Submission: Article for Discoveries

Substitution and site-specific selection driving B cell affinity maturation is consistent across individuals

CONNOR O. MCCOY[1], TREVOR BEDFORD[1], VLADIMIR N. MININ[2], PHILIP BRADLEY[1], HARLAN ROBINS[1], FREDERICK A. MATSEN IV[1]

[1] *Program in Computational Biology, Fred Hutchinson Cancer Research Center, Seattle, WA, 91802, USA*

[2] *Department of Statistics, University of Washington, Seattle, WA 98195*

Corresponding author: Frederick A Matsen, Fred Hutchinson Cancer Research Center, Mail stop: M1-B514, Seattle, WA 98109-1024, USA; E-mail: matsen@fhcrc.org.

ABSTRACT

The antibody repertoire of each individual is continuously updated by the evolutionary process of B cell receptor mutation and selection. It has recently become possible to gain detailed information concerning this process through high-throughput sequencing. Here, we develop modern statistical molecular evolution methods for the analysis of B cell sequence data, and then apply them to a very deep short-read data set of B cell receptors. We find that the substitution process is conserved across individuals but varies significantly across gene segments. We investigate selection on B cell receptors using a novel method that side-steps the difficulties encountered by previous work in differentiating between selection and motif-driven mutation; this is done through stochastic mapping and empirical Bayes estimators that compare the evolution of in-frame and out-of-frame rearrangements. We use this new method to derive a per-residue map of selection, which we find is dominated by negative selection, though not uniformly so.

INTRODUCTION

Antibodies encoded by somatically modified human B cell receptor (BCR) genes bind a vast array of antigens, initiating an immune response or directly neutralizing their target. This diversity is made possible by the processes of *VDJ recombination*, in which random joining of V-, D-, and J-genes generates an initial combinatorial diversity of BCR sequences, and *affinity maturation*, which further modifies these sequences. The affinity maturation process, in which antibodies increase binding affinity for their cognate antigens, is essential to mounting a precise humoral immune response. Affinity maturation proceeds via a nucleotide substitution process that combines Darwinian mutation and selection processes. Mutational diversity is generated by *somatic hypermutation* (SHM), in which a targeted molecular mechanism mutates the BCR sequence. This diversity is then passed through a selective sieve in which B cells that bind well to antigen are stimulated to reproduce, while those that do not bind well or bind to self are marked for destruction. The combination of VDJ recombination and affinity maturation enables B cells to respond to an almost limitless diversity of antigens. Understanding the substitution process and selective forces shaping the diversity of the memory B cell repertoire thus has implications for disease prophylaxis and treatment, including for the promise of rational vaccine design.

It has recently become possible to gain detailed information about the B cell repertoire using high-throughput sequencing (Boyd *et al.*, 2009; Wu *et al.*, 2010; Larimore *et al.*, 2012; DeKosky *et al.*, 2013; Robins, 2013). Recent reviews have highlighted the need for new computational tools that make use of BCR sequence data to bring new insight, including the need for reproducible computational pipelines (Mehr *et al.*, 2012; Six *et al.*, 2013; Warren *et al.*, 2013; Georgiou *et al.*, 2014). Rigorous analysis of the B cell repertoire will require statistical analysis of how evolutionary processes define affinity maturation. Statistical nucleotide molecular evolution models are often described in terms of three interrelated processes: mutation, the process generating diversity, selection, the process determining survival or loss of mutations, and substitution, the observed process of evolution that follows from the first two processes. Although researchers have made many observations concerning the affinity maturation process, this work has not yet been done using rigorous statistical criteria. One major vein of research has focused on how

nucleotide mutation rates depend on the identity of surrounding nucleotides (reviewed in Delker *et al.* (2009); see also Yaari *et al.* (2013)), but little has been done concerning other aspects of the process, such as how the substitution process differs between gene segments.

Along with mutation, selection due to competition for antigen binding forms the other key part of the affinity maturation process. Inference of selective pressures in this context is complicated by nucleotide context-dependent mutation, leading some authors to proclaim that such selection inference is not possible (Dunn-Walters and Spencer, 1998). Indeed, if one does not correct for context-dependent mutation bias, interactions between those motifs and the genetic code can lead to false positive identification of selective pressure. Previous work has developed methodology to analyze selection on sequence tracts in this context (reviewed in Discussion), but no methods have yet achieved the goal of rigorous per-residue selection estimates. This has, however, been recently identified as an important goal (Yaari *et al.*, 2013). Such per-residue estimates of selection would form a foundation for rational vaccine design by giving information on to what degree residues should be considered mutable versus being so essential to structure that they cannot be mutated.

The ensemble of germline V, D, and J genes that rearrange to encode antibodies (equivalently: immunoglobulins) are divided into nested sets. They can first be identified by their *locus*: IGH, denoting the heavy chain, IGK, denoting the kappa light chain, or IGL, denoting the lambda light chain. Our dataset contains solely the IGH locus, so we will frequently omit the locus prefix for simplicity. Genes within a locus can be first subdivided by their *segment*, which is whether they are a V, D, or J gene. IGHV genes are further divided into *subgroups* which share at least 75% nucleotide identity. Genes also have polymorphisms that are grouped into *alleles*, which represent polymorphisms of the gene between individuals (Lefranc *et al.*, 2008).

VDJ recombination does not always produce a functional antibody, such as when the V and J segments are not in the same reading frame after recombination (an *out-of-frame* rearrangement) or when the receptor sequence contains a premature stop codon. However, each B cell carries two copies of the IGH locus, with one on each chromosome. If the rearrangement on the first locus fails to produce a viable antibody, the second locus will rearrange; if this second rearrangement is successful, the antibody encoded by the second rearrangement will be produced by the cell (Corcoran, 2005). If this second rearrangement does not produce a viable antibody the cell dies.

When assaying the BCR repertoire through sequencing, some of the sequences will be from cells for which the first rearrangement was successful, while others will be from cells with one productive and one out-of-frame rearrangement. Although the out-of-frame rearrangements from the second type of cell do not produce viable antibodies, their DNA gets sequenced along with the productive rearrangements. Since SHM rarely introduces insertions or deletions (we observe whole codon insertion deletion events in between 0.013% to 0.014% of memory sequences within templated segments), it is appropriate to assume that observed frame shifts in sequences are dominated by out-of-frame rearrangement events. However, because they are not expressed, but rather are carried along in cells with a separate functional rearrangement, they have no selective constraints. For this

reason, we use sequences from out-of-frame rearrangements as a proxy for the neutral mutation process in affinity maturation.

In this paper, we develop modern statistical molecular evolution methods for the analysis of high-throughput B cell sequence data, and then apply them to a very deep short-read data set of B cell receptors. Specifically, we first apply model selection criteria to find patterns in the affinity maturation single-nucleotide substitution process and find that it is similar across individuals but varies significantly across gene segments. Next, we use principal components analysis on substitution matrices to investigate how substitution processes vary between V genes and find that the primary source of variation is whether or not a sequence produces a functional receptor. Finally, we develop the first statistical methodology and corresponding software for comprehensive per-residue selection estimates for B cell receptors. We leverage out-of-frame rearrangements carried along in B cells with a productively rearranged receptor on the second chromosome to estimate evolutionary rates under neutrality, thus avoiding difficulties encountered by previous work in differentiating between selection and motif-driven mutation. A key part of our method is our extension of the "counting renaissance" method for selection inference (Lemey *et al.*, 2012) for non-constant sequencing coverage and a star-tree phylogeny. Using this modified method, we are able to efficiently derive a per-residue map of selection on more than 15 million B cell receptor sequences; we find that selection is dominated by negative selection with patterns that are consistent among individuals in our study.

RESULTS

Substitution model inference and testing.

name	branch length	GTR transition matrix	across-site rate variation (discrete Gamma)	total parameters
$t_i\, Q_i\, \Gamma_i$	One branch length per *segment* per sequence ($n \times 3$)	One matrix per segment (8×3)	One distribution per segment (3)	$3n + 27$
$t_r\, Q_i\, \Gamma_i$	One branch length per sequence (n) + relative rate between segments (2)	One matrix per segment (8×3)	One distribution per segment (3)	$n + 29$
$t_r\, Q_i\, \Gamma_s$	One branch length per sequence (n) + relative rate between segments (2)	One matrix per segment (8×3)	One shared distribution (1)	$n + 27$
$t_r\, Q_s\, \Gamma_s$	One branch length per sequence (n) + relative rate between segments (2)	One shared matrix (8)	One shared distribution (1)	$n + 11$

TABLE 1. The models of molecular evolution evaluated, including the number of free parameters introduced in parentheses.

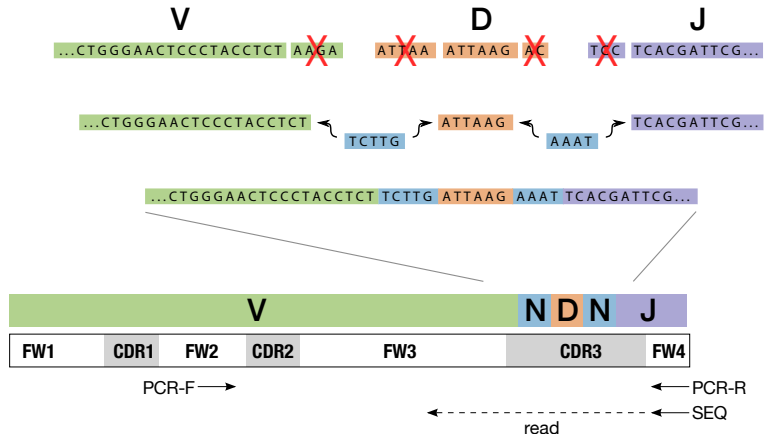

FIGURE 1. The VDJ recombination process and location of sequenced amplicon. In VDJ recombination, individual V, D, and J genes are randomly selected from a number of copies of each. These genes are then joined together via a process that deletes some randomly distributed number of nucleotides on their boundaries then joins them together with random "non templated insertions" (N, blue). The specificity of an antibody is primarily determined by the region defined by the heavy chain recombination site, referred to as the third complementarity determining region (CDR3). The sequence data for this study started in the fourth framework (FW) region and continued into the third. Amplification was via a forward primer in the FW2 region and a reverse primer in the FW4 region.

The setting of B cell affinity maturation is substantially different than that typically encountered in molecular evolution studies, and hence there are some differences between our model fitting procedure compared to common practice. For B cell receptors outside of nontemplated insertions, the root state is the V, D, and J genes encoded in the germline from which a sequenced BCR derives. Thus, we analyze substitutions that have occurred in evolution away from the germline-encoded segments of observed BCR reads, ignoring sites comprising nontemplated insertions. We apply our methods to a data set of memory B cell populations isolated from three healthy volunteers, referred throughout the manuscript as individuals A, B and C. Deep sequencing these B cell populations resulted in 15,023,951 (Tab. S1) unique 130bp reads after pre-processing that spanned the third heavy chain complementarity determining region (CDR3) region (Fig. 1). The CDR3 region of an antibody is generally sufficient to uniquely identify its specificity (Xu and Davis, 2000). Although there are certainly some clones in our data set that derive from a single rearrangement event but differ due to somatic hypermutation, the probability that a given pair of sequences derives from a single common ancestor is small: targeted searches for clonally related antibodies during infection have identified them at 0.003% to 0.5%(Zhu *et al.*, 2013). The classical situation for molecular evolution, on the other hand, assumes all sequences in a

data set have common ancestry. Additionally, we encountered significant computational barriers analyzing the volume of sequences available, and in fact we believe the 15 million unique sequences used in this study to be the largest number analyzed in a molecular evolution study from a single data set to date.

