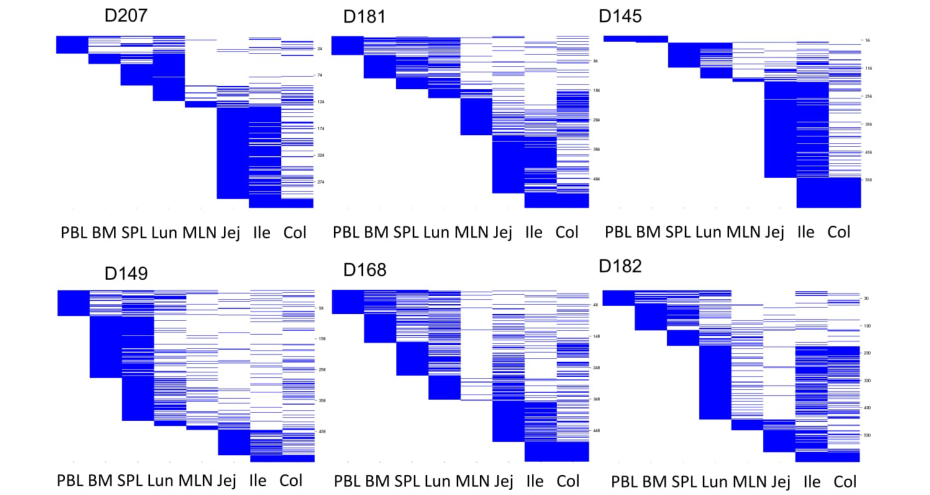
C_{20} clones (including clones in the blood): ***a***, Single tissue clones; ***b***, Two-tissue clones; ***c,*** Regional (found in 3-5 tissues) and ***d***, Global clones (found in 6-8 tissues). Tissues that contain only a single instance of a clone are not counted towards the distribution pattern of that clone. Each line is a clone. The position along the line represents a tissue. The size of the circles at each position represents the total number of instances the clones have in each tissue. The colored wedges represent the fraction of sequencing libraries that contain at least one instance of the clone in each tissue. A sample clone is shown for each tissue distribution pattern. The number to the left of each line is that clone's unique clone ID in http://immunedb.com/tissue-atlas . The number to the right is the number of different clones with this tissue distribution pattern. p = peripheral blood; b = bone marrow; s = spleen; l = lung; m = mesenteric lymph node; j = jejunum; i= ileum; c = colon.



Supplementary Figure 7

Effects of Clone Size Thresholding on Clone Tissue Membership.

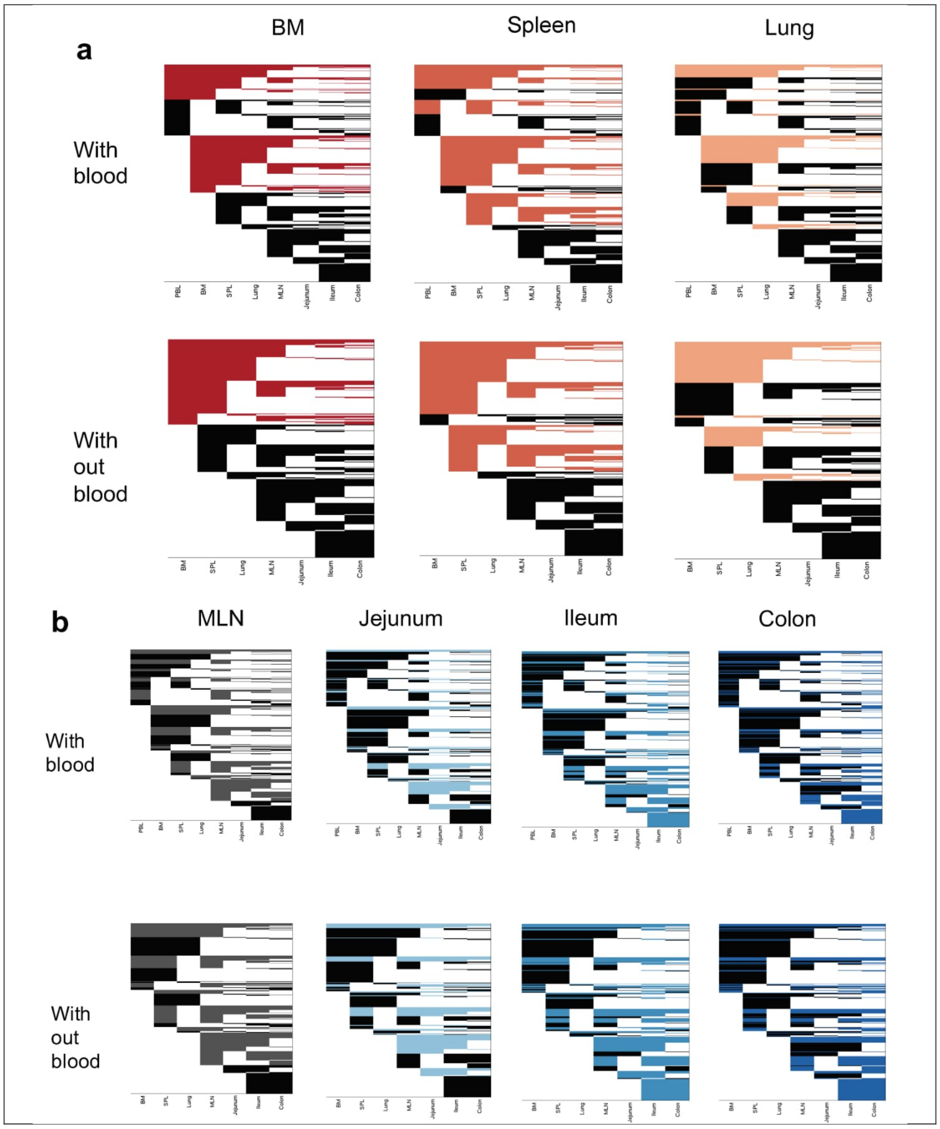
a, Similar blood and GI tract networks are seen at different clone size cut-offs. Overlapping clones are plotted at different clone size cut-offs (greater than or equal to 0.01%, 0.005% and 0.001% of total copies of all of the sequencing libraries within each donor). Each horizontal line represents a clone. Networks of overlapping clones were visualized using MiXCR and VDJtools software (see **Methods**). Only clones from D207 that overlapped between two or more tissues are shown. PBL = peripheral blood; BM = bone marrow; SPL = spleen; Lun = lung; MLN = mesenteric lymph node; Jej = jejunum; Ile = ileum; Col = colon. ***b,*** Breadth of tissue membership using different CDR3 identity thresholds. The tissue distribution of C_{20} clones in the two most deeply sequenced donors was studied using: 85% amino acid identity in CDR3 (the cut-off used in the main manuscript), 90% amino acid identity and no requirement for any CDR3 identity ("bucket"). All of the sequences were collapsed at these different CDR3 similarity thresholds into clones. Then C_{20} clones were analyzed for their tissue representation: single tissue, double (2 tissues), regional (3-5 tissues) and global (6-8 tissues). The C_{20} clones were recalculated for each threshold. ***c,*** Comparable partitioning of clones from D207 into blood and GI tract networks at two different CDR3 amino acid similarity thresholds. Networks of C_{20} clones with 85% CDR3 identity (left image is from **Fig. 1c** in the main manuscript) and 90% CDR3 amino acid identity (right panel) were analyzed using the cosine statistic (see **Methods**). Cosine similarity for two-tissue comparisons is plotted using the color scheme defined in the legend.



