SUBSTITUTION AND SITE-SPECIFIC SELECTION DRIVING B CELL AFFINITY MATURATION IS CONSISTENT ACROSS INDIVIDUALS

CONNOR O. MCCOY, TREVOR BEDFORD, VLADIMIR N. MININ, HARLAN ROBINS, AND FREDERICK A. MATSEN IV

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 purifying selection, though not uniformly so.

INTRODUCTION

The spectrum of antibodies encoded by somatically rearranged human B cell receptor (BCR) genes can bind a vast array of pathogens, initiating an immune reponse or directly neutralizing their target. This diversity is made possible by the processes of *VDJ recombination*, in which combinatorial joining of V-, D-, and J-genes generates a vastly broader collection of unique BCR sequences than could be directly encoded in the genome, and then further diversification and selection by *affinity maturation*. 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 [1–5]. Recent reviews have highlighted the need for new computational tools that make use of BCR sequence data

Date: March 13, 2014.

to bring new insight, including the need for reproducible computational pipelines [6–9]. 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, 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 [10]; see also [11]), 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 [12]. 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 [11]. 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 are divided into nested sets. They can be first identified by their *segment*, which is whether they are a V, D, or J gene. Each set of genes for a given segment is divided into *subgroups* which share at least 75% nucleotide identity. Genes also have polymorphisms that are grouped into *alleles*, which are variants of the gene between individuals.

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, first we 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. Second, we develop the first statistical methodology for comprehensive per-residue selection estimates for B cell receptors. We avoid difficulties encountered by previous work in differentiating between selection and motif-driven mutation by developing statistical means to compare in-frame and out-of-frame rearrangements. A key part of our method is an empirical Bayes regularization process for selection inference, which we develop for non-constant sequencing coverage and adapt to the case of a fixed ancestral sequence. Using this new method, we are able to derive a per-residue map of selection, which we find is dominated by purifying 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.

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 changes that have occurred in going from germline sequences to 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 unique 130bp reads spanning the heavy chain CDR3 region. 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% [13]. 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 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, we performed model fitting on a set of two-taxon trees, each connecting an observed read to its best scoring germline sequence according to Smith-Waterman alignment. This best scoring sequence was taken to be its ancestor. We then calculated model likelihood as the product of likelihoods over these two taxon trees. We evaluated the fit of models with varying complexity, ranging from a simple model with shared branch lengths and rates for the three independent segments of the BCR to a complex model with independent branch lengths for each segment (Tab. 1). For the underlying nucleotide substitution model, we fit a general time-reversible (GTR) nucleotide model [14] to subsets of the data, using 20,000 unique sequences from each individual. The choice of a reversible model, rather than a more general model, was based on the similarity of base frequencies between the germline sequences and observed reads (Tab. S1).

	model	log likelihood	df	AIC	ΔAIC
A	$t_r\, Q_i\, \Gamma_i$	-650,916	20,029	1,341,890	
	$t_r\, Q_i\, \Gamma_s$	-651,307	20,027	1,342,668	779
	$t_r\, Q_s\, \Gamma_s$	-663,795	20,009	1,367,608	25,719
	$t_i\, Q_i\, \Gamma_i$	-626,713	58,546	1,370,518	28,628
B	$t_r\, Q_i\, \Gamma_i$	-481,265	20,029	1,002,589	
	$t_r\, Q_i\, \Gamma_s$	-481,497	20,027	1,003,048	459
	$t_r\, Q_s\, \Gamma_s$	-490,356	20,009	1,020,729	18,141
	$t_i\, Q_i\, \Gamma_i$	-457,040	58,889	1,031,857	29,269
C	$t_r\, Q_i\, \Gamma_i$	-542,673	20,029	1,125,403	
	$t_r\, Q_i\, \Gamma_s$	-542,814	20,027	1,125,683	279
	$t_r\, Q_s\, \Gamma_s$	-551,763	20,009	1,143,545	18,141
	$t_i\, Q_i\, \Gamma_i$	-519,186	59,125	1,156,621	31,218

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 rates of evolution (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) [15] (Tab. 2) and the Bayesian Information Criterion [16] (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 0.062, 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. 1). 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 IGHV, IGHD, and IGHJ segments for each individual was very similar.

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. 2, 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 [17].

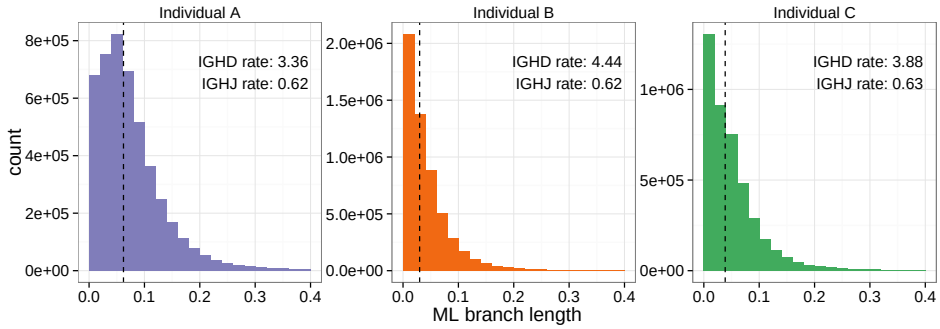

FIGURE 1. Distribution of maximum likelihood branch lengths estimated under the $t_r\, Q_i\, \Gamma_i$ model. Branch lengths are measured in terms of substitutions per site, and rates given for IGHD and IGHJ are relative to a fixed rate of 1 for IGHV.

