PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Research

CrossMark
click for updates

Cite this article: McCoy CO, Bedford T, Minin VN, Bradley P, Robins H, Matsen IV FA. 2015 Quantifying evolutionary constraints on B-cell affinity maturation. *Phil. Trans. R. Soc. B* **370**: 20140244.
http://dx.doi.org/10.1098/rstb.2014.0244

Accepted: 22 May 2015

One contribution of 13 to a theme issue 'The dynamics of antibody repertoires'.

Subject Areas:
immunology, evolution

Keywords:
B cell, antibody, immunoglobulin, affinity maturation, molecular evolution, natural selection

Author for correspondence:
Frederick A. Matsen IV
e-mail: matsen@fredhutch.org

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2014.0244 or via http://rstb.royalsocietypublishing.org.

THE ROYAL SOCIETY
PUBLISHING

Quantifying evolutionary constraints on B-cell affinity maturation

Connor O. McCoy[1], Trevor Bedford[2], Vladimir N. Minin[3], Philip Bradley[1], Harlan Robins[1] and Frederick A. Matsen IV[1]

[1]Computational Biology, and [2]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[3]Departments of Statistics and Biology, University of Washington, Seattle, WA, USA

TB, 0000-0002-4039-5794; PB, 0000-0002-0224-6464; HR, 0000-0001-5225-7076; FAM, 0000-0003-0607-6025

The antibody repertoire of each individual is continuously updated by the evolutionary process of B-cell receptor (BCR) 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 dataset of BCRs. We find that the substitution process is conserved across individuals but varies significantly across gene segments. We investigate selection on BCRs 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 provides a more nuanced view of the constraints on framework and variable regions.

1. 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 divide, whereas 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.

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 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 or loss of mutations and substitution, the observed process of evolution that follows from the first two processes. One major vein of research has focused on how nucleotide mutation rates depend on the identity of surrounding nucleotides (reviewed in [10]; see also [11,12]), 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 owing 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 [13]. 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 analyse selection on sequence tracts in this context (reviewed in §3b), but no methods have yet achieved the goal of statistical per-residue selection estimates. This has, however, been recently identified as an important goal [11]. Such selection estimates could be used to better direct generation of synthetic libraries of antibodies for high-throughput screening. Another application would be to the engineering of antibody Fc regions with specific properties, such as for bispecific monoclonal antibodies or antibody-derived fragments, while preserving overall stability.

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 [14].

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 [15]. 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. As SHM rarely introduces insertions or deletions (we observe whole codon insertion deletion events in between 0.013 and 0.014% of memory sequences within templated segments), it is appropriate to assume that observed frameshifts 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 dataset of BCRs. Specifically, we first apply model selection criteria to identify patterns in the single-nucleotide substitution process that occurs during affinity maturation and find that they are similar across individuals but vary significantly across gene segments. Next, we investigate how substitution processes vary between V genes and find that the primary source of variation is whether a sequence produces a functional receptor. Finally, we develop the first statistical methodology and corresponding software for comprehensive per-residue selection estimates for BCRs. 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 [16] 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 BCR sequences; we find that selection is dominated by negative selection with patterns that are consistent among individuals in our study.

2. Results

(a) Substitution model inference and testing

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 (table 1). For the underlying nucleotide substitution model, we fit a general time-reversible (GTR) nucleotide model [17] 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 and observed sequences (electronic supplementary material, table S3). We modelled substitution rate heterogeneity across sites using a four-category discretized Gamma distribution [18] with fixed mean 1.0.

We find that the best performing model (denoted $t_rQ_i\Gamma_i$, table 2) is one in which the branch length separating a sequenced BCR from its germline counterpart is estimated independently for each observed sequence, 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) [19] (table 2) and the Bayesian information criterion [20]

Table 1. Models and model testing results. The models of molecular evolution evaluated, including the number of free parameters introduced in parentheses.

name	branch length	GTR transition matrix	across-site rate variation (discrete Gamma)	total parameters
$t_iQ_i\Gamma_i$	one branch length per *segment* per sequence ($n \times 3$)	one matrix per segment (8×3)	one distribution per segment (3)	$3n + 27$
$T_rQ_i\Gamma_i$	one branch length per sequence (n) + relative rate between segments (2)	one matrix per segment (8×3)	one distribution per segment (3)	$n + 29$
$t_rQ_i\Gamma_s$	one branch length per sequence (n) + relative rate between segments (2)	one matrix per segment (8×3)	one shared distribution (1)	$n + 27$
$t_rQ_s\Gamma_s$	one branch length per sequence (n) + relative rate between segments (2)	one shared matrix (8)	one shared distribution (1)	$n + 11$

Table 2. Models and model testing results. Models show identical ranking across individuals. Columns include the log-likelihood (LogL), number of degrees of freedom (d.f.), Akaike information criterion (AIC) and difference of AIC from the top model (ΔIC).

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

(data not shown) identified the same rank order of support; this ordering was also identical for each of the three individuals. Other than the $t_iQ_i\Gamma_i$ 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 probably aided by the large volume of sequence data available.

Next, we fit the best-scoring model ($t_rQ_i\Gamma_i$) to the full dataset 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 (figure 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 V, D and J segments for each individual was very similar. We note that the sorting procedure used to separate memory from naive B cells provided memory cells at approximately 97% purity, so these divergence estimates may be conservative because of low levels of contamination from the naive 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 (electronic supplementary material, figures S1 and S2). 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 [21].