Supplementary Figure 8

Blood and GI Tract Networks are Present in All Donors.

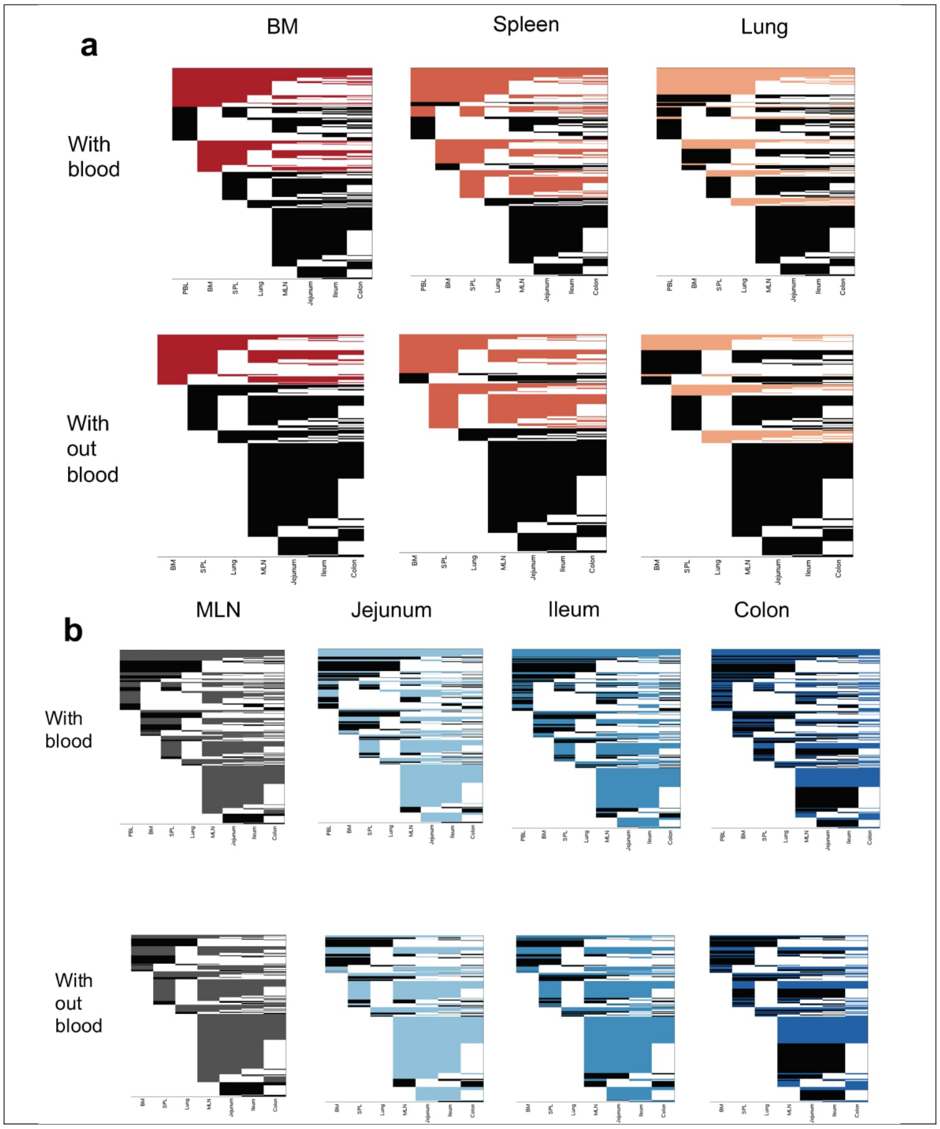
Clonal overlap was evaluated in all donors with clones that comprised 0.01% or more of total copies. Each horizontal line represents a clone. Networks of overlapping clones were visualized using MiXCR and VDJtools software (see **Methods**). Only clones that overlapped between two or more tissues are shown. The MLN sample in D168 was hypocellular. PBL = peripheral blood; BM = bone marrow; SPL = spleen; Lun = lung; MLN = mesenteric lymph node; Jej = jejunum; Ile = ileum; Col = colon.


a
BM
Spleen
Lung
With blood
PBL
BM
SPL
Lung
MLN
Jejunum
Ileum
Colon
With out blood
BM
SPL
Lung
MLN
Jejunum
Ileum
Colon
b
MLN
Jejunum
Ileum
Colon
With blood
PBL
BM
SPL
Lung
MLN
Jejunum
Ileum
Colon
With out blood
BM
SPL
Lung
MLN
Jejunum
Ileum
Colon