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 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 [18] (see Materials and Methods). We find that matrices 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. 3). The Euclidean distance between these transformed discrete probability distributions and the Hamming distance between germline IGHV genes showed significant, but moderate, correlation (Spearman's $\rho = 0.20$, $p < 10^{-15}$; Fig. S2).

Selection. Because of the large volume of sequences to analyze, we needed a mechanism to detect selection that could be run on over 15 million sequences. Classical means of estimating selection by codon model fitting [19, 20] could not be used, even in their most recent and much more efficient recent incarnation [21]. Instead, we used a version of the renaissance counting approach [22], which we modified to work under varying levels of coverage and fixed initial state. As above, we use a two taxon tree for each read, here consisting of the read and matching IGHV segment. The counting renaissance method simulates ancestral substitutions under a simple (nucleotide) model to estimate parameters of a more complex (codon) model.

A key part of the renaissance counting approach is an empirical Bayes regularization procedure [23]. This procedure uses the entire collection of sites to inform parameter estimation for each site individually. This effectively shares data across sites, allowing inferences about sites which either display few substitutions

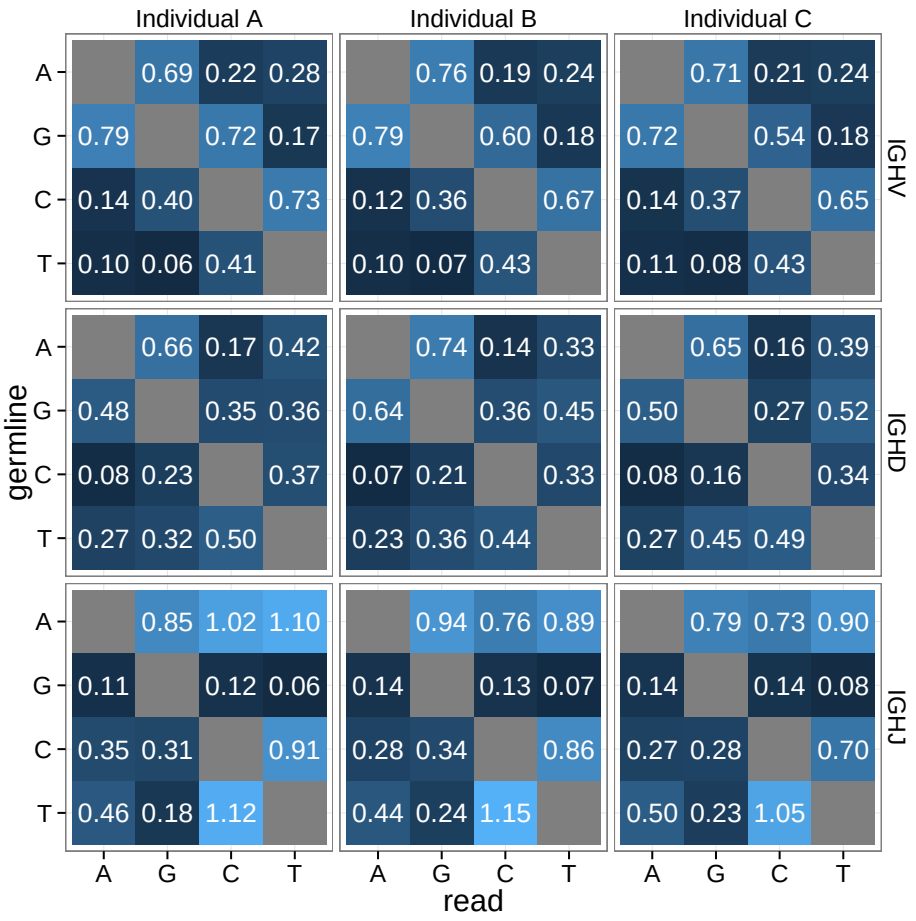

FIGURE 2. GTR coefficients for the $t_r\, Q_i\, \Gamma_i$ model estimated under maximum likelihood. Rows index the nucleotide found in the germline sequence, whereas columns index the nucleotide found in the observed sequence.

or have less read coverage. For example, we might never observe a given site to be mutated in our data. Here, the empirical Bayes procedure would shrink estimates of selection ratios at this site toward the population average.

Previous work [22] used this strategy to infer selection pressure in the case of constant sequencing coverage for a given region. However, our data set had very uneven read depth coverage, and by truncating to the shortest read we would have lost valuable information. Thus, we developed a new method to infer selection pressure in the case of non-constant coverage, and validated it via simulation (see Materials and Methods). We infer selection using a nonsynonymous-synonymous ratio which controls for background mutation rate via out-of-frame sequences. We do so because motif-driven evolution combined with the genetic code can result in high dN/dS ratios that are not attributable to selection.

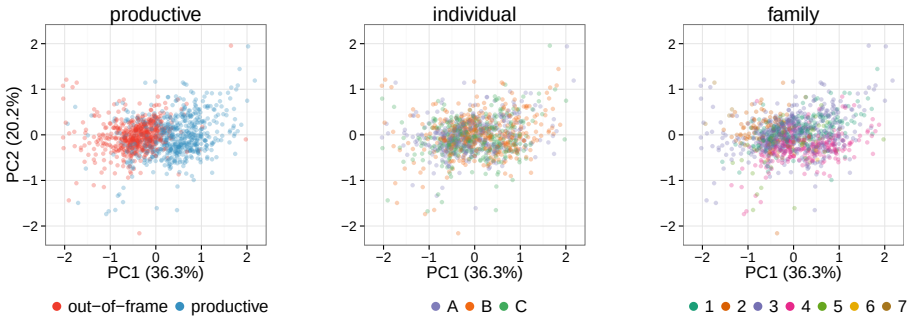

FIGURE 3. First two components of PCA performed on centered log-transformed median-time transition matrices for V gene segments. Points plotted in a random order, with 22 outliers removed for clarity.