For these reasons, our analyses were performed on a set of pairwise alignments, each representing a two taxon tree containing an observed read and its best scoring germline sequence according to Smith-Waterman alignment. This is equivalent to using a rooted "star" tree where the root state is known.

We evaluated the fit of nested models with varying complexity, ranging from a simple model with shared branch lengths and substitution processes for the three independent segments of the BCR, to a complex model with completely separate substitution processes and branch lengths for each segment (Tab. 1). For the underlying nucleotide substitution model, we fit a general time-reversible (GTR) nucleotide model (Lanave *et al.*, 1984) with instantaneous rate matrix Q to subsets of the data, using 20,000 unique sequences from each individual. The choice of a stationary and reversible model, rather than a more general model, was based on the similarity of base frequencies between the germline sequences and observed reads (Tab. S2). We modeled substitution rate heterogeneity across sites using a four-category discretized Gamma distribution (Yang, 1994a) with fixed mean 1.0.

	model	log likelihood	df	AIC	ΔAIC
A	$t_r\, Q_i\, \Gamma_i$	-687,582	20,029	1,415,222	0
	$t_r\, Q_i\, \Gamma_s$	-687,980	20,027	1,416,014	793
	$t_r\, Q_s\, \Gamma_s$	-700,818	20,009	1,441,654	26,433
	$t_i\, Q_i\, \Gamma_i$	-662,417	60,027	1,444,888	29,666
B	$t_r\, Q_i\, \Gamma_i$	-507,980	20,029	1,056,017	0
	$t_r\, Q_i\, \Gamma_s$	-508,229	20,027	1,056,512	494
	$t_r\, Q_s\, \Gamma_s$	-517,320	20,009	1,074,658	18,641
	$t_i\, Q_i\, \Gamma_i$	-482,963	60,027	1,085,979	29,962
C	$t_r\, Q_i\, \Gamma_i$	-563,181	20,029	1,166,420	0
	$t_r\, Q_i\, \Gamma_s$	-563,291	20,027	1,166,637	217
	$t_r\, Q_s\, \Gamma_s$	-572,530	20,009	1,185,078	18,659
	$t_i\, Q_i\, \Gamma_i$	-539,018	60,027	1,198,090	31,671

TABLE 2. Models show identical ranking across individuals. Models described in Table 1. Columns include the number of degrees of freedom (df), Akaike Information Criterion (AIC), and difference of AIC from the top model (ΔAIC).

We find that the best performing model (denoted $t_r\, Q_i\, \Gamma_i$, Tab. 2) is one in which the branch length separating a sequenced BCR from its germline counterpart is estimated independently for each read, but that V, D and J regions differ systematically in their relative amounts of sequence change (denoted t_r). Additionally, this model uses separate GTR transition matrices for V, D and J regions (denoted Q_i) and uses separate distributions for across-site rate variation for V, D and J regions (denoted Γ_i). Looking across models, both the Akaike Information Criterion (AIC)(Akaike, 1974) (Tab. 2) and the Bayesian Information Criterion (Schwarz,

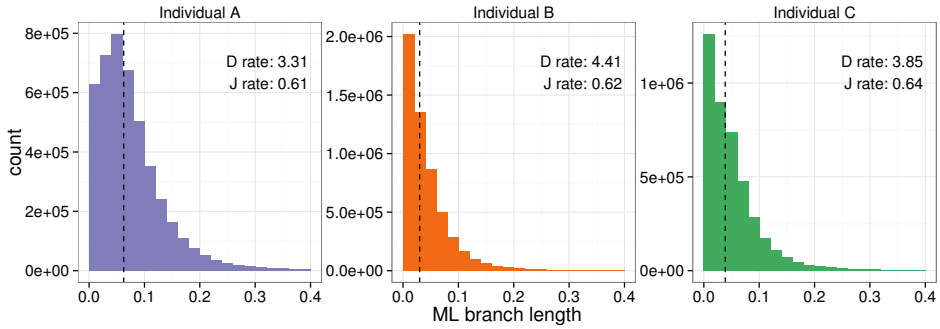

FIGURE 2. Distribution of maximum likelihood branch lengths estimated under the $t_r\, Q_i\, \Gamma_i$ model. Branch lengths are measured in terms of substitutions per site, and rates given for the D and J segments are relative to a fixed rate of 1 for the V segment.

1978) (data not shown) identified the same rank order of support; this ordering was also identical for each of the three individuals. Other than the $t_i\, Q_i\, \Gamma_i$ model, in which branch length is estimated independently across gene segments, models are ranked in terms of decreasing complexity. The finding that a complex model fits better than simpler models is likely aided by the large volume of sequence data available.

Next, we fit the best-scoring model ($t_r\, Q_i\, \Gamma_i$) to the full data set for each individual. The median distance to germline was 0.063, 0.030, and 0.039 substitutions per site for individuals A, B, and C, respectively. The distribution of branch lengths appears nearly exponential for individuals B and C, with many sequences close to germline and few distant from germline sequences (Fig. 2). Individual A displayed a higher substitution load and a non-zero mode. Despite these differences in evolutionary distance, the relative rate of substitution between the V, D, and J segments for each individual was very similar. We note that the sorting procedure used to separate memory from naïve B cells provided memory cells at approximately 97% purity, so these divergence estimates may be conservative due to low levels of contamination from the naïve repertoire.

Coefficients from the GTR models for the same gene segment across individuals were quite similar to one another, while models for different gene segments within an individual showed striking differences (Fig. 3, S1). However, overall correlations of GTR parameters between individuals were very high, yielding correlation coefficients between $\rho = 0.988$ and $\rho = 0.994$. We observe an enrichment of transitions relative to transversions in all segments, as previously described (Teng and Papavasiliou, 2007).

Next we compared the evolutionary process between various groupings of sequences to learn what determines the characteristics of this evolutionary process. We focused on the V gene segment, as it had the most coverage in our dataset, and partitioned the sequences by whether they were in-frame, then by individual, and then by gene subgroup. We fit the $t_r\, Q_i\, \Gamma_i$ model to 1000 sequences from

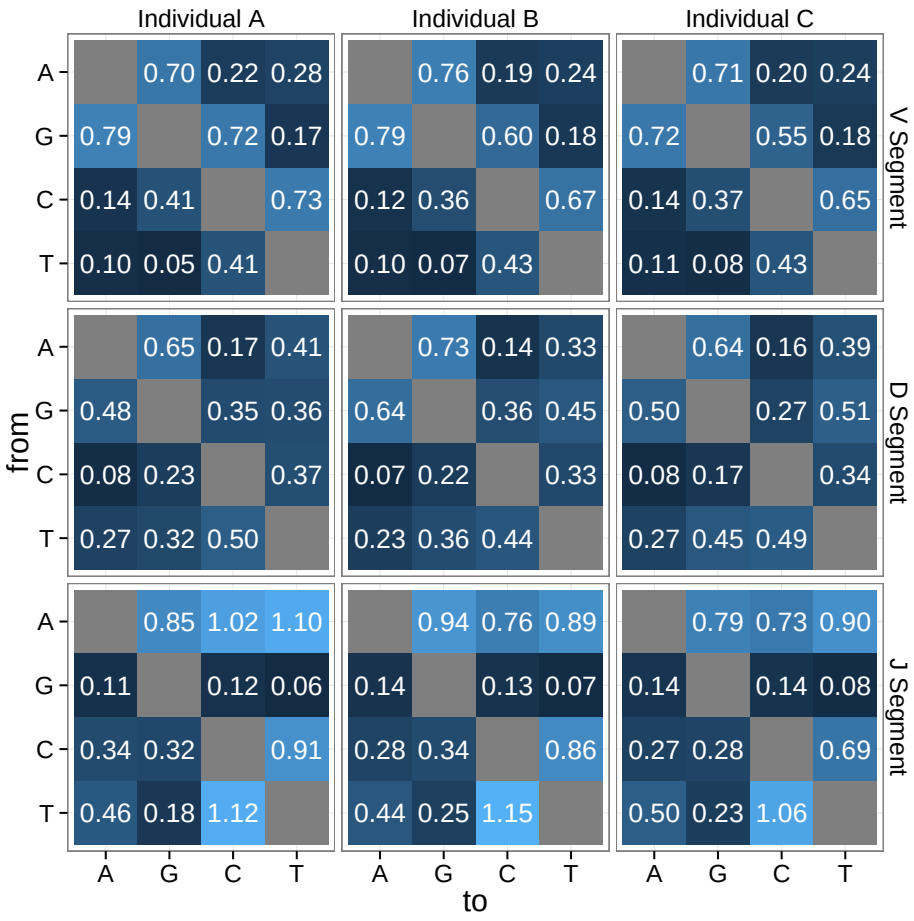

FIGURE 3. GTR coefficients for the $t_r\, Q_i\, \Gamma_i$ model estimated under maximum likelihood. Rows index the nucleotide found in the germline sequence, whereas columns index the nucleotide found in the observed sequence.

each set of the partition and calculated the transition probability matrix (P) associated with the median branch length across all sequences given an equiprobable starting state. These matrices were then analyzed with a variant of compositional principal components analysis (Aitchison, 1983) (see Materials and Methods). We find that substitution models are influenced by in- versus out-of-frame sequence status, find no evidence for models clustering by individual, and see some limited evidence for clustering by gene subgroup (Fig. 4). The Euclidean distance between these transformed discrete probability distributions and the Hamming distance between germline V genes showed significant, but moderate, correlation (Spearman's $\rho = 0.20$, $p < 10^{-15}$; Fig. S2).

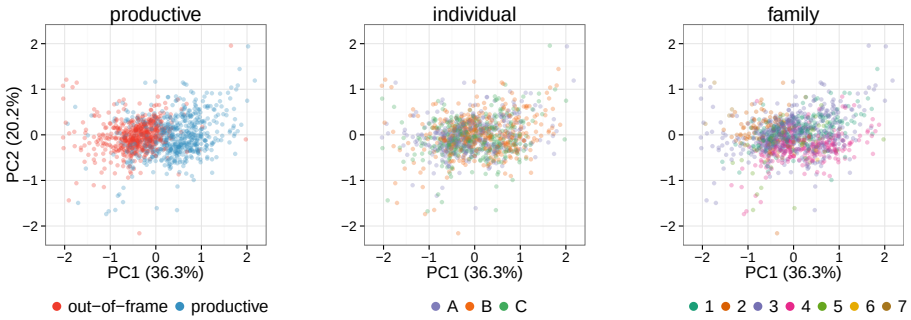

FIGURE 4. First (x-axis) and second (y-axis) principal components from PCA performed on centered log-transformed median-time transition matrices for V gene segments. Points plotted in a random order, with 22 outliers removed for clarity.

Selection. Here we describe our methods for estimating the per-site ratio ω of nonsynonymous to synonymous substitution rates. The primary challenge for selection inference for B cell receptors is that nucleotide context is known to have a very strong impact on mutation rates (reviewed in Teng and Papavasiliou (2007)); these context-specific mutations combined with the structure of the genetic code can result in extreme dN/dS ratios using the classical definition that are not attributable to selection. To address this problem, we infer ω using a nonsynonymous-synonymous ratio which controls for background mutation rate via out-of-frame sequences. Specifically, as described in the methods section, we define the dN/dS ratio ω at a site l as

$$\omega_l = \frac{\lambda_l^{(N,P)}/\lambda_l^{(N,O)}}{\lambda_l^{(S,P)}/\lambda_l^{(S,O)}}.$$

Here the $\lambda_l^{(\cdot,\cdot)}$ are the rates of mutation inferred using the renaissance counting procedure (described next), where N signifies nonsynonymous, S synonymous, P productive, and O out-of-frame. We continue the tradition of calling the selection coefficient ω in this context, even though it is a slightly different definition than previously used.