Supplementary Figure 9

Clonal Lineage Networks in D207 with and without Blood Clones

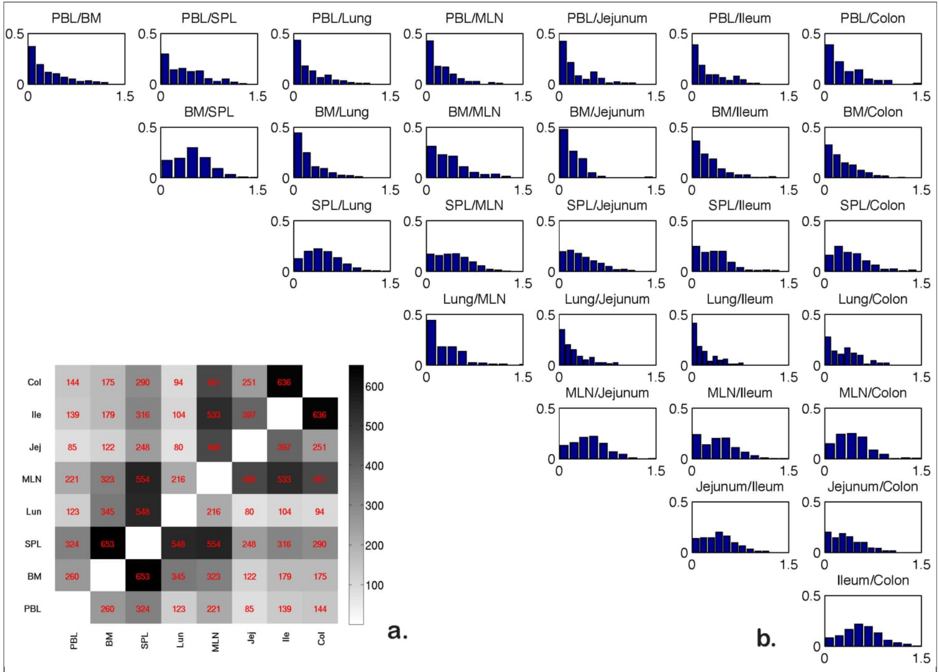
Each horizontal line represents a clone with at least 20 unique sequence instances in one tissue that overlaps with at least one other tissue, each column represents a tissue. All sub plots are identical, but are colored differently to highlight clones that reside in different tissues. In each instance, the clones that overlap in a specific tissue are colored as indicated by the tissue listed at the top of each panel. All other clones are denoted in black. ***a,*** Blood tissue clones highlighted; ***b,*** GI tract clones highlighted. Top row: all C_{20} clones. Bottom row: blood C_{20} clones computationally excluded (any clone with sequences found in the blood are removed from the analysis). Clones are ordered by their tissue membership.


a
BM
Spleen
Lung
With blood
PBL
BM
SPL
Lung
MLN
Jejunum
Ileum
Colon
With out blood
BM
SPL
Lung
MLN
Jejunum
Ileum
Colon
b
MLN
Jejunum
Ileum
Colon
With blood
PBL
BM
SPL
Lung
MLN
Jejunum
Ileum
Colon
With out blood
BM
SPL
Lung
MLN
Jejunum
Ileum
Colon

Supplementary Figure 10

Clonal Lineage Networks in D181 with and without Blood Clones.

Each horizontal line represents a clone with at least 20 unique sequence instances in one tissue that overlaps with at least one other tissue, each column represents a tissue. All sub plots are identical, but are colored differently to highlight clones that reside in different tissues. In each instance, the clones that overlap in a specific tissue are colored as indicated by the tissue listed at the top of each panel. All other clones are denoted in black. ***a,*** Blood tissue clones highlighted; ***b,*** GI tract clones highlighted. Top row: all C_{20} clones. Bottom row: blood C_{20} clones computationally excluded (any clone with sequences found in the blood are removed from the analysis). Clones are ordered by their tissue membership.



Supplementary Figure 11

Sequence Mixing within Clonal Lineages.

a (inset), Numbers of clones used for the clumpiness analysis. Shown are the numbers of clones with nodes in each pair of tissues. This analysis is restricted to C_{20} clones in D207; these numbers correspond to the clones analyzed in **Fig. 3b**. ***b,*** Sequence mixing within clonal lineages varies between different tissue types. The frequency distributions of clumpiness values between two tissues for all clones with nodes in each tissue pair (combinations are listed at the top of each plot). This analysis was performed on C_{20} clones in D207.

Discussion

Summary

The diversity of cells and clones in the B-cell repertoire is key to the immune system's ability to defend against pathogens by responding specifically to the antigens they produce [3-5]. It has therefore been the underlying theme of my thesis that analyzing repertoire diversity will lead to a greater understanding of the different states of immune system and its modes of function. In the previous sections I presented a series of tools that can computationally analyze the diversity of the human B-cell repertoire, and, with some modification, immune repertoires in general. My immune repertoire diversity analyses consists of three major aims: (1) identification of the germline V gene source; (2)quantifying competition between clones (clonal shift), and (3) quantifying competition within clones (clonal drift).

I have also applied the tools that I and others have developed to show that:

(1) Most VH alleles and some genes cannot be discriminated at a confidence above our threshold of statistical significance from other members of the same VH gene/family, even at low levels of somatic hypermutation [41].

(2) That descriptions of VH footprints in normal repertoires can be explained as sampling artifacts resulting from VH similarities in FWR3 and the non-templated nucleotide patterns at the V(D) (N1) and DJ (N2) junctions [42].

(3) Tissue-based B-cell repertoires of humans segregate between gut and blood-rich tissues. The clones that are found in the gut are more mixed between the different component tissues

(jejunum, colon, ileum). Additionally, B cells found in the tissues and most especially the jejunum are highly mutated [43].

(4) B-cell depletion-resistant clones in Sjogren's syndrome exhibit chronic negative selection of mutations of their rearranged antibody genes [45].

In my thesis, I have created several computational tools for analysis and visualization of high throughput, immune-repertoire profiling data including:

(1) A tool to rapidly and reliably identify germline genes using conserved anchor positions in the V gene sequence.

(2) A tool to estimate the probability of confusing similar V genes and output "V-ties" based on sequence length and mutation frequency.

(3) A tool to calculate the additional number of samples needed to achieve certain coverage of repertoire and the diversity of the repertoire given different numbers of sub-samples, as well as visualizing the results.

(4) A tool to calculate the cosine similarity between different compartments of the human body. The tool can be used on any two-sample comparison such as tissue, B-cell subset, etc.

(5) A tool to compare germline sequences to rearranged V region sequences, and output mutations and their frequencies in a clone.