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 $t_rQ_i\Gamma_i$ 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 analysed with a variant of compositional principal components analysis [22] (see §4 Material and methods). We find that substitution models are influenced by in-frame versus out-of-frame sequence status, find no evidence for models clustering by individual, and see some limited evidence for clustering by gene subgroup (figure 2). The Euclidean distance between these transformed discrete probability distributions and the Hamming distance between germline V genes showed significant, but moderate, correlation (Spearman's $\rho = 0.20$, $p < 10^{-15}$; electronic supplementary material, figure S3).

(b) Natural selection

The primary challenge for BCR selection inference is that nucleotide context is known to have a very strong impact on mutation rates (reviewed in [21]). These context-specific mutations combined with the structure of the genetic code can result in extreme $\mathrm{d}N/\mathrm{d}S$ ratios using the classical definition that are not attributable to selection. To address this problem, we infer the selection coefficient ω using a non-synonymous–synonymous ratio which controls for background mutation rate via out-of-frame sequences (3). We continue the tradition of calling the selection coefficient ω in this context, even though it is a slightly different definition than previously used.

We apply this method to our dataset results in the first per-site and per-gene maps quantifying selection in the B-cell

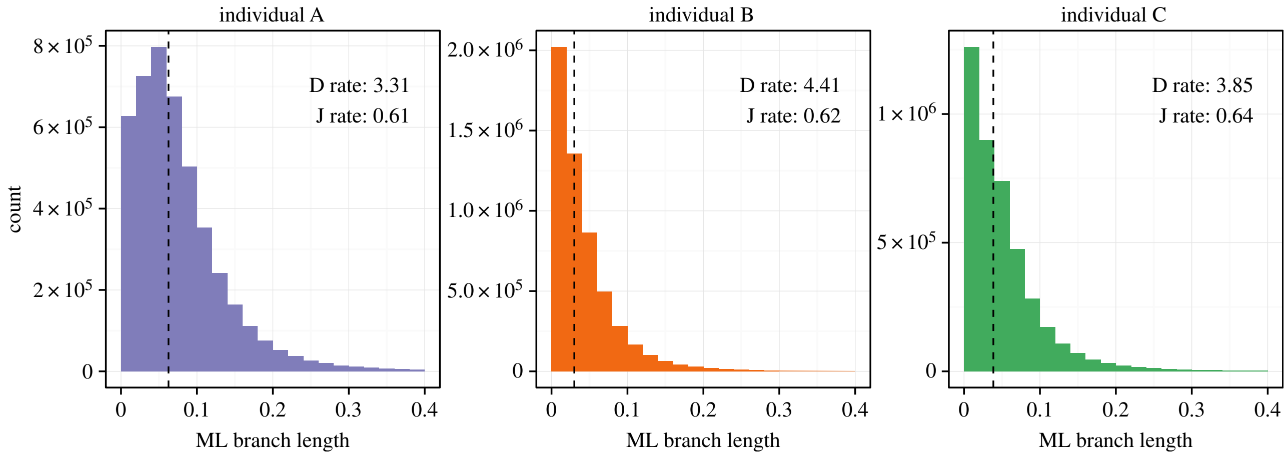

Figure 1. Distribution of maximum-likelihood branch lengths estimated under the $t_r Q_i \Gamma_i$ model. Branch lengths are measured in terms of substitutions per site, and rates given for the D and J segments are relative to a fixed rate of unity for the V segment.

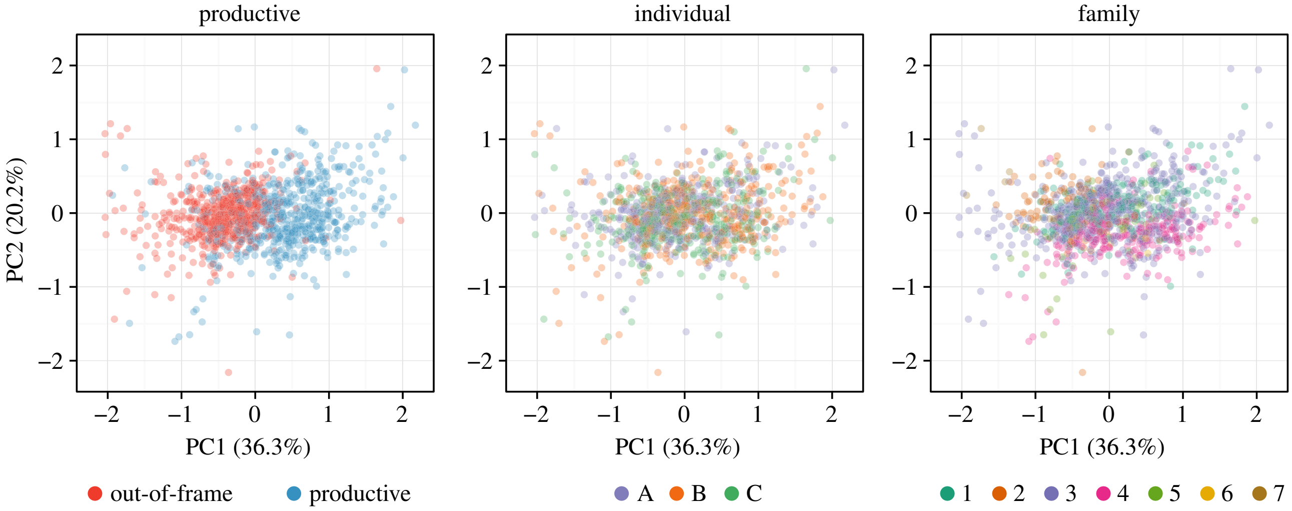

Figure 2. First (*x*-axis) and second (*y*-axis) principal components from PCA performed on centred log-transformed median-time transition matrices for V gene segments. Points plotted in a random order, with 22 outliers removed for clarity.