Because of the large volume of sequences to analyze, we also needed a mechanism to detect selection that could be run on over 15 million sequences. Classical means of estimating selection by codon model fitting (Goldman and Yang, 1994; Muse and Gaut, 1994) could not be used, even in their most recent and much more efficient recent incarnation (Murrell *et al.*, 2013). Instead, we used the renaissance counting approach (Lemey *et al.*, 2012), which we modified to work under varying levels of coverage. The counting renaissance method imputes ancestral substitutions under a simple (nucleotide) model to estimate parameters of a more complex (codon) model. As above, we use a two taxon tree for each read, here consisting of the read and best matching V segment.

A key part of the renaissance counting approach is an empirical Bayes regularization procedure (Robbins, 1956). This procedure uses the entire collection of sites

to inform substitution rate estimation at each site individually, effectively sharing data across sites, allowing inference at sites which either display few substitutions or have less read coverage. For example, with little coverage at a particular site, we might never observe it to be mutated in our data. At this site the empirical Bayes procedure would raise the estimate of substitution rate toward the population average across sites. Our data set had very uneven read depth coverage due to the varying length of the CDR3 region, and by truncating to the shortest read we would have lost valuable information. To address this, we extended the regularization procedure to the case of non-constant coverage, and validated it via simulation (see Materials and Methods). We use the phrase "Bayesian regularization" to describe this shrinkage of site-specific estimates of $\lambda_l, l = 1, \ldots, L$ to the overall mean, which is a common feature of all Bayesian hierarchical models.

Applying this method to our data set results in the first per-site and per-gene maps quantifying selection in the B cell repertoire (results uploaded to the Dryad data repository associated with this paper). Sites were classified as negatively or positively selected based on whether the 95% Bayesian credible interval (BCI) excludes one: sites for which the lower endpoint of the ω BCI is greater than one are classified as being under positive selection, while sites for which the upper endpoint of their ω BCI is less than one are classified as being under negative selection. We employ site numbering according to the IMGT unique numbering for the V domain (Lefranc *et al.*, 2003).

IGHV3-23*01 is the most frequent V gene/allele combination in our dataset, and it displays patterns that are consistent with the other genes. Specifically, we see significant variation in the synonymous substitution rate (right panels, Fig. 5a) even in out-of-frame sequences, which is presumably due to motif-driven mutation. Thus, if we had directly applied traditional means of estimating selection by comparing the rate of nonsynonymous and synonymous substitutions, we would have falsely identified sites as being under strong selection. We found the selection inferences made using out-of-frame sequences stay much closer to neutral (Fig. 5b).

We note extensive negative selection in the residues immediately preceding the CDR3 (Fig. 6). The amino acid profile for these sites shows a distinct preference for a tyrosine or rarely a phenylalanine two residues before the start of the CDR3 at site 102 (Fig. S3). It shows a preference for a tyrosine or more rarely a phenylalanine or a histidine in the residue just before the start of the CDR3 at site 103. We do not know the significance of the selection for aromatic amino acids in this region.

Overall we see extensive selection in our sequenced region (Fig. 7). The mean ω estimate across sites with at least 100 productive and out-of-frame reads aligned was 0.907. 65.6% of sites had a median $\omega < 1$ with a wide distribution of median ω values and confidence interval widths. However, many of them were observed to be positively, negatively, and neutrally evolving with narrow confidence intervals (Fig. 7, left column). 30.6% of sites were confidently classified as being under negative selection (Fig. 7, right column).

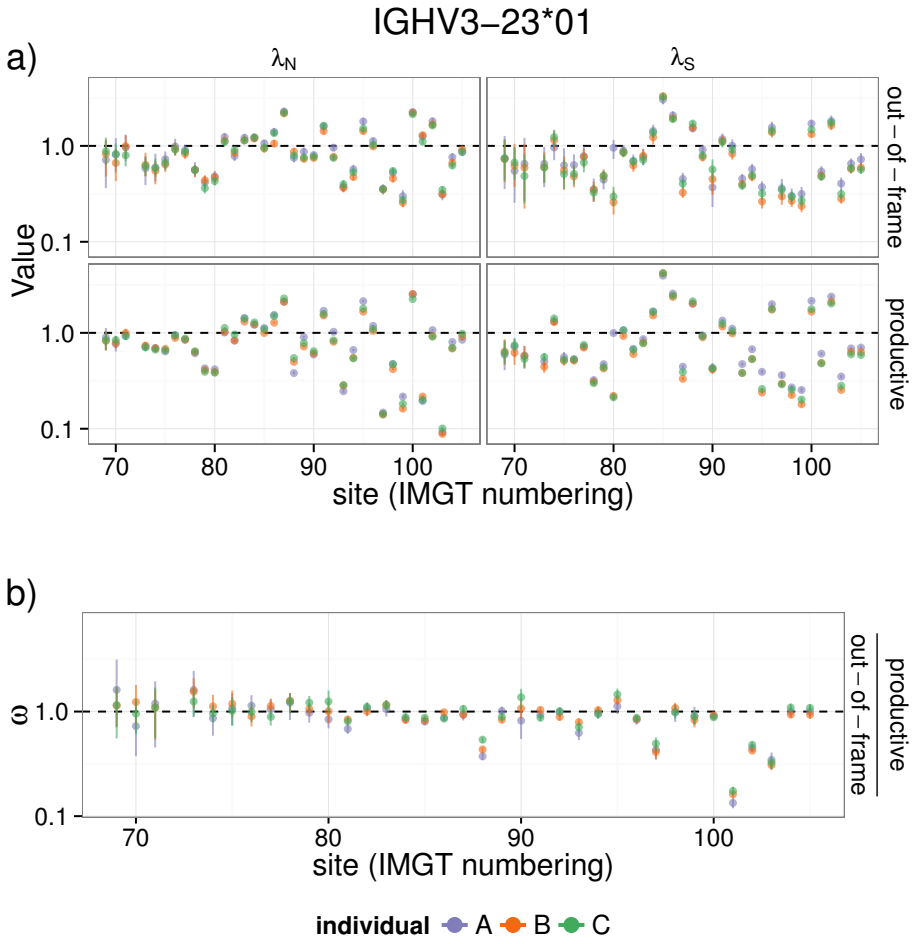

FIGURE 5. a) Comparison of nonsynonymous (λ_N) and synonymous (λ_S) rates in productive and out-of-frame sequences. b) ω estimates using unproductive rearragements as a proxy for the neutral process. Both panels use data from IGHV3-23*01, the most frequent V gene/allele combination.

Because amino acids interior to the protein could be important for protein stability compared to exposed ones, we hypothesized that residues under negative selection would be more internal to the antibody protein than those under neutral or positive selection, and that the inverse would be true for residues under positive selection. To test this, we mapped our ω estimates onto antibody structures (Fig. 8) and calculated the exposure of each amino acid position in the structure using the solvent-accessible surface area (SASA) using ROSETTA3 (Leaver-Fay *et al.*, 2011). The normalized SASA was well correlated with the classification of each site: sites classified as being under positive selection were most exposed in the protein structure, followed by neutral sites, then negatively selected sites (Fig. S6). Differences

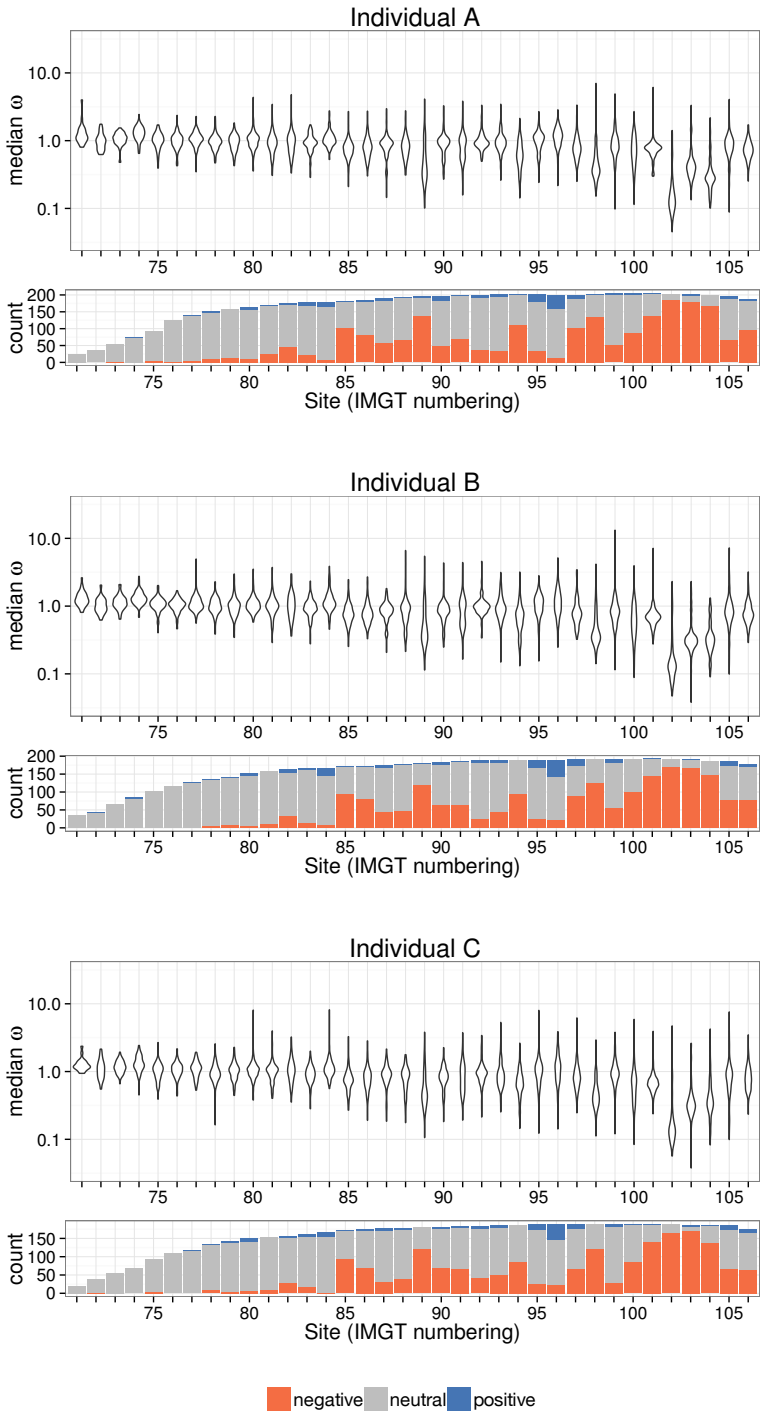

FIGURE 6. Site-specific estimates of the selection coefficient ω. Violin plots show distribution of median ω estimates across V genes at each site. Bar plots show count of V genes classified as undergoing negative, neutral, or positive selection. Only sites with at least 100 productive and out-of-frame reads aligned were considered.