Applying this method to our data set results in a collection of per-site and per-gene maps quantifying selection (results uploaded to Dryad data repository associated with this paper). We employ site numbering according to the IMGT unique numbering for the V domain [24]. Sites were classified as purifying or diversifying based on whether the 95% Bayesian credible interval (BCI) excludes one; sites satisfying this criteria that have the lower bound of their ω BCI greater than one are classified as being under diversifying selection and those that have the upper bound of their ω BCI less than one are classified as being under purifying selection.

IGHV3-23*01 is the most frequent V gene/allele combination, and it displays patterns that are consistent with the other genes. Specifically, we see significant variation in the synonymous substitution rate (right panels, Fig. 4a), 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. Indeed, at those sites where the synonymous rate is very high or very low, we see spiking of the unconditional/conditional ratio which is equivalent to the classical dN/dS ratio (upper panel, Fig. 4b). On the other hand, the selection inferences made using out-of-frame sequences stay much closer to neutral (lower panel, Fig.4b).

We note extensive purifying selection in the residues immediately preceding the CDR3 (Fig. 5). 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. 6). The mean ω estimate across sites with at least 100 productive and out-of-frame reads aligned

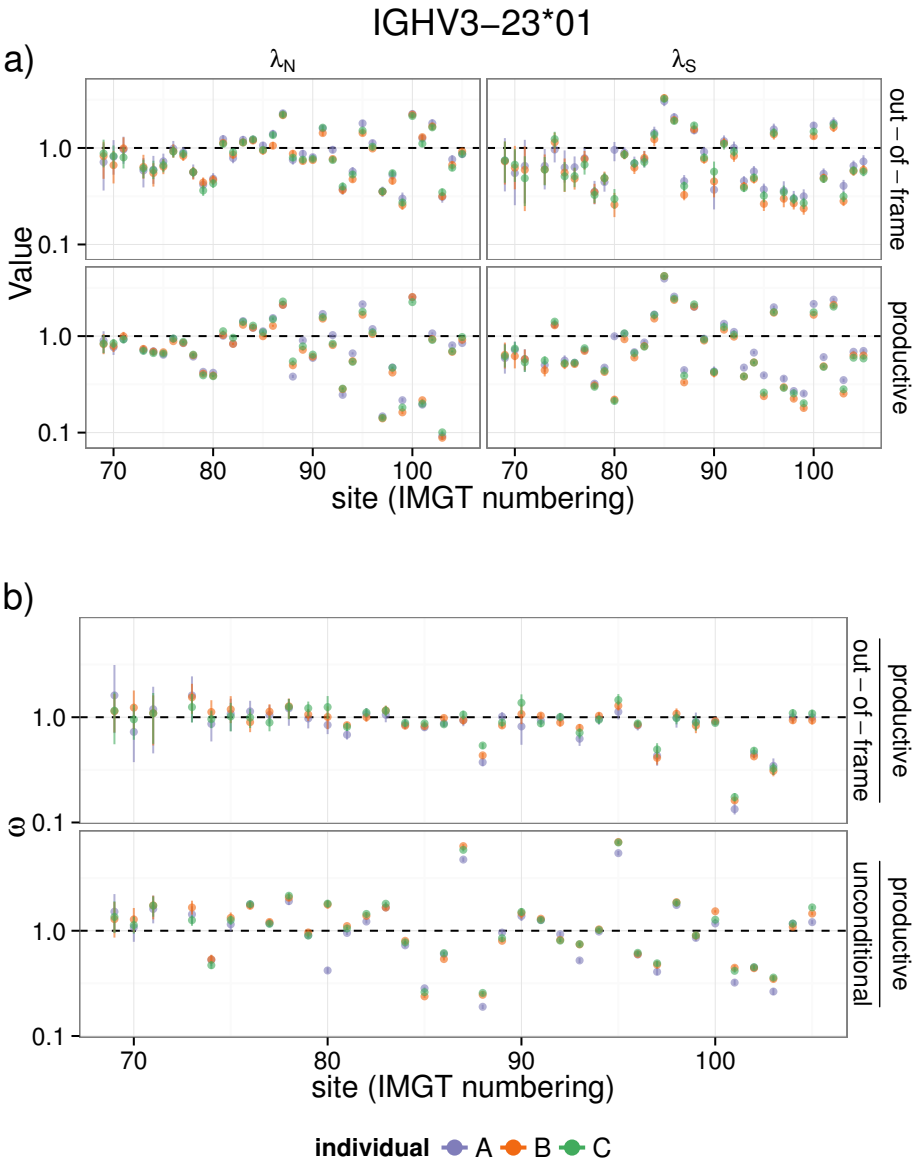

FIGURE 4. a) Comparison of nonsynonymous (λ_N) and synonymous (λ_S) rates in productive and out-of-frame sequences. b) Comparison of ω estimates using either unconditional substitution rates or unproductive rearragements as a proxy for the neutral process. Both panels use data from IGHV3-23*01, the most frequent V gene/allele combination.