repertoire [23,24]. Sites were classified as negatively or positively selected based on whether the 95% Bayesian credible interval (BCI) excludes 1.0: sites for which the lower endpoint of the ω BCI is greater than 1.0 are classified as being under positive selection, whereas sites for which the upper endpoint of their ω BCI is less than 1.0 are classified as being under negative selection. We employ site numbering according to the IMGT unique numbering for the V domain [25].

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, figure 3*a*) even in out-of-frame sequences, which is presumably because of motif-driven mutation. Thus, if we had directly applied traditional means of estimating selection by comparing the rate of non-synonymous and synonymous substitutions, we would have falsely identified sites as being under strong selection. By contrast, the selection inferences made using out-of-frame sequences stay much closer to neutral (figure 3*b*).

We note extensive negative selection in the residues immediately preceding the third heavy chain complementarity-determining region (CDR3; figure 4). 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 (electronic supplementary material, figure S4). 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. These aromatic positions likely play important structural roles in the antibody complex: site 102 is buried in the core of the heavy chain and makes extensive van der Waals interactions as well as a sidechain–backbone hydrogen bond, while site 103 forms part of the interface between the heavy and light chains (see further description of structural results below).

Overall, we see extensive selection in our sequenced region (figure 5). The mean ω estimate across sites with at least 100 productive and out-of-frame sequences 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 (figure 5, left column); 30.6% of sites were confidently classified as being under negative selection (figure 5, right column).

Because amino acids interior to the protein could be important for protein stability compared with exposed ones, we hypothesized that residues under negative selection would be more internal to the antibody protein than those

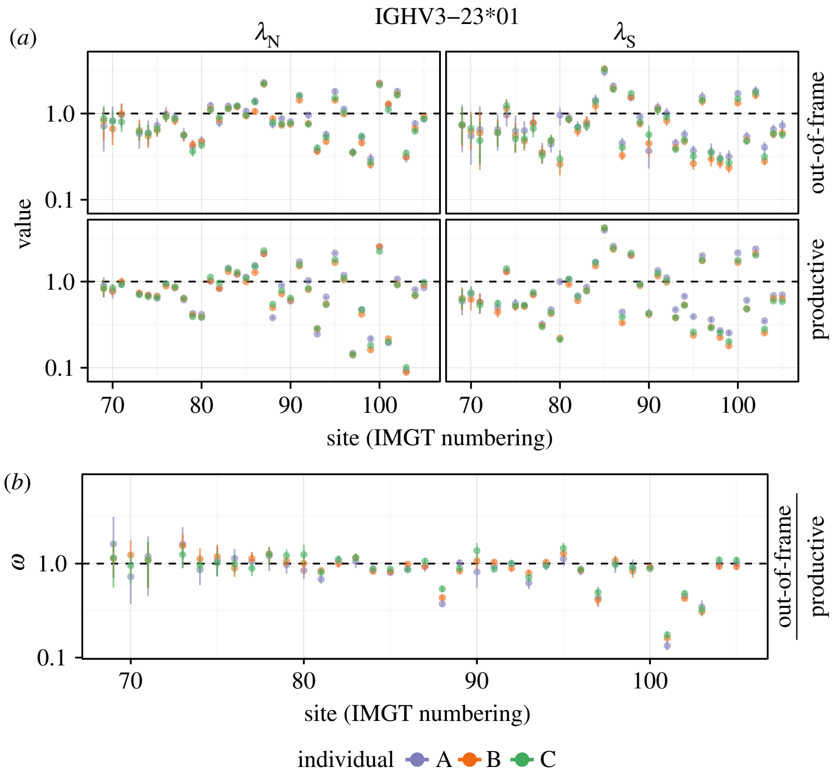

Figure 3. (*a*) Comparison of non-synonymous (λ_N) and synonymous (λ_S) rates in productive and out-of-frame sequences. (*b*) ω estimates using unproductive rearrangements as a proxy for the neutral process. Both panels use data from IGHV3-23*01, the most frequent V gene/allele combination.

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 (figure 6) and calculated the exposure of each amino acid position in the structure using the solvent-accessible surface area (SASA) using ROSETTA3 [27]. 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 (electronic supplementary material, figure S7). Differences in surface accessibility were significant between the three groups, with p-values ranging from less than 0.002 for the comparison of positive versus neutral sites to less than 10^{-15} for the comparison of negative versus neutral sites (Wilcoxon rank-sum test [28]).

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 (electronic supplementary material, figure S5), with between-individual coefficients of determination R^2 between 0.628 and 0.687 for site-specific ω values.

3. Discussion

We find different patterns of substitution across the V, D and J regions which is consistent among individuals (electronic supplementary material, figure S1) even though those individuals have differing levels of substitution (figure 1). We find that the dominant factor determining the V segment substitution process is whether 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 with common examples of Darwinian evolution, such as seen in *Drosophila* [29] and mammals [30], where most genes show ω less than 0.1. However, we note that although our estimates of ω are comparable with 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.