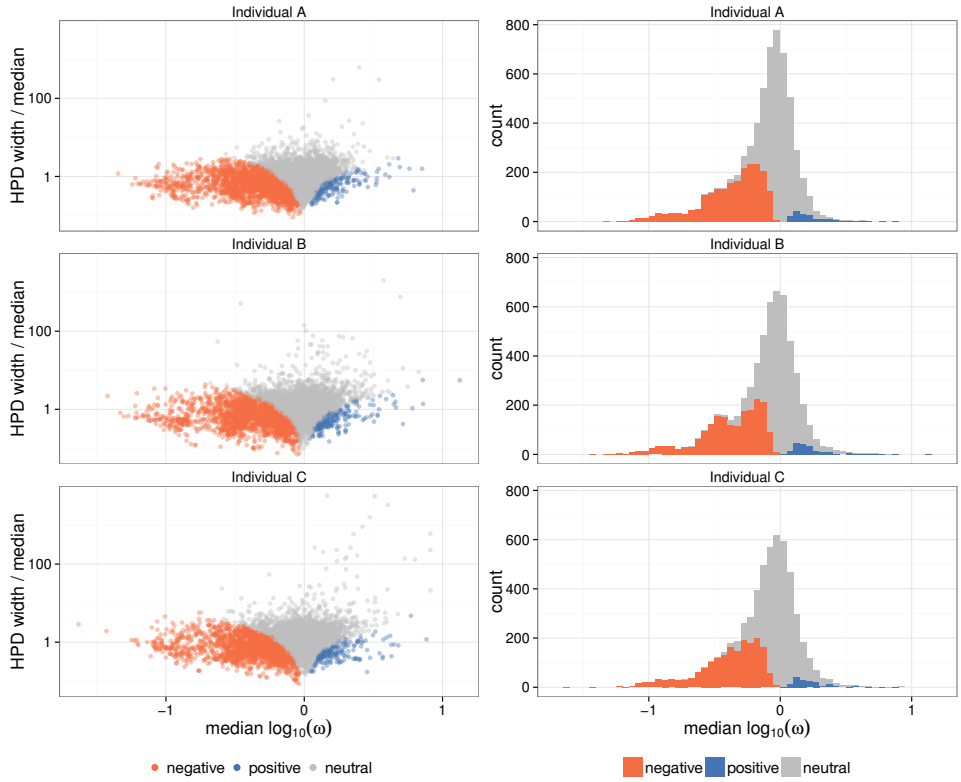

FIGURE 7. Site-specific selection estimates partitioned by individual and gene. Sites classified as negatively selected or positively selected based on whether the 95% Bayesian credible interval excludes 1 and in what direction. Left: comparison of ω estimate and relative width of BCI region. Right: distribution of site-specific selection estimates.

in surface accessibility were significant between the three groups, with p-values ranging from < 0.002 for the comparison of positive vs. neutral sites to $< 10^{-15}$ for the comparison of negative vs. neutral sites (Wilcoxon rank-sum test (Hollander and Wolfe, 1973)).

Despite the three individuals surveyed here presumably having quite different immune histories, we observe remarkable consistency in substitution and selection within the memory B cell repertoire. Indeed, we see a very strong correlation of median selection estimates between individuals (Fig. S4), with between-individual coefficients of determination R^2 of between 0.628 and 0.687 for site-specific ω values.

DISCUSSION

B cells have a complex developmental pathway, the last step of which is affinity maturation by somatic hypermutation and selection. In order to understand this

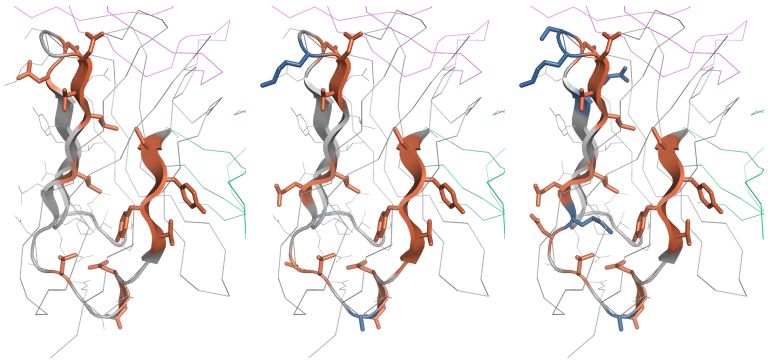

FIGURE 8. An IGHV3-23*01 (the most frequent V gene/allele combination) heavy chain antibody in complex with IL-17A (PDB ID 2VXS; Gerhardt *et al.* (2009)), with sites colored by ω classification in each of the three individuals sampled. The bound antigen is shown in pink (top), and the light chain in green (right). The heavy chain structure is shown as a thin gray line at sites which could not be classified due to insufficient coverage. When there is sufficient coverage it is shown as a cartoon (thick lines or arrows representing beta sheets) which is colored gray at neutral sites, red at negatively selected sites, and blue at positively selected sites.

process, as well as to reconstruct the events that led to a given collection of B cell receptors in a principled fashion, statistical models of these processes are needed. In this paper we have provided rigorous model development in a framework which considers each site independently of surrounding context.

Our biological results can be summarized as follows. We find different patterns of substitution across the V, D and J regions which is consistent among individuals (Fig. 3) even though those individuals have differing levels of substitution (Fig. 2). We find that the dominant factor determining the V segment substitution process is whether it is out-of-frame or productive, with the gene identity being a contributing factor. The pattern of selective pressure is consistent across individuals, and shows especially strong pressure near the boundary between the V gene and the CDR3. Selection estimates for BCRs are still high, with average ω of ≈ 0.9, compared to common examples of Darwinian evolution, such as seen in *Drosophila* (Clark *et al.*, 2007) and mammals (Lindblad-Toh *et al.*, 2011), where most genes show ω less than 0.1. However, we note that although our estimates of ω are comparable to more traditional estimates, we calculate ω slightly differently, using out-of-frame sequences as a control for motif-driven evolution. Finally, the patterns of selective pressure we observed correlated with levels of surface exposure in published antibody structures: highly conserved sites were more frequently found internally, while residues we classified as positively selected were more exposed.

We note that our analyses are based on data from only three individuals. It is possible that including more individuals would reveal variation in the mutation

process. However, we note that these three unrelated individuals had an extraordinary level of agreement, which cannot be explained by sequencing error.

Substitution process. The mutational process underlying somatic hypermutation has been extensively studied, especially how mutation rate depends on the surrounding nucleotide context. A considerable amount of our knowledge about this system comes from laboratory-based studies that express activation-induced deaminase (AID) in model systems (reviewed in Teng and Papavasiliou (2007)). This is very interesting and useful, but begs the question of how the process proceeds *in vivo* in human beings. New sequencing technology provides the raw material for the study of human blood samples in great depth; here we leverage a data set representing the most extensive sampling of the human B cell repertoire so far.

Recent work using deep sequencing has addressed mutation context specificity (Yaari *et al.*, 2013). Yaari *et al.* (2013) tabulated substitutions across all nucleotide substrings of length 5 ("5-mers") and used these to calculate a 1,024$\times$4 substitution rate matrix, in which the columns of the matrix correspond to substitutions of the central nucleotide of the 5-mer. Although they use a corpus of over a million processed reads (which were then subset to synonymous substitutions), they were not able to infer all values for this large matrix. In fact, there were no data whatsoever for 70% of the rows of this matrix; for these rows they used various methods to impute substitution rates in the absence of observations. They also inferred a matrix for substitutions into the various 5-mers, and used the same approaches to infer substitution rates for the 35% of rows that had no data. Thus, this work represents an interesting and highly ambitious project to fit an over-parameterized model.

Here, we take a different approach in terms of strategy and goals for our inference of substitution models. We use model-selection criteria to infer statistical models of substitution whose complexity is guided by the data and consider aspects of the substitution process besides the nucleotide context. Our model inferences show that the best-fitting model allows for a single branch length per sequence, a global multiplier for per-segment differences, a per-segment substitution model, and a per-segment rate variation model across sites (Tab. 2). These between-segment differences are certainly due in part to base composition, which also differs significantly between segments and is similar between individuals (Table S2). Another contributing factor is probably similarity of local nucleotide context between the genes of a given segment compared to between segments; these nucleotide contexts are known to impact AID-induced somatic hypermutation (reviewed in Teng and Papavasiliou (2007)). We also note that the entirety of the D segment lies within the CDR3 region, and is thus more likely to directly contact an epitope; not surprisingly, we observe higher substitution rates within that segment. By analyzing distances between GTR substitution rate matrices we find that the most important difference between them is determined by whether they are productive or non-productive (Fig. 4), and we find a significant correlation between sequence identity and substitution matrix. This analysis is inspired by previous work on "evolutionary fingerprinting," in which genes are characterized by their substitution patterns rather than their sequence identity (Kosakovsky Pond *et al.*, 2010).

Our motivation in closely examining the substitution and selection processes in a context-independent manner is not to make a full description of this clearly context-dependent process, but rather to provide a solid framework for future study and to enable downstream comparative analyses such as Figure 4. In doing so we focus on other aspects of the process, such as rate variation among sites and how the substitution process differs between genes and segments. The class of models we consider here can be used directly for maximum-likelihood and Bayesian reconstruction of B cell lineages within the context of standard phylogenetics packages, which would not be true if we inferred context-dependent patterns of substitution. This will permit likelihood-based lineage inference for B cell receptors, which will enable researchers to leverage decades of research in statistical phylogenetics. Motif-dependence of mutation is a very interesting and important topic; we are in the process of doing a statistical analysis of motif-dependent substitution to follow up on the present work.

Selection process. The role of selection in B cell receptor development has stimulated continuous interest since the pioneering 1985 paper of Clarke and colleagues (Clarke *et al.*, 1985), however methods for the analysis of antigen selection have developed in parallel to related work in the population genetics and molecular evolution community. Work on the selection process for BCRs has focused on aggregate statistics to infer selection for entire sequences or sequence tracts, and there has been a lively debate about the relative merits of these tests (Chang and Casali, 1994; Lossos *et al.*, 2000; Bose and Sinha, 2005; Hershberg *et al.*, 2008; Yaari *et al.*, 2012). Recent work has offered methods that evaluate selection on a per-sequence basis (Yaari *et al.*, 2012). There have also been efforts to infer selection based on lineage shape (Steiman-Shimony *et al.*, 2006; Abraham *et al.*, 2006; Barak *et al.*, 2008; Shahaf *et al.*, 2008; Uduman *et al.*, 2014), which has been a common approach in macroevolutionary studies (reviewed in Mooers and Heard (1997)) and more recently in population genetics (Drummond and Suchard, 2008; Li and Wiehe, 2013).

In this work, we develop the first means of inferring per-residue selection using high-throughput sequence data with non-uniform coverage. Our method sidesteps the difficulties encountered by previous work in differentiating between selection and motif-driven mutation in B cell receptors (Chang and Casali, 1994; Dunn-Walters and Spencer, 1998; Lossos *et al.*, 2000; Bose and Sinha, 2005; Hershberg *et al.*, 2008; Yaari *et al.*, 2012) by developing statistical means to compare in-frame and out-of-frame rearrangements. Also, in contrast to this previous work on B cell selection, it provides a per-residue selection map rather than selection estimates for entire sequence tracts. The closest previous work has gotten to a per-residue selection inference map is the publication of per-residue histograms of substitutions stratified into synonymous and nonsynonymous groups (Dunn-Walters and Spencer, 1998). We also note that Yaari *et al.* (2013) indicated some per-site selection results by finding three sites that were three standard deviations away from the regression line when comparing the observed substitution frequency to an expected mutability based on nucleotide context.

We use out-of-frame rearrangements as our selection-free control population. These sequences do not create functional IGH proteins, but may be carried in heterozygous B cells which do have a productively rearranged IGH allele. Thus

they undergo SHM, but any selection occurs on the level of the productively rearranged allele, not on the residues in the unproductive allele. We acknowledge that some out-of-frame sequences could still feel the impact of selection, which would occur if the sequences accrue frameshift mutations in the process of affinity maturation. However, it is thought that SHM is primarily a process of point mutation (Teng and Papavasiliou, 2007), and indeed, we observe whole codon indels in only 0.013%–0.014% of memory sequences within templated segments. Still, if a weaker version of selection was occurring on the out-of-frame sequences compared to the productive ones then this would simply make our estimates of selection conservative, pulling estimates of ω closer to 1, and yet our selection estimates are confidently classified as non-neutral for a substantial fraction of sites (Fig. 7).