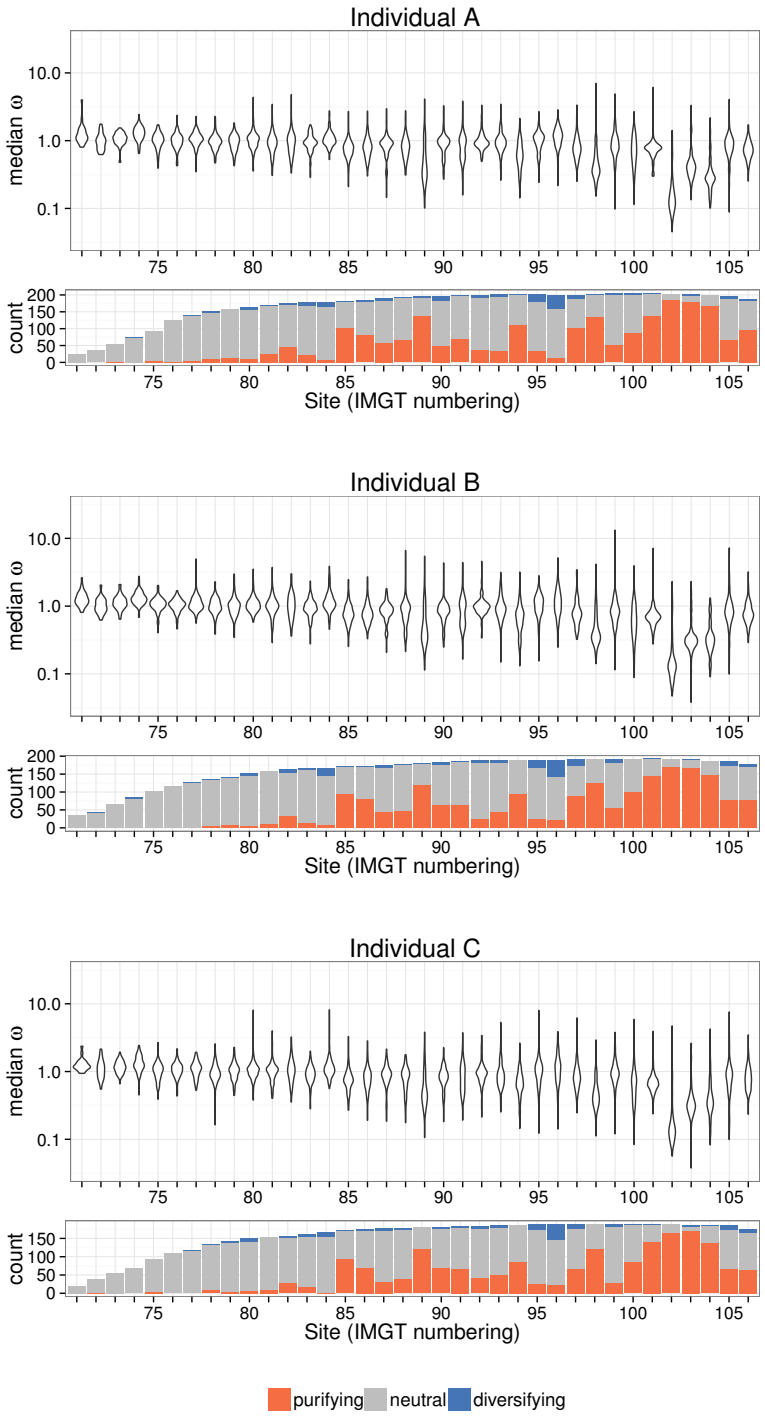

FIGURE 5. Site-specific estimates of ω. Violin plots show distribution of median ω estimates across IGHV genes at each site. Bar plots show count of IGHV genes classified as undergoing purifying, neutral, or diversifying selection.

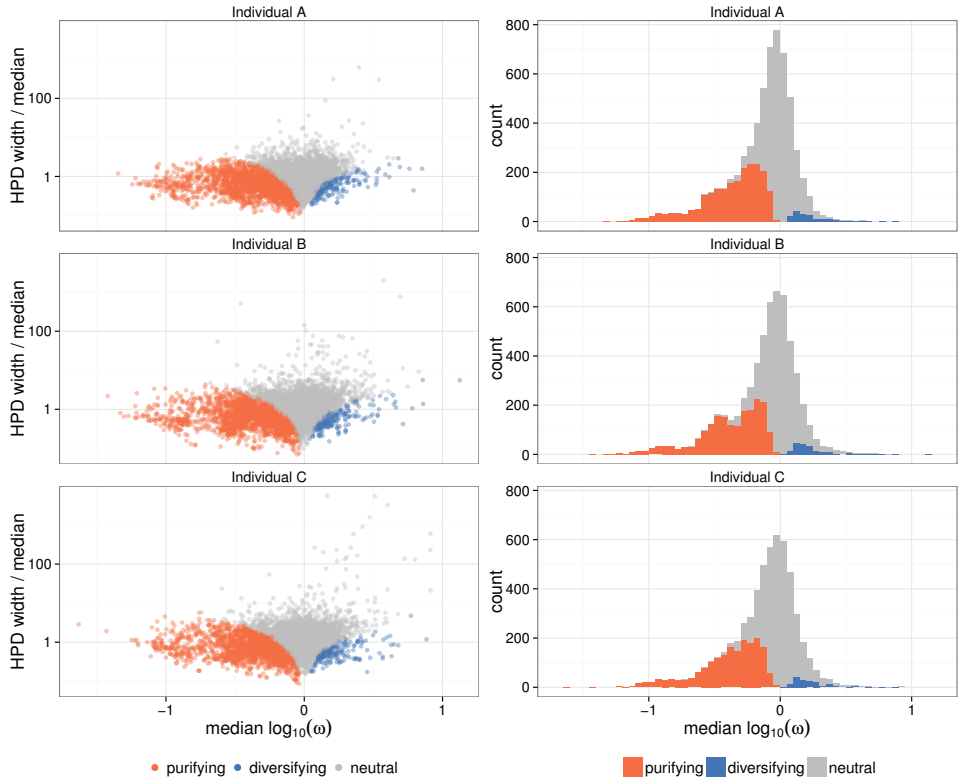

FIGURE 6. Site-specific selection estimates partitioned by individual and gene. Sites classified as purifying or diversifying based on whether the 95% Bayesian credible interval excludes 1 and in what direction. Left: comparison of ω estimate and relative width of HPD region. Right: distribution of site-specific selection estimates.

was 0.907, with 65.6% of sites having a median $\omega < 1$; 30.6% of sites were confidently classified as being under purifying selection.