(a) Substitution process

We closely examine the substitution and selection processes in a context-independent manner, 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 (figure 2). Our model selection shows 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 (table 2). These between-segment differences are certainly due in part to base composition, which also differs significantly between segments and is similar between individuals (electronic supplementary material, table S3). Another contributing factor is probably similarity

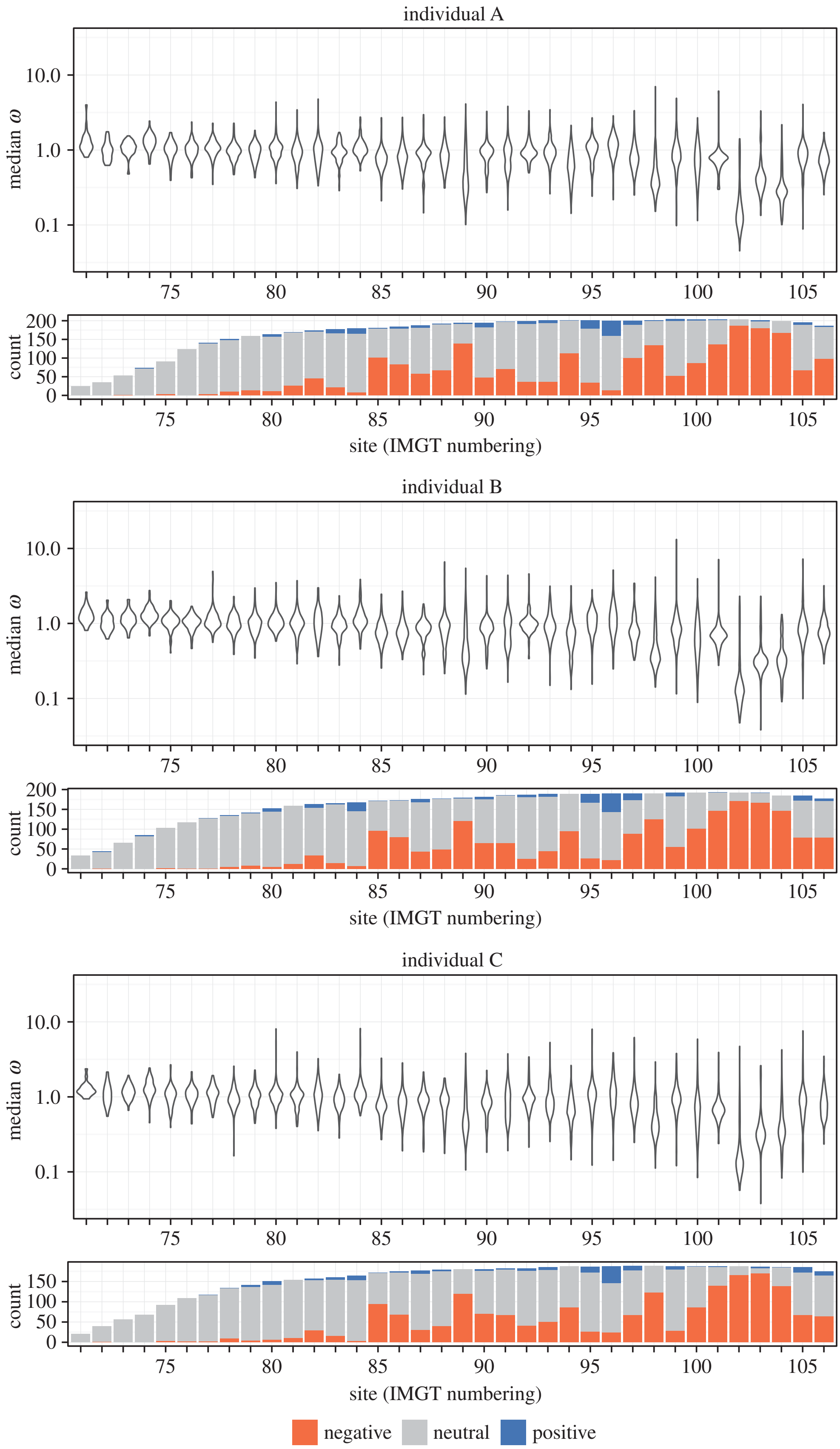

Figure 4. Site-specific estimates of the selection coefficient *ω*. Violin plots show distribution of median *ω* estimates across V genes at each site. Bar plots show count of V genes classified as undergoing negative, neutral, or positive selection. Only sites with at least 100 productive and out-of-frame observed sequences aligned were considered. The sites with IMGT numbers less than or equal to 104 are traditionally designated 'framework'.

of local nucleotide context between the genes of a given segment compared with between segments; these nucleotide contexts are known to impact AID-induced somatic hypermutation (reviewed in [21]). 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 analysing 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 (figure 2),