High throughput sequencing provides a means to sequence large numbers of clones at low cost. But the large quantity of data generated requires computational tools for analyzing data in a convenient and efficient way. All of the tools I have developed are available on my github site at https://github.com/bochaozhang and on our lab github site at https://github.com/DrexelSystemsImmunologyLab. The integration of some of my tools into our analysis pipeline permits an integrated analysis and visualization of the immune repertoire, as well as analysis of clones across different data sets. It also allows for modification of different analysis parameters, such as how clones are defined and how clone sizes are calculated. These tools are fundamental to all immune repertoire studies and have been integrated into ImmuneDB [40], providing an easy and reliable way to analyze large quantities of sequencing data and describe immune repertoire diversity at the sequence and clonal levels.

Identification of the nearest germline V gene

Identification of the nearest corresponding germline V gene is the crucial first step in many analyses of gene rearrangement data. In general we consider the V(D)J recombination of B cell unique to a clone. For this reason the first step of clonal analysis is the association of each B cell receptor sequence to the germline genes segments that comprise its V(D)J or VJ recombination [46]. Making such associations and estimating the accuracy of identification is difficult because germline genes can be highly similar [17] and may undergo somatic mutation [47]. In addition, while high-throughput sequencing methods generate large numbers of sequences at a low cost, providing a way to essentially map the immune repertoire, they can have high sequence error rates. For this reason, it is critical to categorize, as much as possible, the reason for uncertainty in germline gene assignment. What is error and what is a-priori uncertainty? Some germline V genes cannot be well discriminated beforehand. However, they are not completely unknowable as they are similar only to other specific V genes and can be discriminated from most other germline V genes.

I showed that my Anchor method is the fastest way to identify clones, as it only relies on a string comparison for the initial alignment. In the minority of sequences that lack the anchor sequence, more time-consuming alignments can be deployed such as local alignment or IgBLAST with a short window length. After the first round of germline V gene identification with my Anchor method, I estimated the expected variability in our data. Based on the alignment length and mutation frequency of the first-round identification, I could calculate the likelihood of error due to mutation (or other sequence changes) using a simple hypergeometric test. V genes that I calculate as being impossible to distinguish at a given length/mutation/error rate will be identified as coming from one of several potential germline sources. Such V genes that cannot be discriminated are called V-ties. When constructing clones in later steps these V genes will be put in the same clone as they are indistinguishable. One of the advantages of assigning a germline identified at a level above that of the gene but below that of the family (i.e. can be one of a small set of genes from the same VH family) is helpful when determining mutations later. We can ignore the positions that differ between the confused genes and which remain unassigned and focus only on positions where we can identify somatic change.

V-ties are also helpful to germline gene identification as they can reduce noise in mutations and give more confidence to identification. For example, not using V-ties could lead to an over-estimation of the clonal diversity. Sequences with same CDR3 and J gene, and V genes that may be confused with others may be put in different clones when using traditional methods of germline assignment. But if these sequences were first assigned V ties as germlines they would be put in the same clones with an appropriate ambiguous but clearly defined V gene assignment. This can further reduce the noise caused by misidentification. Hence, the use of V-ties and my Anchor method yields faster and more reliable identifications while determining at what level of categorization (family, gene or allele) the identifications are definitive [41].

VH footprints

Another stage in receptor formation is receptor editing [48]. It has been observed that in some cases a non-functional V(D)J recombination or one that is self-reactive can be saved by switching out the V gene segment [49]. In some cases this switching out step leaves a “footprint” of the old VH gene at the 3’ end of the new VH gene [27]. Given my previous analysis of V gene similarities, I suspected that often identification of such footprints could be confounded by the high level of similarity of the FWR3 of different VH [41, 50] and the heightened induced variability of CDR3 [51]. Therefore, I asked to what extent these observed footprint patterns could be explained by chance differences in the FWR3 and CDR3. To answer this question, I compared the patterns of footprints in in-frame and out-of-frame V region sequences derived from 42,221 unique sequences of peripheral blood B-cells from a healthy human adult subject. In both in-frame and out-of-frame CDR3s there was a positive correlation between the number of footprints observed and CDR3 length. Moreover, footprints were found at both junctions of V(D) and (D)J. Finally, the occurrence of footprints from in-frame and out-of-frame CDR3s fit well with a Poisson distribution. Therefore, we concluded that the patterns identified as VH footprints may be the results of random noise. Thus, while VH replacement certainly occurs [48], footprint analysis is a poor measure of its frequency at least in normal immune repertoires [42].

Quantification of competition between clones

With the identification of germline genes and the establishment of clones, I can study the differential selection and competition of clones. Differences in the form of clonal competition can influence the clonal makeup of B-cell repertoires in different regions of the human body [52]. The tissue distribution and trafficking of B-cell clones influences how infections are controlled throughout the body. Animal studies indicate that tissue localization of B-cells and plasma cells is important for protective immunity and homeostasis of bacterial microflora [53-55]. However, tissue-based B-cell subsets are not well understood in humans. Unlike laboratory mice, humans are outbred, and live for decades in diverse environments with exposures to many different antigens and pathogens. Furthermore, most studies of human B cells have sampled only the blood or tonsils; the latter are often inflamed when removed, hardly a physiologic specimen. Consequently, how clones are localized to specific regions or tissues in the human body is not well understood for B cells. To make a snapshot of this localization, I focused on the level of B-cell clonal overlap among different tissues.

To truly describe difference in overlap we must be certain that clones are not missed due to lack of sampling. Proving a negative is impossible, the next best thing we can do is to estimate our sufficiency of sampling. To do so I chose to perform by sample rarefaction analysis [31]. Since the unit of sampling is not an individual clone but a sample of clones from a single tissue or anatomic compartment, it is best to use sample-based rarefaction. Applying this method, I found that an additional 5-70 samples were needed to reach 0.75 coverage with clones of at least two sequence instances in D207, our most deeply sequenced donor. Peripheral blood, as expected, had the most diverse repertoire (perhaps in part because there is extensive mixing in the blood), needed more than 150 additional samples to reach 0.75 coverage. Such numbers of samples were impossible to acquire. Therefore, we instead opted to exclude the smaller clones from our analyses. The immune repertoire has a much greater number of small clones than large clones [56]. By increasing the clone size threshold and excluding smaller clones, the repertoire diversity was substantially reduced and it was easier to achieve adequate coverage. As a result, with clones of at least 20 sequence instances, most tissues were sufficiently sampled in D207 to reach 0.75 coverage.