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 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. 2) even though those individuals have differing levels of substitution (Fig. 1). 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* [25] and mammals [26], 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.

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 [17]). 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 [11]. Yaari et al. [11] 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). 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. 3), and we find a significant correlation between sequence identity and substitution matrix.

Our multivariate analysis of GTR matrices is inspired by previous work on "evolutionary fingerprinting," in which genes are characterized by their substitution patterns rather than their sequence identity [27]. However, our interests and methods differ from this previous work, even beyond that Kosakovsky Pond et al. [27] considered pathogen evolution and we are looking at B cells. Whereas their paper was interested in patterns of selection, we investigate single nucleotide transition matrices. Because these matrices have many parameters and are combinations of probability vectors, we use a variant of compositional PCA rather than kernel PCA with earth-mover's distance as they did.

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 3. 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 [28], 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 [29–33]. Recent work has offered methods that evaluate selection on a per-sequence basis [33]. There have also been efforts to infer selection based on lineage shape [34–38], which has been a common approach in macroevolutionary studies (reviewed in [39]) and more recently in population genetics [40, 41].

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 [12, 29–33] 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 [12]. We also note that [11] 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 [17]. Furthermore, 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. 6).

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 purifying selection among sites for which selection could be confidently classified. We see an pattern of quite strong purifying selection in the region around the beginning of the CDR3. This agrees with recent work that also found strong purifying selection in one site near the beginning of the CDR3 [11]. 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 [11].

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 and adapt to the case of a fixed ancestral sequence. 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. 440ml of blood was drawn from three healthy volunteers under IRB protocol at the Fred Hutchinson Cancer Research Center. Over 10 million naïve (CD19+CD27-IgD+IgM+) and over 10 million memory (CD19+CD27+) B cells were then bead purified from each donor. Genomic DNA was extracted and the ImmunoSeq assay was performed on the six samples at Adaptive Biotechnologies in Seattle, WA. Each sample was divided amongst the wells on two 96 well plates and bar-coded individually. The complete description of the experiment and a full set of visualization and analysis tools will published in a separate manuscript.

Preprocessing, alignment and germline assignment. We corrected PCR and sequencing errors using the methods in [42]. Briefly, we clustered all sequences into

groups with Hamming distance less than or equal to two, and the underlying sequence in each cluster was inferred using parsimony. Groups with only one member were discarded. We used only sequences from the memory B cell population.

Each sequence read was aligned to each IGHV gene using Smith-Waterman algorithm with an affine gap penalty [43]. The 3' portion of the sequence not included in the best IGHV alignment was next aligned to all D and J genes available from the IMGT database [44]. 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. The 18 IGHV 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. 1) 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 describes how rates-across-sites variation is modeled in terms of a four category discrete gamma distribution [45]. 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.

Maximum likelihood values of substitution model parameters and branch lengths were estimated using a combination of Bio++ [46], BEAGLE [47], with model optimization via the BOBYQA algorithm [48] as implemented in NLopt [49], and branch length optimization via Brent's method [50]. Optimization alternated between substitution model parameters and branch lengths until the change in log-likelihood at a given iteration was less than 0.5. 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 [18]. 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` [51], 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 IGHV genes using the alignment available from the IMGT database [44]. 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 [22], to accommodate pairwise analyses of a large number of sequences with a known ancestral sequence and non-constant site coverage. The original counting renaissance method starts by assuming a codon position-specific HKY substitution model [52] 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 parameters of the HKY model, along with the relative rates between positions, to the maximum likelihood estimates produced using the whole dataset. *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 the posterior distribution of query branch lengths, the counting renaissance proceeds by imputing unobserved codon substitutions conditional on the observed data and without such conditioning. We review these two imputation steps in the next section.

Sampling codon substitutions conditional on data. For each codon position l and posterior sample j, counts of synonymous ($C_{jl}^{(S-C)}$) and non-synonymous ($C_{jl}^{(N-C)}$) substitutions at each site were imputed using stochastic mapping as described in [22]. Counts unconditional on the data ($C_{jl}^{(S-U)}$, $C_{jl}^{(N-U)}$) were imputed using the following modification of the original counting renaissance. In the classical setting of estimating dN and dS (reviewed in [53]), the root state of the tree is unknown. In the case of somatically hypermutated B cell receptors, however, the root state is known to be the germline segment. Thus, we do not think of the root state as being sampled, and the "unconditional" counts are conditioned on the ancestral state. This is implemented in the BEAST by constructing a tree such that the reference sequence is fixed with a negligible branch length to the root. The branch length t to the query is then sampled via MCMC (Fig. 7(a)). When sampling "unconditional" counts, we only sample substitutions on the pendant edge leading to the query sequence. These counts are conditional on the state at the root, which is fixed to the state of the reference sequence at each site (Fig. 7(b)). This differs from the original implementation (Fig. 7(c)), which sampled the starting state for each site from the assumed initial distribution (e.g., codon frequencies induced by codon position-specific nucleotide models) [22].

For N MCMC iterations based on an alignment of L codons, the result of this procedure was four $N \times L$ matrices, each containing the number of events of a given type 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.