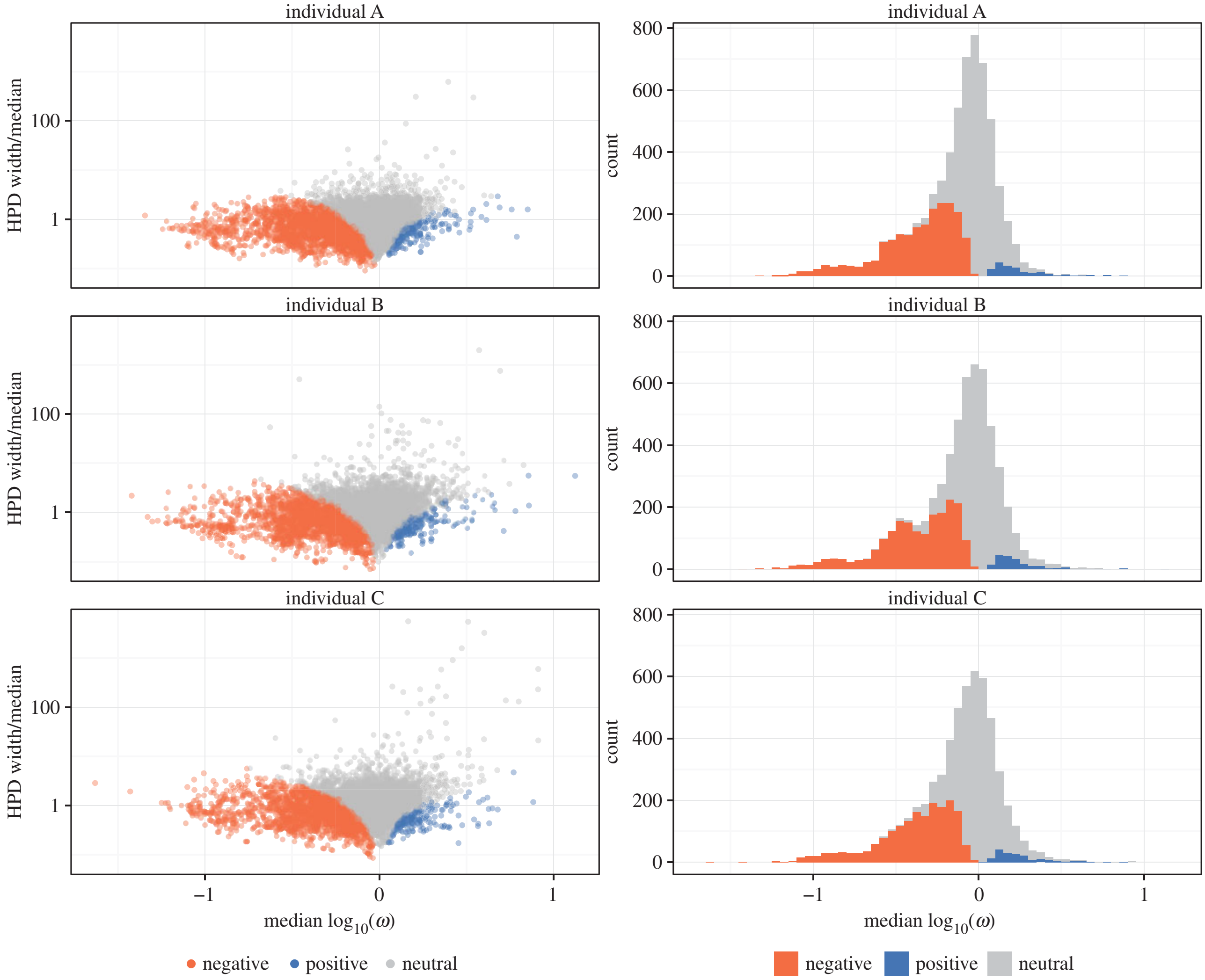

Figure 5. Site-specific selection estimates partitioned by individual and gene. Sites classified as negatively selected or positively selected based on whether the 95% BCI excludes unity and in what direction. (*a*,*c*,*e*) Comparison of ω estimate and relative width of BCI region and (*b*,*d*,*f*) distribution of site-specific selection estimates.

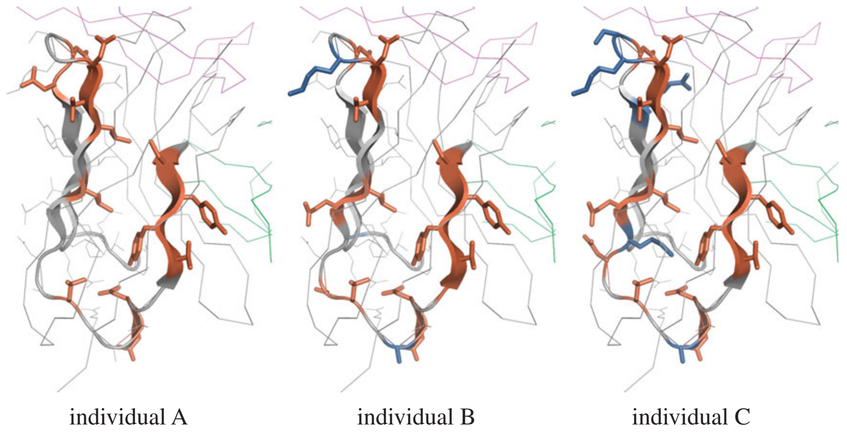

Figure 6. An IGHV3-23*01 (the most frequent V gene/allele combination) heavy chain antibody in complex with IL-17A (PDB ID 2VXS; [26]), with sites coloured by ω classification in each of the three individuals sampled. The bound antigen is shown in pink (top), and the light chain in green (right). The heavy chain structure is shown as a thin grey 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 β sheets) which is coloured grey at neutral sites, red at negatively selected sites and blue at positively selected sites.

which is presumably because of the impact of natural selection. We also find a significant correlation between sequence identity and substitution matrix (cf. [31]). In a related though distinct vein, Mirsky *et al.* [32] develop an amino acid substitution model for BCR sequences, which analogously aggregates information across positions.