In applying our methodology to IGHV sequences, we gain a high resolution per-gene map of selective forces on B cell receptors, which is dominated by negative selection among sites for which selection could be confidently classified. We see an pattern of quite strong negative selection in the region around the beginning of the CDR3. This agrees with recent work that also found strong negative selection in one site near the beginning of the CDR3 (Yaari *et al.*, 2013). As also indicated by these authors, our results provide a more nuanced view into the constraints on B cell receptor sequences rather than the traditional framework/CDR designations (Yaari *et al.*, 2013).

In conclusion, our work puts down a solid foundation of statistical models for future molecular evolutionary studies of B cell receptors. By focusing on context-independent models, we are able to do a statistical model inference procedure including a number of aspects of the molecular evolution process that have not been considered before. We find that a moderately parameter-rich model of substitution and rate variation fits the data best; this non-trivial structure to the substitution process can be leveraged in future studies. We perform selection inference using an empirical Bayes regularization process of stochastic mapping, which we develop for non-constant sequencing coverage. By applying this new method, we are able to derive a per-residue map of selection without the confounding effects of context-dependent substitution. We find that selection is primarily purifying, with a pattern that is consistent among individuals.

Materials and Methods

Data set. The complete description of the experiment will be published separately (manuscript in preparation). However, here we give a brief overview of the data in order to facilitate understanding of our analysis and to emphasize that the experimental design has a number of features that greatly reduce errors in sequencing and quantification. 400cc of blood was drawn from three healthy volunteers under IRB protocol at the Fred Hutchinson Cancer Research Center. $CD19^+$ cells were obtained by bead purification then flow sorted to isolate over 10 million naïve ($CD19^+CD27^-IgD^+IgM^+$) and over 10 million memory ($CD19^+CD27^+$) B cells, with greater than 97% purity. Genomic DNA was extracted and the ImmunoSeq assay described in Larimore *et al.* (2012) was performed on the six samples at Adaptive Biotechnologies in Seattle, WA.

The experiments and preprocessing were carefully designed to give an accurate quantification of error-corrected reads. To mitigate preferential amplification

of some V/J pairs through primer bias, the PCR amplification was performed using primers optimized via a large collection of synthetic templates (Carlson *et al.*, 2013). To reduce sequencing errors and provide accurate quantification, each sample was divided amongst the wells on two 96 well plates and bar-coded individually. These templates were then amplified and "over-sequenced" (Table S1), such that each template was sequenced an average of about 10 times; these various sequences representing the same template can then be collapsed into a consensus sequence with greatly reduced sequencing error. Following Robins *et al.* (2009) for each well separately, we clustered all sequences into groups by joining reads with Hamming distance less than or equal to two, and inferred the underlying sequence in each cluster using parsimony. Groups with only one member were discarded. In order for two distinct sequences to be collapsed into one, they would have to co-occur exclusively in the same wells.

The full data set will be made public upon publication of the manuscript describing the experiment. We have made a virtual machine available for running our analyses, which are described next, on a subset of the data (Supplemental Methods).

Alignment and germline assignment. Each sequence read was first aligned to each V gene using Smith-Waterman algorithm with an affine gap penalty (Gotoh, 1982). The 3′ portion of the sequence not included in the best V gene alignment was next aligned to all D and J genes available from the IMGT database (Lefranc *et al.*, 2008). The best scoring V, D, and J alignment for each read was taken to be the germline alignment, and the corresponding germline sequence was taken to be the ancestral sequence for that read; in the case of ties, one germline sequence was chosen randomly among those alleles present at abundance $\geq 10\%$. Sequences were classified as productive or out-of-frame based on whether the V and J segments were in the same frame; all sequences with stop codons were removed, as these sequences could result from either an unproductive rearrangement event or inactivation due to a lethal mutation. The 18 V gene polymorphisms present at the highest frequency in the naïve populations of the individuals surveyed which were not represented in the IMGT database were added to the list of candidates for alignment.

Substitution models, fitting and analysis. Substitution models are summarized in Table 1 and described in detail here. We will use n for the number of reads. Our models are characterized by three components. First, the subscript of t describes how branch length assignments are allowed to vary across segments of a single sequence. The t_i model allows branch lengths to vary independently, resulting in $3n$ parameters. The t_r model has two global per-segment multipliers to define the branch lengths (see, e.g. Fig. 2) with the V segment rate fixed at 1, resulting in $n+2$ parameters. The subscript of Q describes how rate matrices are fit. The Q_i model allows an independent global GTR rate matrix for each segment, with a total of 24 parameters. The Q_r model just has one GTR rate matrix overall, with 8 parameters. The subscript of Γ denotes how across-site substitution rate variation is modeled in terms of a four category discrete gamma distribution (Yang, 1994a). The Γ_i model allows an independent rates across sites parameter for each read, with 3 parameters. The Γ_s has a global rates across sites parameter, with 1 parameter. Given these choices concerning how the data was partitioned and parametrized, the standard

phylogenetic likelihood function was used as described in the original literature (Felsenstein, 1981; Tavaré, 1986; Yang, 1994b) and in books (e.g. Salemi *et al.* (2009); Felsenstein (2004)).

Maximum likelihood values of substitution model parameters and branch lengths were estimated using a combination of Bio++ (Guéguen *et al.*, 2013) and BEAGLE (Ayres *et al.*, 2011), with model optimization via the BOBYQA algorithm (Powell, 2009) as implemented in NLopt (Johnson, 2010), and branch length optimization via Brent's method (Brent, 1973). Optimization alternated between substitution model parameters and branch lengths until the change in log-likelihood at a given iteration was less than 0.001. Our software to perform this optimization is available from `https://github.com/cmccoy/fit-star`.

For the principal components analysis on substitution matrices, we first obtained the median branch length $\hat{t}$ across all sequences for all individuals. We then calculated the corresponding transition matrix for each model given equiprobable starting state: $e^{Q\hat{t}}\,\mathrm{diag}(0.25)$. These were then projected onto the first two principal components, adapting suggestions for doing PCA in the simplex (Aitchison, 1983). Specifically, each row of these matrices, as a discrete probability distribution, is a point in the simplex. Hence we applied a centered log transformation to each row of this matrix using the `clr` function of the R package `compositions`(van den Boogaart and Tolosana-Delgado, 2008), and followed with standard principal components analysis.

To compare distance between inferred models and sequence distance, we calculated the Hamming distance between all pairs of V genes using the alignment available from the IMGT database (Lefranc *et al.*, 2008). To obtain distances between models, we calculated the Euclidean distance calculated between pairs of the transformed probability vectors used in the PCA analysis above.

Selection analysis.

Bayesian inference of star-shaped phylogeny. To determine the site-specific selection pressure for each V gene, we extended the counting renaissance method, described in Lemey *et al.* (2012), to accommodate pairwise analyses of a large number of sequences with a known ancestral sequence and non-constant site coverage. The counting renaissance method starts by assuming a separate HKY substitution model (Hasegawa *et al.*, 1985) for each of the three codon positions and uses Markov chain Monte Carlo (MCMC) to approximate the posterior distribution of model parameters that include substitution rates and phylogenetic tree with branch lengths. Since in our analyses we assumed that query sequences are related by a star-shaped phylogeny, our model parameters included only HKY model parameters and branch lengths leading to all the query sequences. Moreover, we fixed the parameters of the HKY model, along with the relative rates between codon positions, to the maximum likelihood estimates produced using the whole dataset. We note that we could have fit per-codon-position GTR models and used them for stochastic mapping, however such a model would still be substantially misspecified compared to a codon model and thus we decided to follow Lemey *et al.* (2012) and use HKY for the mapping. *A priori*, we assumed that branch lengths leading to the query sequence independently follow an exponential distribution with mean 0.1. We performed 20,000 iterations of MCMC, scaling the branch length leading to

the observed sequence at each iteration, and sampling every 40 iterations to generate a total of 500 samples. Given each posterior sample of query branch lengths, the counting renaissance method draws a sample of ancestral substitutions conditional on the observed data using a simple per-codon-position nucleotide model; the resulting sampled ancestral sequences are then used to count synonymous and nonsynonymous mutations.

Sampling codon substitutions. For each codon position l and posterior sample j, counts of synonymous ($C_{jl}^{(S)}$) and nonsynonymous ($C_{jl}^{(N)}$) substitutions at each site were imputed using stochastic mapping as described in Lemey *et al.* (2012).

For N MCMC iterations based on an alignment of L codons, the result of this procedure was two $N \times L$ matrices, each containing the number of synonymous and nonsynonymous events at each codon position in each posterior sample. Counts of each substitution type along with the total branch length for each site were aggregated across sequences from the same gene by element-wise addition.

Empirical Bayes regularization. The varying length of the CDR3, combined with short reads, leads to quite skewed coverage of sites stratified by gene. We modified the empirical Bayes regularization procedure of the original counting renaissance method (Lemey *et al.*, 2012) to account for varying depth of observation as follows. First, we define a branch length leading to query sequence i for site l as

$$t_{il} = \begin{cases} t_i, & \text{if any residues in the observed sequence } i \text{ align to codon position } l \\ 0, & \text{otherwise} \end{cases}$$

We assume that substitution counts for site l come from a Poisson process with rate $\lambda_l t_l$:

$$C_l \sim \text{Poisson}(\lambda_l t_l),$$

where $t_l = \sum_{i=1}^{n} t_{il}$.

As in the original counting renaissance, we assume that the site-specific rates λ_l come from a Gamma distribution with shape α and rate β:

$$\lambda_l \sim \text{Gamma}(\alpha, \beta).$$

We fix α and β to their maximum likelihood estimates $\hat{\alpha}$ and $\hat{\beta}$ by treating sampled branch lengths and counts as fixed and maximizing the likelihood function

$$\mathcal{L}(\alpha, \beta) = \left(\frac{\beta^\alpha}{\Gamma(\alpha)} \right)^L \prod_l \frac{t_l^{C_l}}{\Gamma(C_l + 1)} \frac{\Gamma(C_l + \alpha)}{(t_l + \beta)^{C_l + \alpha}}. \tag{1}$$

We provide a derivation of this likelihood function below. In contrast to (Lemey *et al.*, 2012), we do not have closed-form solutions for the maximum likelihood or method of moments estimators of α and β in this slightly more complex setting. However it does not add a substantial computational burden to estimate these parameters numerically via the BOBYQA optimizer (Powell, 2009).

Given $\hat{\alpha}$ and $\hat{\beta}$, we draw rates λ_l from the posterior:

$$\lambda_l \mid C_l \sim \text{Gamma}(C_l + \hat{\alpha}, t_l + \hat{\beta}), \tag{2}$$

derived below.