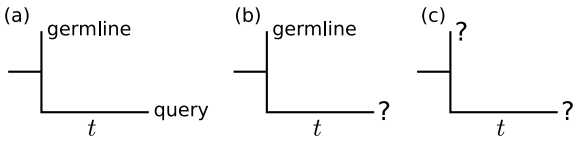

FIGURE 7. Different ways of conditioning for doing mutation sampling, shown as a phylogenetic tree with one branch length zero (leading to a germline sequence) and the other of length t. (a) Conditioning on the germline sequence and the query (observation). (b) Conditioning on the germline sequence only, which is our notion of "unconditional" counts. (c) The original implementation of renaissance counting, which did not condition on any sequences because it did not have a fixed ancestral sequence.

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 [22] 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 [22], 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 [48].

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 maximimizing 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 [54]. 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 [22], 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 a rate unconditional (U) on the data: $\omega_l^{RC} = \frac{\lambda_l^{(N-C)}/\lambda_l^{(N-U)}}{\lambda_l^{(S-C)}/\lambda_l^{(S-U)}}$. As SHM is highly context-specific, we used 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:

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

where I and O refer to in-frame and out-of-frame rearrangements, respectively.

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 $\text{Gamma}(\alpha, \beta)$ 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 [55] implementation of the counting renaissance to generate samples of conditional and unconditional counts for both synonymous and nonsynonymous substitutions at each site. We modified BEAST 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 [56] 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 for this work 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).

COMPETING INTERESTS

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

REFERENCES

[1] Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, et al. (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. Sci Transl Med 1: 12ra23–12ra23.

[2] Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, et al. (2010) High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. Blood 116: 1070–1078.

[3] Larimore K, McCormick MW, Robins HS, Greenberg PD (2012) Shaping of human germline IgH repertoires revealed by deep sequencing. J Immunol 189: 3221–3230.

[4] DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, et al. (2013) High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. Nat Biotechnol 31: 166–169.

[5] Robins H (2013) Immunosequencing: applications of immune repertoire deep sequencing. Curr Opin Immunol 25: 646–652.
[6] Mehr R, Sternberg-Simon M, Michaeli M, Pickman Y (2012) Models and methods for analysis of lymphocyte repertoire generation, development, selection and evolution. Immunol Lett 148: 11–22.
[7] Six A, Mariotti-Ferrandiz ME, Chaara W, Magadan S, Pham HP, et al. (2013) The past, present, and future of immune repertoire biology - the rise of Next-Generation repertoire analysis. Front Immunol 4: 413.
[8] Warren EH, Matsen FA 4th, Chou J (2013) High-throughput sequencing of B- and T-lymphocyte antigen receptors in hematology. Blood 122: 19–22.
[9] Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, et al. (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotechnol .
[10] Delker RK, Fugmann SD, Papavasiliou FN (2009) A coming-of-age story: activation-induced cytidine deaminase turns 10. Nat Immunol 10: 1147–1153.
[11] Yaari G, Vander Heiden JA, Uduman M, Gadala-Maria D, Gupta N, et al. (2013) Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. Front Immunol 4: 358.
[12] Dunn-Walters DK, 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: 339–345.
[13] Zhu J, Ofek G, Yang Y, Zhang B, Louder MK, et al. (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 .
[14] Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20: 86–93.
[15] Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr 19: 716–723.
[16] Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6: 461–464.
[17] Teng G, Papavasiliou FN (2007) Immunoglobulin somatic hypermutation. Annu Rev Genet 41: 107–120.
[18] Aitchison J (1983) Principal component analysis of compositional data. Biometrika 70: 57–65.
[19] Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725–736.
[20] Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11: 715–724.
[21] Murrell B, Moola S, Mabona A, Weighill T, Sheward D, et al. (2013) FUBAR: A fast, unconstrained bayesian AppRoximation for inferring selection. Mol Biol Evol 30: 1196–1205.
[22] Lemey P, Minin VN, Bielejec F, Kosakovsky Pond SL, Suchard Ma (2012) A counting renaissance: combining stochastic mapping and empirical Bayes to quickly detect amino acid sites under positive selection. Bioinformatics 28: 3248–3256.
[23] 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.
[24] Lefranc MP, Pommié C, Ruiz M, Giudicelli V, Foulquier E, et al. (2003) IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol 27: 55–77.
[25] Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203–218.
[26] Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476–482.
[27] Kosakovsky Pond SL, Scheffler K, Gravenor MB, Poon AF, Frost SD (2010) Evolutionary fingerprinting of genes. Mol Biol Evol 27: 520–536.
[28] Clarke SH, Huppi K, Ruezinsky D, Staudt L, Gerhard W, et al. (1985) Inter- and intraclonal diversity in the antibody response to influenza hemagglutinin. J Exp Med 161: 687–704.
[29] Chang B, 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: 367–373.