(b) Selection process

The role of selection in BCR development has stimulated continuous interest since the pioneering 1985 paper of Clarke *et al.* [33]; 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 [34–38]. Recent work has offered methods that evaluate selection on a per-sequence basis [38]. There have also been efforts to infer selection based on lineage shape [39–44], which has been a common approach in macroevolutionary studies (reviewed in [45]) and more recently in population genetics [46–49].

In this work, we develop the first means of inferring per-residue selection using high-throughput sequence data with non-uniform coverage. Our method side-steps the difficulties encountered by previous work in differentiating between selection and motif-driven mutation in BCRs [11,13,34–38] by developing statistical means to compare in-frame and out-of-frame rearrangements. An alternative means of estimating selection was recently developed by Kepler *et al.* [12] in which a regression model incorporating a detailed model of motif preferences was used to infer selection coefficients for the framework and CDR regions in aggregate. In contrast to the previous work on B-cell selection, our methods provide a *per-residue* selection map for a contiguous stretch of BCR sequence.

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 observe a similarly high proportion of germline-identical sequences for in-frame and out-of-frame subsets in naive cells (electronic supplementary material, table S2); differences from germline derive in part from sequencing and other platform errors that do not depend on frame. For memory cells, we see extensive action of somatic hypermutation, but with a higher proportion of germline-identical out-of-frame sequences than in-frame (electronic supplementary material, table S2). We interpret these additional mutations for in-frame memory sequences as following from the process of affinity maturation for a specific antigen. 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 [21], 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 with the productive ones then this would simply make our estimates of selection conservative, pulling estimates of ω closer to unity, and yet our selection estimates are confidently classified as non-neutral for a substantial fraction of sites (figure 5).

In applying our methodology to IGHV sequences, we gain a high resolution per-gene map of selective forces on BCRs for part of the V gene. This part is primarily in the framework region, which is thought to be under substantial evolutionary constraint to preserve structure. Indeed, we see a pattern of quite strong negative selection in the region around the beginning of the CDR3 (figure 4), agreeing with recent work that found strong negative selection in one site near the beginning of the CDR3 [11]. However, other sites in this section of framework have substantially relaxed selection (figure 4). These results thus provide a more nuanced view into the constraints on BCR sequences rather than the traditional framework/variable designations, as also noted by Yaari *et al.* [11].

This work points the way towards future directions. In this work, we assumed that the size of individual lineages is small compared with the size of the overall repertoire, and thus that lineage structure could be ignored for the purpose of evolutionary model analysis. Ideally, we would reconstruct lineages and then do evolutionary analysis on the tree corresponding to each lineage. However, reconstructing groups of sequences forming a lineage is a challenging prospect on its own, to say nothing of doing phylogenetics on sequences in the presence of strong context-specific mutation-selection patterns, and have left out incorporating those aspects until we have first developed the necessary methods. We have recently developed an HMM framework to analyse VDJ rearrangements [50] and are currently developing and validating ways to use this framework for likelihood-based (as opposed to procedure-based [51,52]) lineage group inference.

4. Material and methods

(a) Dataset

The complete description of the experiment will be published separately [53]. 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. A measure of 400 ml 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 naive ($CD19^+D27^-IgD^+IgM^+$) and over 10 million memory ($CD19^+CD27^+$) B cells, with greater than 97% purity. Genomic DNA was extracted and the ImmunoSeq assay described in [3] was performed on the six samples at Adaptive Biotechnologies in Seattle, WA, USA.

The experiments and preprocessing were carefully designed to give an accurate quantification of error-corrected observed sequences. 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 [54]. To reduce sequencing errors and provide accurate quantification, each sample was divided among the wells on two 96-well plates and bar-coded individually. These templates were then amplified and 'over-sequenced' (electronic supplementary material, table S1), so that an average of almost six reads were present for each template. Following Robins *et al.* [55], reads matching the same template were collapsed into a consensus sequence with reduced sequencing error. Here, we grouped reads from within a well into consensus sequences by joining reads with Hamming distance less than or equal to two, and inferred the consensus sequence in each group using parsimony. Groups with only one member were discarded. This procedure protects against collapsing distinct sequences, as the probability that nearly identical distinct sequences co-occur exclusively in the same wells is small. We acknowledge this procedure may eliminate low frequency variants, but we prioritized accuracy over sensitivity towards these variants; despite this conservative analysis pipeline we observed substantial signal in the data.

Deep sequencing these B-cell populations resulted in 15 023 951 (electronic supplementary material, table S1) unique 130 bp observed sequences after preprocessing that spanned the third heavy chain complementarity determining region (CDR3) region (figure 7). The full dataset is available at http://adaptivebiotech.com/link/mat2015.