The next question was how to quantify the distribution of B-cell clones and map them to different tissues. Simply counting the number of clones would neglect the fact that some clones were more expanded and may play a more important role in the repertoire composition. To answer this question, I used cosine similarity to calculate the normalized number of shared clones between each sample. The cosine similarity not only takes the number of shared clones but also their sizes into account. To take into account the difference in the numbers of sequencing libraries, cosine similarity is calculated with both the number of sequence instances and the number of sequencing libraries as a vector, to address the similarity between tissues. Cosine similarity gives more weight to clones that are more expanded or found in multiple replicate samples and the they will contribute more to the final score. Also, cosine similarity always gives score range from 0 to 1, where 0 means no similarity at all and 1 means completely similarity, providing a normalized score for better visualization and comparison of different tissue pairs within the same figure. Through this analysis, I revealed how clones are distributed across the different tissue pairs.

Analysis of similarities in the B-cell clonal makeup in the blood (PBL), bone marrow, spleen, lung, mesenteric lymph node (MLN), jejunum, ileum and colon, comparing between each pair of tissues, revealed two prominent networks of expanded clones in the six donors. One network comprised the PBL, bone marrow, lung and spleen, and the other network consisted of the jejunum, ileum and colon. Clones in the MLN spanned both the blood-rich sites and the GI tract, but exhibited greater overlap with the GI tract. This supports the existence of two major networks of expanded B-cell clones, one in blood-rich tissues and a separate network in the GI tract. To rule out blood contamination, clones with sequences in PBL samples were removed. The network of overlapping clones within the blood-rich sites was more diminished than the GI tract network.

A limitation of using cosine similarity to assess overlap is that it only takes into account the relationship of two tissues at a time. Clones can and do span more than two tissues. Roughly, the clones can be divided into global (found in six to eight tissues), regional (three to five tissues), two-tissue and single-tissue clones. I visualized the clonal networks using line circle plots that I developed. Both the regional and two-tissue clones echoed the patterns of overlap observed in paired tissue analysis: the clones were usually present in either the GI tract or the blood-rich tissues. Even global clones that spanned essentially all tissues and were generally the largest clones, were never as expanded in both regions. Thus even a clone that is found in both gut and blood-rich tissues is clearly more expanded in only one of them. This I feel confirms the existence of two distinct networks even in more expanded clones that we inferred from the pairwise comparisons. While it still remains an open question what forces are maintaining this segregation and how it relates to antigen specificity, it is clear that we must take it into account when querying the blood for tissue based and especially gut-based immune responses.

Quantification of clonal drift

One of the fascinating aspects of studying B-cell repertoires is the presence of somatic mutations within clonal lineages. Mutations can provide information about antigen experience and the extent of clonal expansion [57, 58]. Comparing the patterns of mutations in members of a clone opens a unique window into the relative history of the individual members of a clone and the overall selection pressures they undergo.

Different mutation and selection pressures can reflect the development of B cells as they age and encounter antigens in the human body. Using sequencing data from patients with Sjögren's syndrome (SjS), I identified large expanded clones that persisted in the blood of one patient despite the fact that the patient received B-cell depletion therapy with rituximab (anti-CD20). The expanded clones harbored large number of synonymous mutations but only a few, non-synonymous mutations, indicating that the expanded clones were under negative selection. This was corroborated by selection analysis. Using BaSELINE [44]. I found that both CDR and FWR exhibited negative selection and that in fact CDRs were the more negatively selected regions. The negative selection in the CDRs may have occurred because the clone had evolved to the point where it could no longer improve its affinity through mutation. Any further non-synonymous mutation thereby may lower the antibody affinity for its antigen, suggesting the clone has high affinity to a specific antigen. Alternatively the receptor is either highly cross reactive or only needs weak but constant activation. Any change in receptivity (through the change of even a single amino acid) might abrogate the functional competence of the receptor.

The analysis of somatic hypermutation also enhanced our understanding of how individual cells compete within clones. One of the major tools we used to analyze the patterns of mutation in a clone was the construction of clonal lineages. Clonal lineage analysis allows us to identify mutations that are especially important in a clone -- for instance, those that are shared by all of the members of a clone (and therefore are found in the trunk of the tree) versus those that are very recent or only appear in some parts of the lineage tree (such as the leaves). One of the issues with studying mutations is how to differentiate between real mutations and artificial mismatches from germline. They could be the result of misidentifying germline, not having the correct germline allele in the germline sequence library and/or sequencing error. To rectify this, we both considered V-ties and filtered out sequences and mutation variants that only occurred once in a person. Despite these corrections, it is clear that many of our mutations may still be errors. For this reason, even as we constructed lineages we only took into account gross structures of the mutation relationships. Even with these severely constrained data sets, several interesting patterns arose. Looking at the overall level of mutation, clones with higher levels of mutation were more frequent in the GI tract tissues than in the PBL or the bone marrow, suggesting that there are more memory B-cells and/or plasma cells (terminally differentiated B cells that secrete antibodies) in the tissues. When we wanted to see which tissues had clones that had expanded the most we measured their height. To cancel out the impact of selection, I also looked at the number of four-fold degenerate synonymous mutations. Counting four-fold synonymous mutations along the longest branches of the lineages in clones, we found that clones from the colon had substantially more of these mutations than clones in lung, spleen, PBL or bone marrow. From this analysis we proposed that the gut clones underwent more immune responses, which would be consistent with them having undergone the greatest number of divisions while the mutation process was engaged [59].

As a final step in analyzing the lineages in different tissues, we used a clumpiness measure estimate the degree of intermingling of sequence variants among different tissues within the leaves of the individual clonal lineages. Clumpiness is a measurement of the tendency of two tissue types to be close on a given lineage tree [39]. In the lineage tree of a clone that spans two or more tissues, a higher clumpiness value indicates that members of the clone exhibit more intermingling (mixing and overlapping of sequence variants in the different tissues). We found clones with PBL and another tissue are the least mixed, followed by clones mixing blood and GI tract tissues, then blood tissue clones and, finally, GI tract clones, which were the most mixed. The fact that identically mutated sequences can often be found at multiple GI sites can be explained if mutated clones disseminate throughout the GI tract, yet also undergo serial rounds of mutation and selection.