[30] Lossos IS, Tibshirani R, Narasimhan B, Levy R (2000) The inference of antigen selection on Ig genes. J Immunol 165: 5122–5126.
[31] Bose B, Sinha S (2005) Problems in using statistical analysis of replacement and silent mutations in antibody genes for determining antigen-driven affinity selection. Immunology 116: 172–183.
[32] Hershberg U, Uduman M, Shlomchik MJ, Kleinstein SH (2008) Improved methods for detecting selection by mutation analysis of Ig V region sequences. Int Immunol 20: 683–694.
[33] Yaari G, Uduman M, Kleinstein SH (2012) Quantifying selection in high-throughput immunoglobulin sequencing data sets. Nucleic Acids Res 40: e134.
[34] Steiman-Shimony A, Edelman H, Hutzler A, Barak M, Zuckerman NS, et al. (2006) Lineage tree analysis of immunoglobulin variable-region gene mutations in autoimmune diseases: Chronic activation, normal selection. Cell Immunol 244: 130–136.
[35] Abraham RS, Manske MK, Zuckerman NS, Sohni A, Edelman H, et al. (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: 69–87.
[36] Barak M, Zuckerman N, Edelman H, Unger R, Mehr R (2008) IgTree (c) : Creating immunoglobulin variable region gene lineage trees. Journal of Immunological Methods 338: 67-74.
[37] Shahaf G, Barak M, Zuckerman NS, Swerdlin N, Gorfine M, et al. (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: 210–222.
[38] Uduman M, Shlomchik MJ, Vigneault F, Church GM, Kleinstein SH (2014) Integrating B cell lineage information into statistical tests for detecting selection in Ig sequences. J Immunol 192: 867–874.
[39] Mooers AO, Heard SB (1997) Inferring evolutionary process from phylogenetic tree shape. Q Rev Biol 72: 31–54.
[40] Drummond AJ, Suchard MA (2008) Fully Bayesian tests of neutrality using genealogical summary statistics. BMC Genet 9: 68.
[41] Li H, Wiehe T (2013) Coalescent tree imbalance and a simple test for selective sweeps based on microsatellite variation. PLoS Comput Biol 9: e1003060.
[42] Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, et al. (2009) Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. Blood 114: 4099–4107.
[43] Gotoh O (1982) An improved algorithm for matching biological sequences. J Mol Biol 162: 705–708.
[44] Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, et al. (2008) IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res 37: D1006–12.
[45] Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39: 306–314.
[46] Guéguen L, Gaillard S, Boussau B, Gouy M, Groussin M, et al. (2013) Bio++: efficient extensible libraries and tools for computational molecular evolution. Mol Biol Evol 30: 1745–1750.
[47] Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, et al. (2011) BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. Syst Biol 61: 170–173.
[48] Powell M (2009) The BOBYQA algorithm for bound constrained optimization without derivatives. Cambridge NA Report NA2009/06, University of Cambridge, Cambridge .
[49] Johnson SG (2010). The NLopt nonlinear-optimization package.
[50] Brent RP (1973) Algorithms for Minimization Without Derivatives. Courier Dover Publications.
[51] van den Boogaart KG, Tolosana-Delgado R (2008) "Compositions": a unified R package to analyze compositional data. Computers & Geosciences 34: 320–338.
[52] Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160–174.
[53] Yang, Bielawski (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496–503.
[54] Bielejec F, Lemey P, Carvalho LM, Baele G, Rambaut A, et al. (2013) piBUSS: a parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios. 1312.4699.
[55] Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29: 1969–1973.
[56] Zaharia M, Chowdhury M, Franklin M, others (2010) Spark: cluster computing with working sets. in cloud computing .

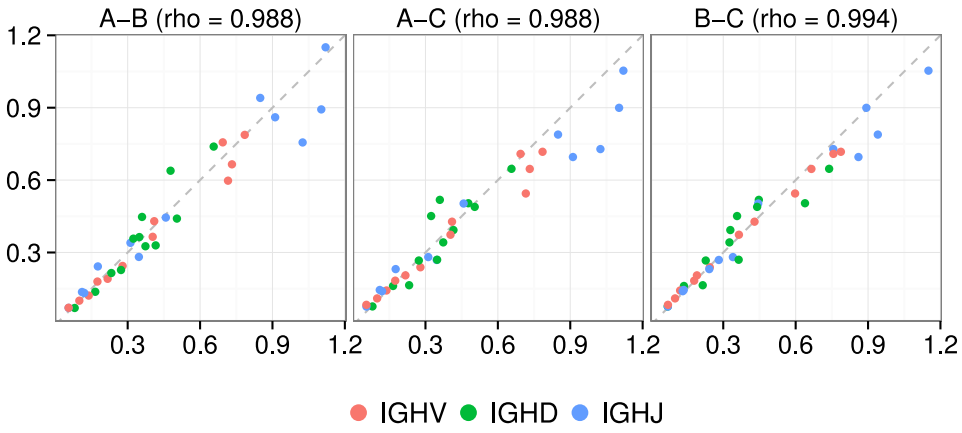

FIGURE S1. Pairwise comparison of off-diagonal entries in maximum-likelihood Q matrices between the three individuals. Coefficients are shown in Fig. 2.

SUPPLEMENTAL INFORMATION

		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 S1. Empirical stationary distribution for germline and observed reads.

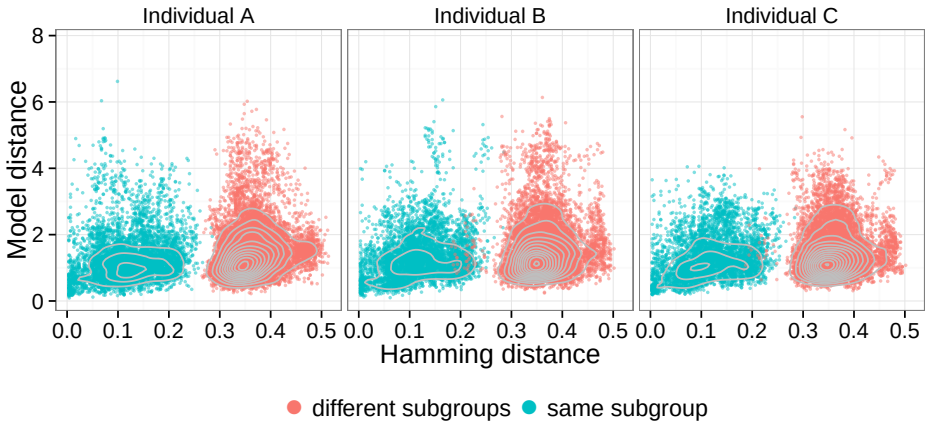

FIGURE S2. Comparison of Hamming distance between IGHV genes (x-axis) and Euclidean distance between centered log-transformed median time transition matrices for productive rearrangements (y-axis). Colors indicate whether the IGHV genes in a comparison come from the same or different subgroups. The correlation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