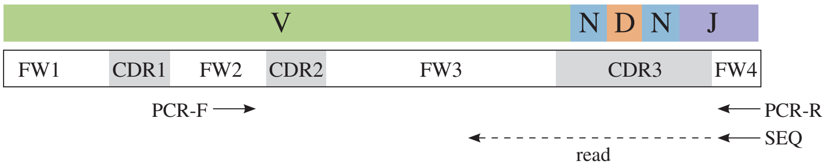

Figure 7. B-cell receptor schematic showing variable (V), diversity (D) and joining (J) gene segments as well as framework (FW) and complementarity-determining regions (CDRs). In VDJ recombination, individual V, D, and J gene segments are randomly selected and joined together via a process that deletes some randomly distributed number of nucleotides on their boundaries, then joins them together with random 'non-templated insertions' (N). 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.

(b) Alignment and germline assignment

Each sequence was first aligned to each V gene using the Smith–Waterman algorithm with an affine gap penalty [56]. 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 [14]. The best-scoring V, D and J alignment for each sequence was taken to be the germline alignment, and the corresponding germline sequence was taken to be the ancestral sequence for that observed sequence; in the case of ties, one germline sequence was chosen randomly among those alleles present at abundance greater than or equal to 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 naive populations of the individuals surveyed which were not represented in the IMGT database were added to the list of candidates for alignment. In contrast to naive sequences which have no mutations across almost all sites, the alleles we added to the germline collection were all present at greater than 30% for the IGHV gene in question.

(c) Substitution models, fitting and analysis

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 with common practice. For BCRs outside non-templated insertions, the root state is the V, D and J genes encoded in the germline from which a sequenced BCR derives. Thus, we analyse substitutions that have occurred in evolution away from the germline-encoded segments of observed BCR sequences, ignoring sites comprising non-templated insertions. The CDR3 region of an antibody is generally sufficient to uniquely identify its specificity [57]. Although there are certainly some clones in our dataset that derive from a single rearrangement event but differ due to somatic hypermutation, the probability that a given pair of distinct 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% [58]. It is a substantial challenge to statistically infer which sequences sit together in a clonal lineage and then to perform phylogenetic analysis on such a large dataset (see work by [12,59,60]) and performing this analysis incorrectly could bias our results. Additionally, we encountered significant computational barriers analysing the volume of sequences available, and adding phylogenetic structure to our analysis may have made the analysis computationally prohibitive even if we had the lineage structure in hand (we believe this is the largest number of sequences from a single dataset analysed in selection study to date).

For these reasons, our analyses were performed on a set of pairwise alignments, each representing a two taxon tree containing an observed sequence 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. This assumption allowed us to focus our attention on the selection inference problem.

Substitution models are summarized in table 1 and described in detail here. We will use n for the number of observed sequences. 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 (e.g. figure 1) with the V segment rate fixed at unity, 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 eight parameters. The subscript of Γ denotes how across-site substitution rate variation is modelled in terms of a four-category discrete Gamma distribution [18]. The Γ_i model allows an independent rates across-sites parameter for each sequence, with three parameters. The Γ_s has a global rates across-sites parameter, with one parameter. Given these choices concerning how the data were partitioned and parametrized, the standard phylogenetic likelihood function was used as described in the original literature [18,61,62] and in books (e.g. [63,64]).

Maximum-likelihood values of substitution model parameters and branch lengths were estimated using a combination of Bio++ [65] and BEAGLE [66], with model optimization via the BOBYQA algorithm [67] as implemented in NLopt [68], and branch length optimization via Brent's method [69]. 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 [22]. Specifically, each row of these matrices, as a discrete probability distribution, is a point in the simplex. Hence, we applied a centred log transformation to each row of this matrix using the `clr` function of the R package `compositions` [70], 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 [14]. To obtain distances between models, we calculated the Euclidean distance calculated between pairs of the transformed probability vectors.

(d) Selection analysis

Because of the large volume of sequences to analyse, we also needed a mechanism to detect selection that could be run on over 15 million sequences, most of which did not share evolutionary history. Classical means of estimating selection by codon model fitting [71,72] could not be used, even in their most recent and much more efficient incarnation [73]. Instead, we used the renaissance counting approach [16], which we modified to work under varying levels of coverage. A key part of the renaissance counting approach is an empirical Bayes regularization procedure [74]. 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 sequencing coverage. We note that obtaining precise per-site selection estimates for hundreds of genes requires a large quantity of sequence data like that which we have here: the read coverage decrease on the 5′ end of the V gene correspondingly increases the width of the error bar (figure 3, [23]), resulting in a decrease of power for selection regime classification (figure 4).

(i) Bayesian inference of selection on a star-shaped phylogeny

To determine the site-specific selection pressure for each V gene, we extended the counting renaissance method, described in [16], 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 [75] 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. As 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 mis-specified compared with a codon model and thus we decided to follow [16] 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 non-synonymous mutations.

(ii) Sampling codon substitutions

For each unique read, for each codon position l and posterior sample j, counts of synonymous ($C_{jl}^{(S)}$) and non-synonymous ($C_{jl}^{(N)}$) substitutions at each site were imputed using stochastic mapping as described above in §4d(i).

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 non-synonymous 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. This took about 5 days on an 194 core cluster launched on Amazon Web Services using `starcluster` (http://star.mit.edu/cluster/).