Summary and Future plans

The study of B cell sequence repertoires in this thesis has opened new ways to study their diversity. At every level of analysis, there are distinctions that we cannot make a priori. However, we can quantify our uncertainty so that at least what we do know and what we do not know are clearly defined. The tools I developed provide investigators with the means to analyze and visualize high-throughput, immune-repertoire profiling data. Given our current ways of sampling the repertoire, there are differences in germline association, and thus clonal relationships, that cannot be identified. On the other hand, they are not completely unknowable as germline V genes are similar only to other specific V genes and can be discriminated from most other germline V genes. Based on the alignment length and mutation frequency of the first-round identification, we can calculate the likelihood of misidentification. V genes that we calculate as being impossible to distinguish at a given length/mutation/error rate will be identified as V-ties. Then these V genes will be put in the same clone as they are indistinguishable later when constructing clones. In this way, one can have more reliable identifications and know at what level of categorization (family, gene or allele) the identifications are definitive. At the level of clonal competition, the sampling coverage of the repertoire is often unknown. Sample-based rarefaction analysis provides a way to estimate how well we have sampled the repertoire and can help us quantify the likelihood of false negatives. Using these tools I identified two major networks B cell clones in the blood-rich and GI tract tissues in human body. The clarity with which we could identify these two networks of expanded clones leads to the next step of research where we will ask if they are the result of different selection pressures in mucosal and lymphoid tissues and intestinal tissues or to some differences in circulation or lymphocyte trafficking? Or do these networks arise because of some other hitherto undescribed environmental effect such as different commensal bacteria?

This atlas of clones also provides normative data for future studies of tissue-specific and response-specific maps of repertoire in the immune response to antigens. At the level of competition within clones, mutation analyses showed that the GI clones were the most mutated, which is consistent with them having undergone the greatest number of immune responses. Here again, using specifically designed tools and computational approaches to quantify relationships of B cells within clonal lineages, selection was found to act differently in the gut and blood tissue clones. Using clumpiness analysis, jejunum, ileum and colon were found to have more mixing while blood tissues developed along separate lineage lines.

One of the obvious future plans is to analyze the directional relationship between GI-tract tissues and blood-rich tissues. We have already made the clones into lineage trees and labeled the nodes by tissue. Our assumption was the naïve B cells circulating in blood would reside in gut and go through somatic mutation and selection locally. It would be interesting to see if in a mixed tissue tree (one that has both blood nodes and gut nodes) the gut nodes are always descendants of blood nodes. I have tried this with our current data, but there were not enough trees that had both blood and gut nodes, and were complex enough to yield a meaningful data for a robust analysis. With more data, the relationship may be clear enough to give significant results. Such an analysis would yield insights into the direction of trafficking and/or maturation of B cell clones as they flow through the body.

Another way to study trafficking is to sample tissues and/or blood at multiple time points. However, with humans, such analysis in tissues would likely be limited to unique circumstances, such as an individual undergoing serial monitoring biopsies of the GI tract for evaluation of inflammation, malignancy or transplantation tolerance. Longitudinal tracking of clones could yield interesting insights into how B cell subsets develop and/or are maintained in humans. This could be performed in blood or, better yet, using blood and bone marrow samples from the same individuals. Another area where longitudinal clone tracking could provide interesting insights is in autoimmune disease, where large clones may be pathogenic. Using the tools I developed to computationally identify and measure the sizes of clones, we could track these clones through different time points in a patient with a B-cell autoimmune disease such as Sjogren's syndrome, and determine if the size of large persistent clones correlate with disease activity or response to therapy. Tracking clones also has obvious importance in the clinical evaluation of patients with B-cell malignancies such as acute lymphoblastic leukemia, chronic lymphoblastic leukemia, lymphoma and multiple myeloma. My analysis of VHR footprints could be extended to evaluate diversification of acute lymphoblastic leukemia clones by VHR, improving the ability to not only diagnose but also monitor disease (particularly in patients where there is evolution of the ALL B cell clone by further rearrangement).

Another approach to studying B cell maturation and function is to use flow cytometry to evaluate the immunophenotype of the cells and then corroborate the phenotypic data with sequence analysis by sequencing sorted subsets of B cells. For example, one can distinguish memory B cells from naïve B cells through the expression of the cell surface marker CD27 and the presence of class switching (from IgM to other heavy chain isotypes). One can also look for plasma cells, which have a distinctive phenotype. This type of analysis is under way, in collaboration with the Shlomchik laboratory (University of Pittburgh). Consistent with our prediction that the GI tract has more memory B cells, we observe more B cells that express CD27 and exhibit evidence of class switching (IgG or IgA+) in the GI tract, compared to the peripheral blood. This finding is also consistent with the literature, in which class-switched B cells and plasma cells are more abundant in the tissues. Coupling immunophenotype to the BCR or TCR is just the beginning. With the advent of single cell approaches, it may be possible in the very near future to integrate the phenotype, BCR or TCR (with paired chain sequencing) and other features of the cell, such as its transcriptome or epigenome [60].

Bibliography

1. Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, Ni I, Mei L, Sundar PD, Day GM, Cox D, Rajpal A, Pons J, *Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire.* Proc Natl Acad Sci USA, 2009. 106(48): 20216-20221.
2. Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, Riddell SR, Warren EH, Carlson CS, *Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells.* Blood, 2009. 114(19): 4099-4107.
3. Tonegawa S, *Somatic generation of antibody diversity.* Nature, 1983. 302(5909): 575-581.
4. Kronenberg M, Siu G, Hood LE, Shastri N, *The molecular genetics of the T-cell antigen receptor and T-cell antigen recognition.* Annu Rev Immunol, 1986. 4: 529-591.
5. Mauri C, Bosma A, *Immune regulatory function of B cells.* Annu Rev Immunol, 2012. 30: 221-241.
6. Schroeder HW Jr, Cavacini L, *Structure and function of immunoglobulins.* J Allergy Clin Immunol, 2010. 125(2 Suppl 2): S41 – S52.
7. Gibson KL, Wu YC, Barnett Y, Duggan O, Vaughan R, Kondeatis E, Nilsson BO, Wikby A, Kipling D, Dunn-Walters DK, *B-cell diversity decreases in old age and is correlated with poor health status.* Aging cell, 2009. 8(1): 18-25.
8. Sherwood AM, Emerson RO, Scherer D, Habermann N, Buck K, Staffa J, Desmarais C, Halama N, Jaeger D, Schirmacher P, Herpel E, Kloor M, Ulrich A, Schneider M, Ulrich CM, Robins H, *Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of T cell receptor sequences that differ from the T cells in adjacent mucosal tissue.* Cancer Immunol Immunother, 2013. 62(9): 1453-1461.
9. Brown SD, Hapgood G, Steidl C, Weng AP, Savage KJ, Holt RA, *Defining the clonality of peripheral T cell lymphomas using RNA-seq.* Bioinformatics, 2017. 33(8): 1111-1115.