Estimation of α and β by maximizing likelihood (1) fails when the sample variance of the observed counts $C_1 \dots C_L$, weighted by the site-specific branch length sums, $t_1 \dots t_L$, is less than the corresponding weighted sample mean. In these cases, we assume that the observed counts are drawn from Poisson distributions with site-specific rate λt_l:

$$C_l \sim \text{Poisson}(\lambda t_l),$$

where here λ is shared across sites, and is estimated from the data by maximizing the likelihood

$$L(\lambda) = \prod_l^L \frac{(\lambda t_l)^{C_l}}{C_l!} e^{-\lambda t_l}.$$

Simulations. To validate this method, we simulated 1000 sequences of 100 codon sites each under the GY94 model and a star-like phylogeny with branch lengths fixed to 0.05 using piBUSS (Bielejec *et al.*, 2013). We varied ω over the alignment, with 85 sites having $\omega = 0.1$, 5 sites having $\omega = 1$, and 10 sites under positive selection - $\omega = 10$. We next introduced varying coverage over the alignment: sequences were truncated such that no sequences covered the first 10 codons, only half of the sequences had coverage over the next 40 codons, and all sequences covered the remaining 50 codons (Fig. S5, bottom panel). Estimates of ω were more accurate with higher site coverage (Fig. S5, top panel). Of note, as a result of the empirical Bayes regularization, even some sites with no coverage were classified as being under purifying selection. In all other analyses, we only report ω estimates for sites covered by at least 100 sequences. Since the starting state is always the germline amino acid, no classifications can be made for sites which are Tryptophan or Methionine in the germline, as all mutations are nonsynonymous for codons encoding those amino acids.

Site-specific estimates of ω. In Lemey *et al.* (2012), the authors arrive at site-specific estimates of ω_l by comparing data-conditioned (C) rates λ_l of nonsynonymous (N) and synonymous (S) substitutions, each normalized by an "unconditional rate" (U): $\omega_l^{RC} = \frac{\lambda_l^{(N,C)}/\lambda_l^{(N,U)}}{\lambda_l^{(S,C)}/\lambda_l^{(S,U)}}$. As SHM is highly context-specific, we chose to use rates inferred from out-of-frame rearrangements in place of the unconditional rates, as these more accurately represent the mutation rates in the absence of selection:

$$\omega_l = \frac{\lambda_l^{(N,P)}/\lambda_l^{(N,O)}}{\lambda_l^{(S,P)}/\lambda_l^{(S,O)}},$$

where P and O refer to productive and out-of-frame rearrangements, respectively.

Structural analysis. For each of the eleven most frequently occurring V genes, we identified the closest structure in the Protein Data Bank (PDB) (Berman *et al.*, 2000) using BLAST (Altschul *et al.*, 1997). Structures were visualized using PyMOL (Delano, 2002). We calculated the SASA for each amino acid position using ROSETTA3 (Leaver-Fay *et al.*, 2011) and normalized these values by dividing them by the fully exposed SASA of the given residue type in an extended chain. Wilcoxon rank-sum tests (Hollander and Wolfe, 1973) between all pairs of selection

classifications (negative, neutral, positive) were used to assess whether the normalized SASA differed between groups. p-values were Bonferroni-corrected (Holm, 1979) to account for multiple comparisons.

Derivation of the Gamma-Poisson marginal likelihood with varying observation depth. Our first task is to write down a likelihood of α and β given a collection of counts. To do so we will marginalize out the rates λ_l when they are drawn from a $\mathrm{Gamma}(\alpha, \beta)$ as above.

The likelihood for a single site is (omitting l for now):

$$\begin{aligned} P(C|t,\alpha,\beta) &= \int_0^\infty P(C|t,\lambda)P(\lambda|\alpha,\beta)d\lambda \\ &= \int_0^\infty \frac{(\lambda t)^C e^{-\lambda t}}{C!} P(\lambda|\alpha,\beta)d\lambda \\ &= \int_0^\infty \frac{(\lambda t)^C e^{-\lambda t}}{C!} \left[\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}\right] d\lambda \\ &= \frac{\beta^\alpha t^C}{C!\,\Gamma(\alpha)} \int_0^\infty \lambda^{C+\alpha-1} e^{-\lambda(t+\beta)} d\lambda. \end{aligned}$$

Letting $\alpha' = C + \alpha$ and $\beta' = t + \beta$, introduce a normalizing constant for the distribution $\mathrm{Gamma}(\alpha', \beta')$:

$$\begin{aligned} P(C|t,\alpha,\beta) &= \frac{\beta^\alpha t^C}{C!\,\Gamma(\alpha)} \frac{\Gamma(\alpha')}{\beta'^{\alpha'}} \int_0^\infty \frac{\beta'^{\alpha'}}{\Gamma(\alpha')} \lambda^{\alpha'-1} e^{-\lambda(\beta')} d\lambda \\ &= \frac{\beta^\alpha t^C}{C!\,\Gamma(\alpha)} \frac{\Gamma(\alpha')}{\beta'^{\alpha'}} \int_0^\infty \mathrm{DGamma}(\lambda;\alpha',\beta') d\lambda. \end{aligned}$$

The integral over the support of the Gamma distribution is 1, so:

$$\begin{aligned} P(C|t,\alpha,\beta) &= \frac{\beta^\alpha t^C}{C!\,\Gamma(\alpha)} \frac{\Gamma(\alpha')}{\beta'^{\alpha'}} \\ &= \frac{\beta^\alpha t^C}{C!\,\Gamma(\alpha)} \frac{\Gamma(C+\alpha)}{(t+\beta)^{C+\alpha}}. \end{aligned}$$

The overall marginal likelihood is the product over such sites:

$$\begin{aligned} \mathcal{L} = P(C_1,\ldots,C_L|t_1,\ldots,t_L,\alpha,\beta) &= \prod_l \frac{\beta^\alpha t_l^{C_l}}{C_l!\,\Gamma(\alpha)} \frac{\Gamma(C_l+\alpha)}{(t_l+\beta)^{C_l+\alpha}} \\ &= \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^L \prod_l \frac{t_l^{C_l}}{C_l!} \frac{\Gamma(C_l+\alpha)}{(t_l+\beta)^{C_l+\alpha}} \\ &= \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^L \prod_l \frac{t_l^{C_l}}{\Gamma(C_l+1)} \frac{\Gamma(C_l+\alpha)}{(t_l+\beta)^{C_l+\alpha}}, \end{aligned}$$

giving (1).

Posterior for λ. Our eventual goal is a regularized posterior estimate of the rates λ_l. For a single site, once again dropping l:

$$P(\lambda|C,t,\hat{\alpha},\hat{\beta}) \propto P(C|\lambda,t)P(\lambda|\hat{\alpha},\hat{\beta}).$$

Substituting in the PDFs for the distributions employed for C and λ:

$$P(\lambda|C,t,\hat{\alpha},\hat{\beta}) \propto \frac{\hat{\beta}^{\hat{\alpha}}t^C}{C!\,\Gamma(\hat{\alpha})}\lambda^{C+\hat{\alpha}-1}e^{-\lambda(t+\hat{\beta})}.$$

As above, we let $\hat{\alpha}' = C + \hat{\alpha}$ and $\hat{\beta}' = t + \hat{\beta}$.

$$\begin{aligned} P(\lambda|C,t,\hat{\alpha},\hat{\beta}) &\propto \frac{\hat{\beta}^{\hat{\alpha}}t^C}{C!\,\Gamma(\hat{\alpha})}\frac{\Gamma(\hat{\alpha}')}{\hat{\beta}'^{\hat{\alpha}'}}\left[\frac{\hat{\beta}'^{\hat{\alpha}'}}{\Gamma(\hat{\alpha}')}\lambda^{\hat{\alpha}'-1}e^{-\lambda(\hat{\beta}')}\right] \\ &\propto \frac{\hat{\beta}^{\hat{\alpha}}t^C}{C!\,\Gamma(\hat{\alpha})}\frac{\Gamma(\hat{\alpha}')}{\hat{\beta}'^{\hat{\alpha}'}}\mathrm{DGamma}(\lambda;\hat{\alpha}',\hat{\beta}') \\ &\propto \mathrm{DGamma}(\lambda;\hat{\alpha}',\hat{\beta}'), \end{aligned}$$

hence these two probability densities are equal, justifying (2).

Implementation details. We used the BEAST (Drummond *et al.*, 2012) implementation of the counting renaissance to sample counts for both synonymous and nonsynonymous substitutions at each site. We extended BEAST version 1.8.0 to generate "unconditional" counts using the germline state as the starting state for simulating along the edge to the query as described above. This process (sampling substitutions for each sequence, then combining counts from sequences mapping to the same IGHV) provides a natural setting for parallelization via the map-reduce model of computation; we used the Apache Spark (Zaharia *et al.*, 2010) framework to distribute work across a cluster running on Amazon EC2. Our software to perform this analysis is available from `https://github.com/cmccoy/startreerenaissance`.

ACKNOWLEDGEMENTS

The molecular work for this project was done by Paul Lindau in the laboratory of Phil Greenberg, and was supported by a grant from the W. M. Keck Foundation. C.O.M. and F.A.M. supported in part by a 2013 new investigator award from the University of Washington Center for AIDS Research (CFAR), an NIH funded program under award number P30AI027757 which is supported by the following NIH Institutes and Centers: NIAID, NCI, NIMH, NIDA, NICHD, NHLBI, NIA, NIGMS, and NIDDK. C.O.M. and F.A.M. were also supported in part by the University of Washington eScience Institute through its Seed Grants program in Translational Health Sciences. V.N.M. was supported in part by the National Science Foundation (DMS-0856099) and National Institute of Health (R01-AI107034). C.O.M. was supported in part by NIH grant AI103981 to Dr. Julie Overbaugh.

COMPETING INTERESTS

H.S.R. owns stock in and consults for Adaptive Biotechnologies. The other authors have no competing interests.

REFERENCES

Abraham, R. S., Manske, M. K., Zuckerman, N. S., Sohni, A., Edelman, H., Shahaf, G., Timm, M. M., Dispenzieri, A., Gertz, M. A., and Mehr, R. 2006. Novel analysis of clonal diversification in blood B cell and bone marrow plasma cell clones in immunoglobulin light chain amyloidosis. *J. Clin. Immunol.*, 27(1): 69–87.

Aitchison, J. 1983. Principal component analysis of compositional data. *Biometrika*, 70(1): 57–65.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, 19(6): 716–723.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25(17): 3389–3402.

Amazon Web Services 2014. Amazon EC2 FAQs.

Ayres, D. L., Darling, A., Zwickl, D. J., Beerli, P., Holder, M. T., Lewis, P. O., Huelsenbeck, J. P., Ronquist, F., Swofford, D. L., Cummings, M. P., Rambaut, A., and Suchard, M. A. 2011. BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.*, 61(1): 170–173.

Barak, M., Zuckerman, N., Edelman, H., Unger, R., and Mehr, R. 2008. IgTree (c) : Creating immunoglobulin variable region gene lineage trees. *Journal of Immunological Methods*, 338(1-2): 67–74.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. 2000. The protein data bank. *Nucleic Acids Res.*, 28(1): 235–242.

Bielejec, F., Lemey, P., Carvalho, L. M., Baele, G., Rambaut, A., and Suchard, M. A. 2013. piBUSS: a parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios.

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

Boyd, S. D., Marshall, E. L., Merker, J. D., Maniar, J. M., Zhang, L. N., Sahaf, B., Jones, C. D., Simen, B. B., Hanczaruk, B., Nguyen, K. D., Nadeau, K. C., Egholm, M., Miklos, D. B., Zehnder, J. L., and Fire, A. Z. 2009. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. *Sci. Transl. Med.*, 1(12): 12ra23–12ra23.

Brent, R. P. 1973. *Algorithms for Minimization Without Derivatives*. Courier Dover Publications.

Carlson, C. S., Emerson, R. O., Sherwood, A. M., Desmarais, C., Chung, M.-W., Parsons, J. M., Steen, M. S., LaMadrid-Herrmannsfeldt, M. A., Williamson, D. W., Livingston, R. J., Wu, D., Wood, B. L., Rieder, M. J., and Robins, H. 2013. Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.*, 4: 2680.