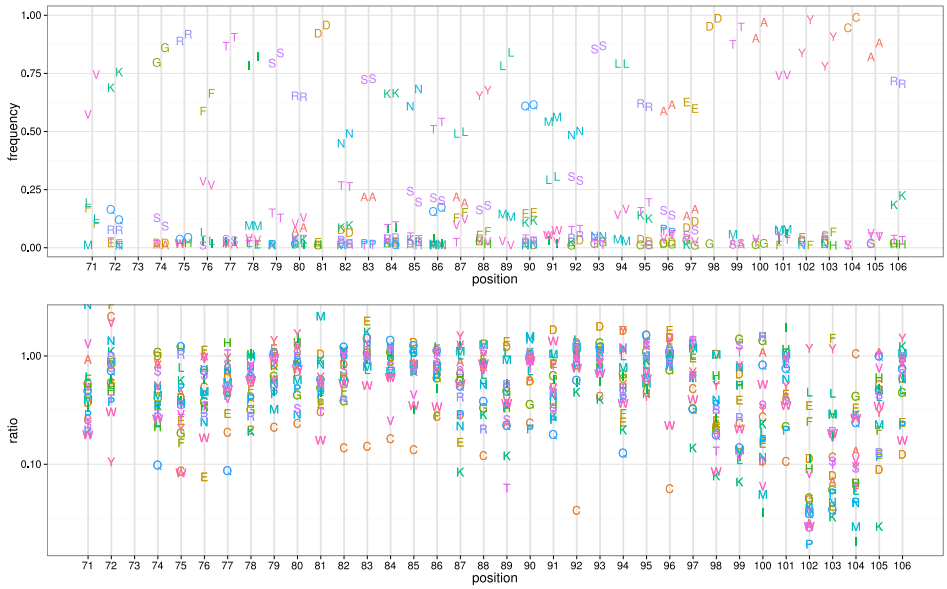

FIGURE S3. Amino acid profiles of out-of-frame and functional B cell sequences as aligned by the IMGT alignment. Top panel: frequency of amino acids per site. Letters to the left of the line show the profile for out-of-frame sequences and those to the right of the line show the profile for functional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

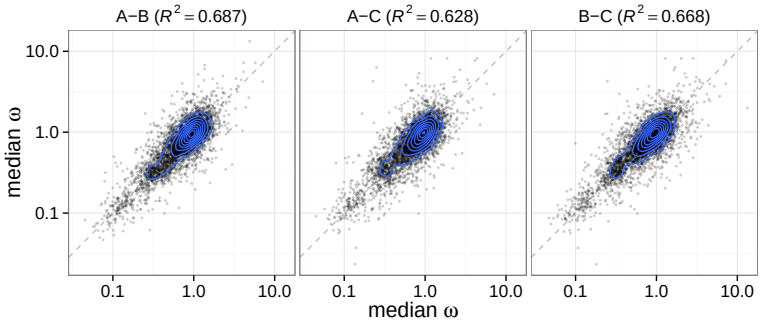

FIGURE S4. Pairwise comparisons of site-specific ω estimates between the three individuals along with the R^2 value from a linear model fit using $\log_{10}(\omega)$.

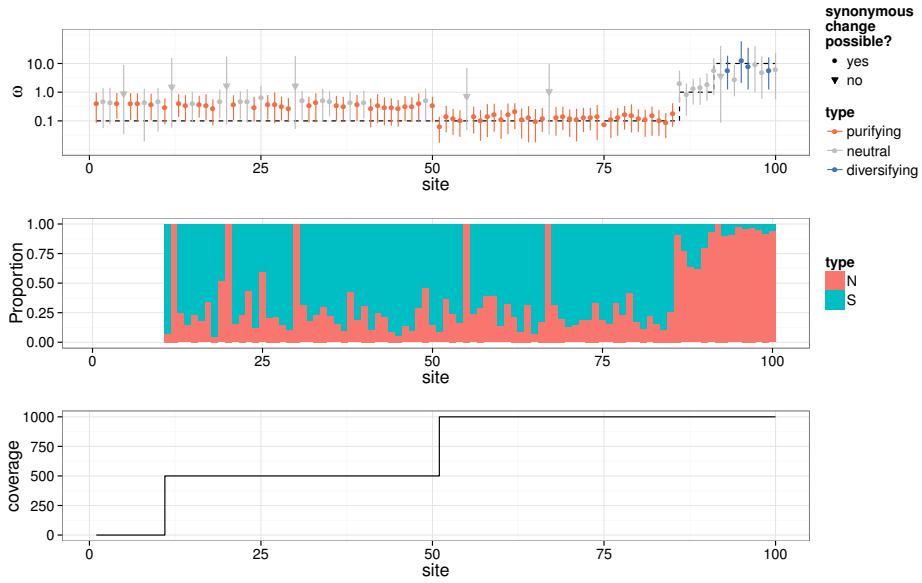

FIGURE S5. Top panel: site-specific ω estimates under simulated data with varying coverage. Inverted trianges show sites where the germline state was Tryptophan or Methionine, from which no synonymous changes are possible. Dashed black line shows simulated ω. Middle panel: proportion (second panel) of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