10. Watanabe T, *Adult T-cell leukemia: molecular basis for clonal expansion and transformation of HTLV-1-infected T cells.* Blood, 2017. 129(9): 1071-1081.

11. Paiva B, Martinez-Lopez J, Corchete LA, Sanchez-Vega B, Rapado I, Puig N, Barrio S, Sanchez ML, Alignani D, Lasa M, García de Coca A, Pardal E, Oriol A, Garcia ME, Escalante F, González-López TJ, Palomera L, Alonso J, Prosper F, Orfao A, Vidriales MB, Mateos MV, Lahuerta JJ, Gutierrez NC, San Miguel JF, *Phenotypic, transcriptomic, and genomic features of clonal plasma cells in light-chain amyloidosis.* Blood, 2016. 127(24): 3035-3039.

12. Barrio S, Shanafelt TD, Ojha J, Chaffee KG, Secreto C, Kortüm KM, Pathangey S, Van-Dyke DL, Slager SL, Fonseca R, Kay NE, Braggio E, *Genomic characterization of high-count MBL cases indicates that early detection of driver mutations and subclonal expansion are predictors of adverse clinical outcome.* Leukemia, 2017. 31(1): 170-176.

13. Mitchell R, Kelly DF, Pollard AJ, Trück J, *Polysaccharide-specific B cell responses to vaccination in humans.* Hum Vaccin Immunother, 2014. 10(6): 1661-1668.

14. Alt FW, Baltimore D, *Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-JH fusions.* Proc Natl Acad Sci USA, 1982. 79(13): 4118-4122.

15. Petrie HT, Livak F, Burtrum D, Mazel S, *T cell receptor gene recombination patterns and mechanisms: cell death, rescue, and T cell production.* J Exp Med, 1995. 182(1): 121-127.

16. Gellert M, *V(D)J recombination: RAG proteins, repair factors, and regulation.* Annu Rev Biochem, 2002. 71: 101-132.

17. Lefranc MP, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G, *IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains.* Dev Comp Immunol, 2003. 27(1): 55-77.

18. Desiderio SV, Yancopoulos GD, Paskind M, Thomas E, Boss MA, Landau N, Alt FW, Baltimore D, *Insertion of N regions into heavy-chain genes is correlated with expression of terminal deoxytransferase in B cells.* Nature, 1984. 311(5988): 752-755.

19. Lieber MR, Ma Y, Pannicke U, Schwarz K, *The mechanism of vertebrate nonhomologous DNA end joining and its role in V(D)J recombination.* DNA Repair (Amst), 2004. 3(8-9): 817-826.

20. Luning Prak ET, Monestier M, Eisenberg RA, *B cell receptor editing in tolerance and autoimmunity.* Ann NY Acad Sci, 2011. 1217: 96-121.

21. Martin A, Bardwell PD, Woo CJ, Fan M, Shulman MJ, Scharff MD, *Activation-induced cytidine deaminase turns on somatic hypermutation in hybridomas.* Nature, 2002. 415(6873): 802-806.

22. Uduman M, Yaari G, Hershberg U, Stern JA, Shlomchik MJ, Kleinstein S, *Detecting selection in immunoglobulin sequences.* Nucleic Acids Res, 2011. 39(Web Server issue): W499-504.

23. Berglund EC, Kiialainen A, Syvänen AC, *Next-generation sequencing technologies and applications for human genetic history and forensics.* Investig Genet, 2011. 2: 23.

24. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M, *Comparison of next-generation sequencing systems.* J Biomed Biotechnol, 2012. 2012: 251364.

25. Ye J, Ma N, Madden TL, Ostell JM, *IgBLAST: an immunoglobulin variable domain sequence analysis tool.* Nucleic Acids Res, 2013. 41(Web Server issue): W34-40.

26. Brochet X, Lefranc MP, Giudicelli V, *IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis.* Nucleic Acids Res, 2008. 36: W503-W508.

27. Zhang Z, Wang YH, Zemlin M, Findley HW, Bridges SL, Burrows PD, Cooper MD, *Molecular mechanism of serial VH gene replacement.* Ann N Y Acad Sci, 2003. 987: 270-273.

28. McKean D, Huppi K, Bell M, Staudt L, Gerhard W, Weigert M, *Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin.* Proc. Natl Acad. Sci. USA., 1984. 81(10): 3180-3184.

29. Burnet FM, *A modification of Jerne's theory of antibody production using the concept of clonal selection.* CA Cancer J Clin, 1976. 26(2): 119-121.

30. Greaves M, Maley CC, *Clonal evolution in cancer.* Nature, 2012. 481(7381): 306-313.

31. Colwell RK, Chao A, Gotelli NJ, Lin S, Mao CX, Chazdon RL, Longino JT, *Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages.* Journal of Plant Ecology, 2012. 5(1): 3-21.

32. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C, *Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform.* Nucleic Acids Res, 2015. 43(6): e37.

33. Shlomchik MJ, Aucoin AH, Pisetsky DS, Weigert MG, *Structure and function of anti-DNA autoantibodies derived from a single autoimmune mouse.* Proc Natl Acad Sci USA, 1987. 84(24): 9150-9154.

34. Chang B, Casali P, *The CDR1 sequences of a major proportion of human germline ig VH genes are inherently susceptible to amino acid replacement.* Immunol Today, 1994. 15(8): 367-373.

35. Lossos IS, Okada CY, Tibshirani R, Warnke R, Vose JM, Greiner TC, Levy R, *Molecular analysis of immunoglobulin genes in diffuse large b-cell lymphomas.* Blood, 2000. 95(5): 1797-1803.

36. Bose B, Sinha S, *Problems in using statistical analysis of replacement and silent mutations in antibody genes for determining antigen-driven affinity selection.* Immunology, 2005. 116(2): 172-183.