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

Clark, A. G., Eisen, M. B., Smith, D. R., Bergman, C. M., Oliver, B., Markow, T. A., Kaufman, T. C., Kellis, M., Gelbart, W., Iyer, V. N., *et al.* 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167): 203–218.

Clarke, S. H., Huppi, K., Ruezinsky, D., Staudt, L., Gerhard, W., and Weigert, M. 1985. Inter- and intraclonal diversity in the antibody response to influenza hemagglutinin. *J. Exp. Med.*, 161(4): 687–704.

Corcoran, A. E. 2005. Immunoglobulin locus silencing and allelic exclusion. *Semin. Immunol.*, 17(2): 141–154.

DeKosky, B. J., Ippolito, G. C., Deschner, R. P., Lavinder, J. J., Wine, Y., Rawlings, B. M., Varadarajan, N., Giesecke, C., Dörner, T., Andrews, S. F., Wilson, P. C., Hunicke-Smith, S. P., Willson, C. G., Ellington, A. D., and Georgiou, G. 2013. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.*, 31(2): 166–169.

Delano, W. L. 2002. The PyMOL molecular graphics system.

Delker, R. K., Fugmann, S. D., and Papavasiliou, F. N. 2009. A coming-of-age story: activation-induced cytidine deaminase turns 10. *Nat. Immunol.*, 10(11): 1147–1153.

Drummond, A. J. and Suchard, M. A. 2008. Fully Bayesian tests of neutrality using genealogical summary statistics. *BMC Genet.*, 9: 68.

Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.*, 29(8): 1969–1973.

Dunn-Walters, D. K. and Spencer, J. 1998. Strong intrinsic biases towards mutation and conservation of bases in human IgVH genes during somatic hypermutation prevent statistical analysis of antigen selection. *Immunology*, 95(3): 339–345.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17(6): 368–376.

Felsenstein, J. 2004. *Inferring phylogenies*. Sinauer Associates.

Georgiou, G., Ippolito, G. C., Beausang, J., Busse, C. E., Wardemann, H., and Quake, S. R. 2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.*

Gerhardt, S., Abbott, W. M., Hargreaves, D., Pauptit, R. A., Davies, R. A., Needham, M. R. C., Langham, C., Barker, W., Aziz, A., Snow, M. J., Dawson, S., Welsh, F., Wilkinson, T., Vaugan, T., Beste, G., Bishop, S., Popovic, B., Rees, G., Sleeman, M., Tuske, S. J., Coales, S. J., Hamuro, Y., and Russell, C. 2009. Structure of IL-17A in complex with a potent, fully human neutralizing antibody. *J. Mol. Biol.*, 394(5): 905–921.

Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, 11(5): 725–736.

Gotoh, O. 1982. An improved algorithm for matching biological sequences. *J. Mol. Biol.*, 162(3): 705–708.

Guéguen, L., Gaillard, S., Boussau, B., Gouy, M., Groussin, M., Rochette, N. C., Bigot, T., Fournier, D., Pouyet, F., Cahais, V., Bernard, A., Scornavacca, C., Nabholz, B., Haudry, A., Dachary, L., Galtier, N., Belkhir, K., and Dutheil, J. Y. 2013. Bio++: efficient extensible libraries and tools for computational molecular evolution. *Mol. Biol. Evol.*, 30(8): 1745–1750.

Hasegawa, M., Kishino, H., and Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22(2): 160–174.

Hershberg, U., Uduman, M., Shlomchik, M. J., and Kleinstein, S. H. 2008. Improved methods for detecting selection by mutation analysis of Ig V region sequences. *Int. Immunol.*, 20(5): 683–694.

Hollander, M. and Wolfe, D. 1973. *Nonparametric Statistical Methods*. Wiley.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scand. Stat. Theory Appl.*

Johnson, S. G. 2010. The NLopt nonlinear-optimization package.

Kosakovsky Pond, S. L., Scheffler, K., Gravenor, M. B., Poon, A. F., and Frost, S. D. 2010. Evolutionary fingerprinting of genes. *Mol. Biol. Evol.*, 27(3): 520–536.

Lanave, C., Preparata, G., Saccone, C., and Serio, G. 1984. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.*, 20(1): 86–93.

Larimore, K., McCormick, M. W., Robins, H. S., and Greenberg, P. D. 2012. Shaping of human germline IgH repertoires revealed by deep sequencing. *J. Immunol.*, 189(6): 3221–3230.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., Davis, I. W., Cooper, S., Treuille, A., Mandell, D. J., Richter, F., Ban, Y.-E. A., Fleishman, S. J., Corn, J. E., Kim, D. E., Lyskov, S., Berrondo, M., Mentzer, S., Popović, Z., Havranek, J. J., Karanicolas, J., Das, R., Meiler, J., Kortemme, T., Gray, J. J., Kuhlman, B., Baker, D., and Bradley, P. 2011. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.*, 487: 545–574.

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

Lefranc, M.-P., 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., and Duroux, P. 2008. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.*, 37(Database issue): D1006–12.

Lemey, P., Minin, V. N., Bielejec, F., Kosakovsky Pond, S. L., and Suchard, M. a. 2012. A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. *Bioinformatics*, 28(24): 3248–3256.

Li, H. and Wiehe, T. 2013. Coalescent tree imbalance and a simple test for selective sweeps based on microsatellite variation. *PLoS Comput. Biol.*, 9(5): e1003060.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., *et al.* 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370): 476–482.

Lossos, I. S., Tibshirani, R., Narasimhan, B., and Levy, R. 2000. The inference of antigen selection on Ig genes. *J. Immunol.*, 165(9): 5122–5126.

Mehr, R., Sternberg-Simon, M., Michaeli, M., and Pickman, Y. 2012. Models and methods for analysis of lymphocyte repertoire generation, development, selection and evolution. *Immunol. Lett.*, 148(1): 11–22.

Mooers, A. O. and Heard, S. B. 1997. Inferring evolutionary process from phylogenetic tree shape. *Q. Rev. Biol.*, 72(1): 31–54.

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., and Scheffler, K. 2013. FUBAR: A fast, unconstrained Bayesian AppRoximation for inferring selection. *Mol. Biol. Evol.*, 30(5): 1196–1205.

Muse, S. V. and Gaut, B. S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.*, 11(5): 715–724.

Powell, M. 2009. The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*.

Robbins, H. 1956. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Robins, H. 2013. Immunosequencing: applications of immune repertoire deep sequencing. *Curr. Opin. Immunol.*, 25(5): 646–652.

Robins, H. S., Campregher, P. V., Srivastava, S. K., Wacher, A., Turtle, C. J., Kahsai, O., Riddell, S. R., Warren, E. H., and Carlson, C. S. 2009. Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood*, 114(19): 4099–4107.

Salemi, M., Lemey, P., and Vandamme, A. M. 2009. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press.

Schwarz, G. 1978. Estimating the dimension of a model. *Ann. Stat.*, 6(2): 461–464.

Shahaf, G., Barak, M., Zuckerman, N. S., Swerdlin, N., Gorfine, M., and Mehr, R. 2008. Antigen-driven selection in germinal centers as reflected by the shape characteristics of immunoglobulin gene lineage trees: a large-scale simulation study. *J. Theor. Biol.*, 255(2): 210–222.

Six, A., Mariotti-Ferrandiz, M. E., Chaara, W., Magadan, S., Pham, H.-P., Lefranc, M.-P., Mora, T., Thomas-Vaslin, V., Walczak, A. M., and Boudinot, P. 2013. The past, present, and future of immune repertoire biology - the rise of Next-Generation repertoire analysis. *Front. Immunol.*, 4: 413.

Steiman-Shimony, A., Edelman, H., Hutzler, A., Barak, M., Zuckerman, N. S., Shahaf, G., Dunn-Walters, D., Stott, D. I., Abraham, R. S., and Mehr, R. 2006. Lineage tree analysis of immunoglobulin variable-region gene mutations in autoimmune diseases: Chronic activation, normal selection. *Cell. Immunol.*, 244(2): 130–136.

Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*.

Teng, G. and Papavasiliou, F. N. 2007. Immunoglobulin somatic hypermutation. *Annu. Rev. Genet.*, 41: 107–120.

Uduman, M., Shlomchik, M. J., Vigneault, F., Church, G. M., and Kleinstein, S. H. 2014. Integrating B cell lineage information into statistical tests for detecting selection in Ig sequences. *J. Immunol.*, 192(3): 867–874.

van den Boogaart, K. G. and Tolosana-Delgado, R. 2008. "Compositions": a unified R package to analyze compositional data. *Computers & Geosciences*, 34(4): 320–338.

Warren, E. H., Matsen, IV, F. A., and Chou, J. 2013. High-throughput sequencing of B- and T-lymphocyte antigen receptors in hematology. *Blood*, 122(1): 19–22.

Wu, Y.-C., Kipling, D., Leong, H. S., Martin, V., Ademokun, A. A., and Dunn-Walters, D. K. 2010. High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood*, 116(7): 1070–1078.

Xu, J. L. and Davis, M. M. 2000. Diversity in the CDR3 region of VH is sufficient for most antibody specificities. *Immunity*, 13(1): 37–45.

Yaari, G., Uduman, M., and Kleinstein, S. H. 2012. Quantifying selection in high-throughput immunoglobulin sequencing data sets. *Nucleic Acids Res.*, 40(17):

e134.

Yaari, G., Vander Heiden, J. A., Uduman, M., Gadala-Maria, D., Gupta, N., Stern, J. N. H., O'Connor, K. C., Hafler, D. A., Laserson, U., Vigneault, F., and Kleinstein, S. H. 2013. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front. Immunol.*, 4: 358.

Yang, Z. 1994a. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39(3): 306–314.

Yang, Z. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39(3): 306–314.

Zaharia, M., Chowdhury, M., Franklin, M., Shenker, S., and Stoica, I. 2010. Spark: cluster computing with working sets. In E. Nahum and D. Xu, editors, *2nd USENIX conference on hot topics in cloud computing*, page 10. static.usenix.org.

Zhu, J., Ofek, G., Yang, Y., Zhang, B., Louder, M. K., Lu, G., McKee, K., Pancera, M., Skinner, J., Zhang, Z., Parks, R., Eudailey, J., Lloyd, K. E., Blinn, J., Alam, S. M., Haynes, B. F., Simek, M., Burton, D. R., Koff, W. C., Mullikin, J. C., Mascola, J. R., Shapiro, L., Kwong, P. D., Becker, J., Benjamin, B., Blakesley, R., Bouffard, G., Brooks, S., Coleman, H., Dekhtyar, M., Gregory, M., Guan, X., Gupta, J., Han, J., Hargrove, A., Ho, S.-L., Johnson, T., Legaspi, R., Lovett, S., Maduro, Q., Masiello, C., Maskeri, B., McDowell, J., Montemayor, C., Mullikin, J., Park, M., Riebow, N., Schandler, K., Schmidt, B., Sison, C., Stantripop, M., Thomas, J., Thomas, P., Vemulapalli, M., and Young, A. 2013. Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proceedings of the National Academy of Sciences*.