37. Yang Z, Bielawski JP, *Statistical methods for detecting molecular adaptation.* Trends Ecol Evol, 2000. 15(12): 496-503.

38. Hurst LD, *The Ka/Ks ratio: diagnosing the form of sequence evolution.* Trends Genet, 2002. 18(9): 486.

39. Schwartz GW, Shokoufandeh A, Ontanon S, Hershberg U, *Using a novel clumpiness measure to unite data with metadata: Finding common sequence patterns in immune receptor germline V genes.* Pattern Recognit Lett, 2016. 74: 24-29.

40. Rosenfeld AM, Meng W, Luning Prak ET, Hershberg U, *ImmuneDB: A system for the analysis and exploration of high-throughput adaptive immune receptor sequencing data.* Bioinformatics, 2017. 33(2): 292-293.

41. Zhang B, Meng W, Prak ET, Hershberg U, *Discrimination of germline V genes at different sequencing lengths and mutational burdens: a new tool for identifying and evaluating the reliability of V gene assignment.* J Immunol Methods, 2015. 427: 105-116.

42. Meng W, Jayaraman S, Zhang B, Schwartz GW, Daber RD, Hershberg U, Garfall AL, Carlson CS, Luning Prak ET, *Trials and tribulations with VH replacement.* Front Immunol, 2014. 5: 10.

43. Zhang B, Meng W, Schwartz GW, Rosenfeld AM, Ren D, Thome JJC, Carpenter DJ, Matsuoka N, Lerner H, Friedman AL, Granot T, Farber DL, Shlomchik MJ, Hershberg U, Luning Prak ET, *An atlas of B-cell clonal distribution in the human body.* Nat Biotechnol, 2017. 35(9): 879-884.

44. Yaari G, Uduman M, Kleinstein SH, *Quantifying selection in high-throughput Immunoglobulin sequencing data sets.* Nucleic Acids Res, 2012. 40(17): e134.

45. Hershberg U, Meng W, Zhang B, Haff N, St Clair EW, Cohen PL, McNair PD, Li L, Levesque MC, Luning Prak ET, *Persistence and selection of an expanded B-cell clone in the setting of rituximab therapy for Sjögren's syndrome.* Arthritis Res Ther, 2014. 16(1): R51.

46. Toby IT, Levin MK, Salinas EA, Christley S, Bhattacharya S, Breden F, Buntzman A, Corrie B, Fonner J, Gupta NT, Hershberg U, Marthandan N, Rosenfeld A, Rounds W, Rubelt F, Scarborough W, Scott JK, Uduman M, Vander Heiden JA, Scheuermann RH,

Monson N, Kleinstein SH, Cowell LG, *VDJML: a file format with tools for capturing the results of inferring immune receptor rearrangements.* BMC Bioinformatics, 2016. 17(Suppl 13): 333.

47. Berek C, Milstein C, *The dynamic nature of the antibody repertoire.* Immunol Rev, 1988. 105: 5-26.

48. Darlow JM, Stott DI, *V(H) replacement in rearranged immunoglobulin genes.* Immunology, 2005. 114(2): 155-165.

49. Zhang Z, Zemlin M, Wang YH, Munfus D, Huye LE, Findley HW, Bridges SL, Roth DB, Burrows PD, Cooper MD, *Contribution of Vh gene replacement to the primary B cell repertoire.* Immunity, 2003. 19(1): 21-31.

50. Schwartz GW, Hershberg U, *Conserved variation: identifying patterns of stability and variability in BCR and TCR V genes with different diversity and richness metrics.* Phys Biol, 2013. 10(3): 035005.

51. Bangs LA, Sanz IE, Teale JM, *Comparison of D, JH, and junctional diversity in the fetal, adult, and aged B cell repertoires.* J Immunol, 1991. 146(6): 1996-2004.

52. Rhiner C, Díaz B, Portela M, Poyatos JF, Fernández-Ruiz I, López-Gay JM, Gerlitz O, Moreno E, *Persistent competition among stem cells and their daughters in the Drosophila ovary germline niche.* Development, 2009. 136(6): 995-1006.

53. Kroese FG, de Waard R, Bos NA, *B-1 cells and their reactivity with the murine intestinal microflora.* Semin Immunol, 1996. 8(1): 11-18.

54. Rhee KJ, Sethupathi P, Driks A, Lanning DK, Knight KL, *Role of commensal bacteria in development of gut-associated lymphoid tissues and preimmune antibody repertoire.* J Immunol, 2004. 172(2): 1118-1124.

55. Lindner C, Thomsen I, Wahl B, Ugur M, Sethi MK, Friedrichsen M, Smoczek A, Ott S, Baumann U, Suerbaum S, Schreiber S, Bleich A, Gaboriau-Routhiau V, Cerf-Bensussan N, Hazanov H, Mehr R, Boysen P, Rosenstiel P, Pabst O, *Diversification of memory B*

cells drives the continuous adaptation of secretory antibodies to gut microbiota. Nat Immunol, 2015. 16(8): 880-888.

56. Desponds J, Mora T, Walczak AM, *Fluctuating fitness shapes the clone-size distribution of immune repertoires.* Proc Natl Acad Sci USA, 2016. 113(2): 274-279.

57. Rudikoff S, Pawlita M, Pumphrey J, Heller M, *Somatic diversification of immunoglobulins.* Proc Natl Acad Sci USA, 1984. 81(7): 2162-2166.

58. Sablitzky F, Wildner G, Rajewsky K, *Somatic mutation and clonal expansion of B cells in an antigen-driven immune response.* EMBO J, 1985. 4(2): 345-350.

59. Holtmeier W, Hennemann A, Caspary WF, *IgA and IgM V(H) repertoires in human colon: evidence for clonally expanded B cells that are widely disseminated.* Gastroenterology, 2000. 119: 1253 – 1266.

60. Su Y, Wei W, Robert L, Xue M, Tsoi J, Garcia-Diaz A, Homet Moreno B, Kim J, Ng RH, Lee JW, Koya RC, Comin-Anduix B, Graeber TG, Ribas A, Heath JR, *Single-cell analysis resolves the cell state transition and signaling dynamics associated with melanoma drug-induced resistance.* Proc Natl Acad Sci USA, 2017. 114(52): 13679-13684.