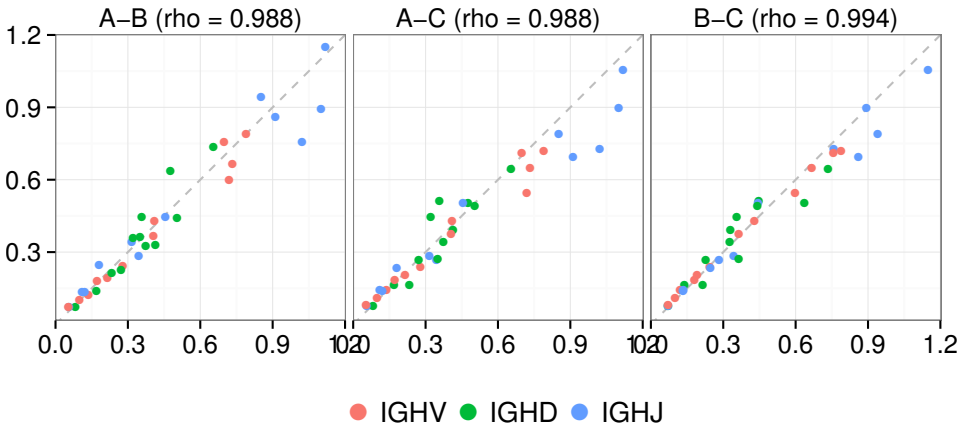

FIGURE S1. Pairwise comparison of off-diagonal entries in maximum-likelihood Q matrices under the $t_r\, Q_i\, \Gamma_i$ model between the three individuals. Coefficients are shown in Fig. 3.

SUPPLEMENTAL INFORMATION

	sequence count		
individual	raw	unique by well	unique overall
A	52,381,123	8,275,848	4,778,427
B	59,241,547	9,820,657	5,826,068
C	66,469,248	8,452,997	4,419,453

TABLE S1. Number of memory BCR sequences obtained by individual. "raw" refers to the number of reads obtained from sequencing, "unique by well" the number of unique reads after performing clustering on reads for each barcoded PCR well, and "unique overall" the total number of unique sequences in the sample.

		Individual A				Individual B				Individual C			
		A	G	C	T	A	G	C	T	A	G	C	T
IGHV	germline	0.283	0.27	0.255	0.192	0.279	0.27	0.261	0.19	0.285	0.268	0.258	0.189
	read	0.277	0.261	0.256	0.206	0.276	0.266	0.261	0.197	0.282	0.265	0.258	0.196
IGHD	germline	0.199	0.328	0.141	0.332	0.196	0.323	0.157	0.324	0.197	0.326	0.153	0.324
	read	0.197	0.315	0.168	0.321	0.198	0.309	0.176	0.317	0.197	0.314	0.172	0.317
IGHJ	germline	0.197	0.428	0.22	0.154	0.2	0.424	0.223	0.154	0.186	0.438	0.225	0.151
	read	0.186	0.433	0.222	0.159	0.193	0.427	0.224	0.156	0.18	0.44	0.227	0.153

TABLE S2. Empirical stationary distribution for germline and observed reads.

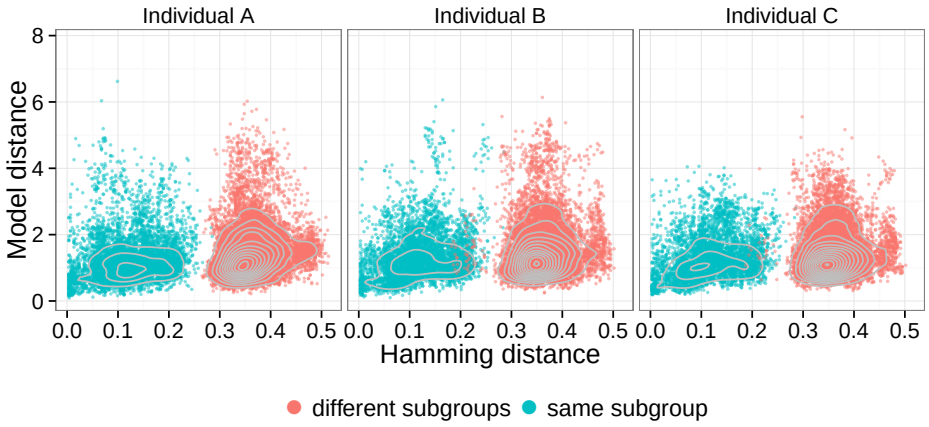

FIGURE S2. Comparison of Hamming distance between V genes (x-axis) and Euclidean distance between centered log-transformed median time transition matrices for productive rearrangements (y-axis). Colors indicate whether the V genes in a comparison come from the same or different subgroups. The correlation between the two was significant ($p < 10^{-15}$, Spearman's $\rho = 0.197$).

SUPPLEMENTAL METHODS

Virtual machine for reproducing analyses. We have made an Amazon Machine Image (AMI) (Amazon Web Services, 2014) available pre-loaded with our analysis pipeline and some example data. To use it, launch an instance of AMI `ami-ab295b9b` in the `us-west-2` region and log in as user `ubuntu` (no password needed: authentication by public key).

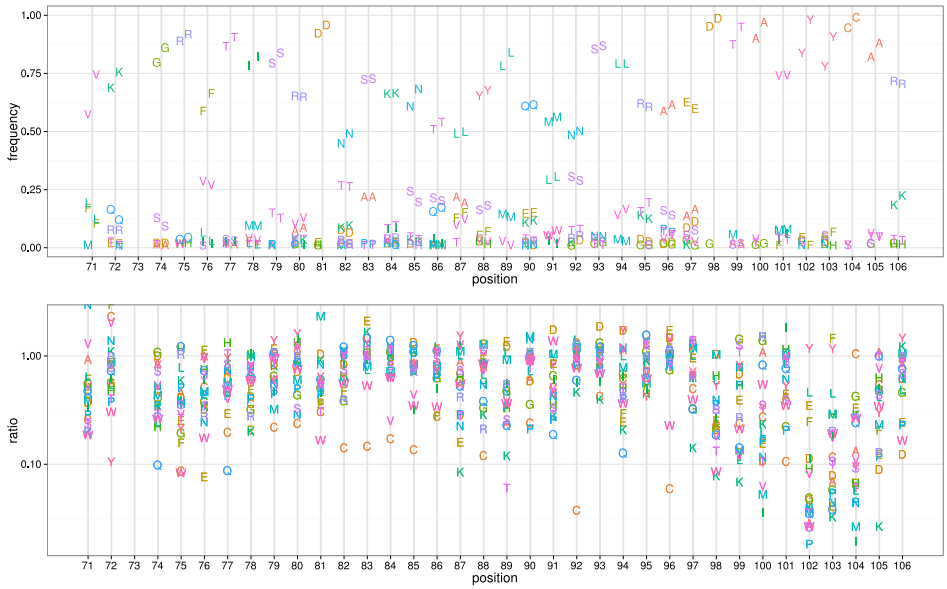

FIGURE S3. Amino acid profiles of out-of-frame and functional B cell sequences as aligned by the IMGT alignment. Top panel: frequency of amino acids per site. Letters to the left of the line show the profile for out-of-frame sequences and those to the right of the line show the profile for functional sequences. Bottom panel: amino acid frequency in functional sequences divided by that in out-of-frame sequences.

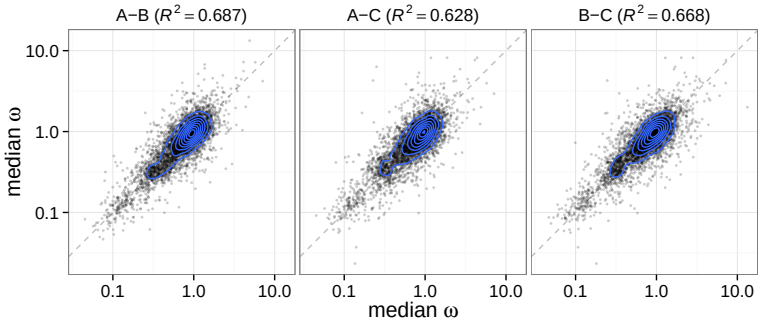

FIGURE S4. Pairwise comparisons of site-specific ω estimates between the three individuals along with the R^2 value from a linear model fit using $\log_{10}(\omega)$ for both the predictor (x-axis) and response (y-axis).

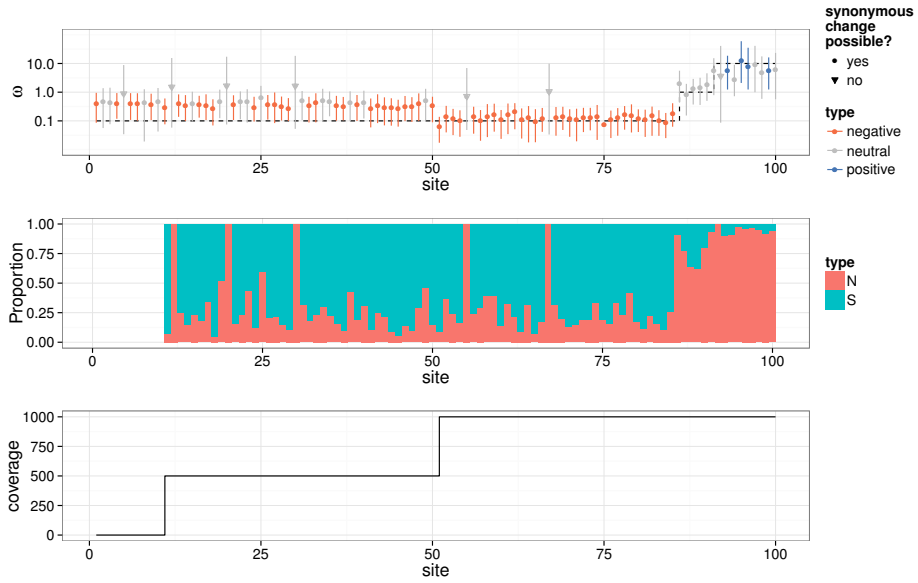

FIGURE S5. Top panel: site-specific ω estimates under simulated data with varying coverage. Inverted trianges show sites where the germline state was Tryptophan or Methionine, from which no synonymous changes are possible. Dashed black line shows simulated ω. Middle panel: proportion (second panel) of mutations at each position which were nonsynonymous (N) or synonymous (S). Bottom panel: read coverage by codon position.

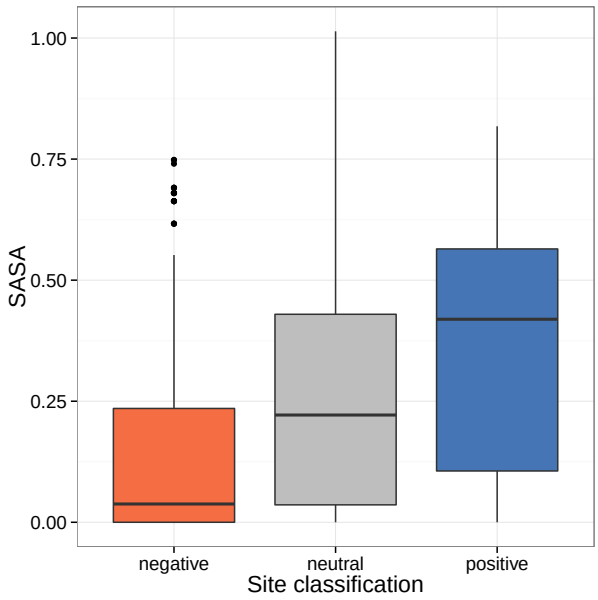

FIGURE S6. Normalized solvent-accessible surface area (SASA) values by per-site ω classification. A SASA value of 1.0 indicates that the residue is fully exposed, while a value of 0.0 indicates that the residue is buried. Sites under negative selection are significantly less exposed than sites under positive selection ($p < 10^{-12}$) or neutral selection ($p < 10^{-15}$) by Bonferroni-corrected Wilcoxon rank-sum test. Neutral sites were less exposed than sites under positive selection ($p < 0.002$).