(iii) Empirical Bayes regularization

The varying length of the CDR3, combined with short observed sequences, leads to quite skewed coverage of sites stratified by gene. We modified the empirical Bayes regularization procedure of the original counting renaissance method [16] to account for varying depth of observation as follows. First, we define a branch length leading to query sequence i for site l as

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

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

$$\mathcal{L}(\alpha, \beta) = \left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right)^{L} \prod_{l} \frac{t_l^{C_l}}{\Gamma(C_l + 1)} \frac{\Gamma(C_l + \alpha)}{(t_l + \beta)^{C_l + \alpha}}. \tag{4.1}$$

We provide a derivation of this likelihood function in the electronic supplementary material. In contrast to [16], 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 [67].

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

$$\lambda_l | C_l \sim \text{Gamma}(C_l + \hat{\alpha}, t_l + \hat{\beta}), \tag{4.2}$$

derived in the electronic supplementary material.

Estimation of α and β by maximizing likelihood (4.1) fails when the sample variance of the observed counts $C_1 \ldots C_L$, weighted by the site-specific branch length sums, $t_1 \ldots t_L$, 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 λ is shared across sites and is estimated from the data by maximizing the likelihood

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

(iv) 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 [76]. 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 (electronic supplementary material, figure S6, bottom panel). Estimates of ω were more accurate with higher site coverage (electronic supplementary material, figure S6, 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. As 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 non-synonymous for codons encoding those amino acids.

(v) Site-specific estimates of ω

In [16], the authors arrive at site-specific estimates of ω_l by comparing data-conditioned (C) rates λ_l of non-synonymous (N) and synonymous (S) substitutions, each normalized by an 'unconditional rate' (U): $\omega_l^{\mathrm{RC}} = (\lambda_l^{(\mathrm{N,C})}/\lambda_l^{(\mathrm{N,U})})/(\lambda_l^{(\mathrm{S,C})}/\lambda_l^{(\mathrm{S,U})})$. 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:

$$\omega_l = \frac{\lambda_l^{(\mathrm{N,P})}/\lambda_l^{(\mathrm{N,O})}}{\lambda_l^{(\mathrm{S,P})}/\lambda_l^{(\mathrm{S,O})}}, \tag{4.3}$$

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

(vi) Implementation

We used the BEAST [77] implementation of the counting renaissance procedure to sample counts for both synonymous and non-synonymous substitutions at each site. We extended BEAST v. 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 in §4d(v). 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 [78] 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.

(e) Structural analysis

For each of the eleven most frequently occurring V genes, we identified the closest structure in the Protein Data Bank (PDB) [79] using BLAST [80]. Structures were visualized using PyMOL [81]. We calculated the normalized SASA for each amino acid position using ROSETTA3 [27] 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 [28] 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 [82] to account for multiple comparisons.

The details of our computational methods are available in the electronic supplementary material.

Data accessibility. Data are available at http://adaptivebiotech.com/link/mat2015. We have made an Amazon Machine Image (AMI) [83] 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).

Authors' contributions. C.M. designed and carried out the computational analysis, developed the statistical methods and wrote the paper; T.B. , V.M. and F.M. designed the computational analysis, developed the statistical methods and wrote the paper; P.B. performed structural analysis; H.R. helped define questions and provided the data.

Competing interests. H.R. owns stock in and consults for Adaptive Biotechnologies. Other authors have no competing interests.

Funding. This work was supported by the National Institute of Health [R01-AI107034 to V.N.M., R01-GM11324601 to F.A.M., R01-AI103981 to Dr Julie Overbaugh] and the National Science Foundation [DMS-0856099 to V.N.M]; C.O.M. and F.A.M. were supported in part by the University of Washington eScience Institute through its Seed Grants program in Translational Health Sciences, and by a 2013 new investigator award from the University of Washington Center for AIDS Research (CFAR), an NIH funded program under award no. P30AI027757 which is supported by the following NIH Institutes and Centers: NIAID, NCI, NIMH, NIDA, NICHD, NHLBI, NIA, NIGMS, and NIDDK. The molecular work for this project was supported by a grant from the W. M. Keck Foundation

Acknowledgements. The molecular work for this project was done by Paul Lindau in the laboratory of Phil Greenberg. The authors are grateful to Bryan Howie and William DeWitt III for helpful discussions.

References

1. Boyd SD *et al*. 2009 Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. *Sci. Transl. Med.* **1**, 12ra23. (doi:10.1126/scitranslmed.3000540)
2. Wu YC, Kipling D, Leong HS, Martin V, Ademokun AA, Dunn-Walters DK. 2010 High-throughput immunoglobulin repertoire analysis distinguishes between human IgM memory and switched memory B-cell populations. *Blood* **116**, 1070–1078. (doi:10.1182/blood-2010-03-275859)
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. (doi:10.4049/jimmunol.1201303)
4. DeKosky BJ *et al*. 2013 High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.* **31**, 166–169. (doi:10.1038/nbt.2492)
5. Robins H. 2013 Immunosequencing: applications of immune repertoire deep sequencing. *Curr. Opin. Immunol.* **25**, 646–652. (doi:10.1016/j.coi.2013.09.017)
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. (doi:10.1016/j.imlet.2012.08.002)
7. Six A *et al*. 2013 The past, present, and future of immune repertoire biology—the rise of next-generation repertoire analysis. *Front. Immunol.* **4**, 413. (doi:10.3389/fimmu.2013.00413)
8. Warren EH, Matsen FA IV, Chou J. 2013 High-throughput sequencing of B- and T-lymphocyte antigen receptors in hematology. *Blood* **122**, 19–22. (doi:10.1182/blood-2013-03-453142)
9. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. 2014 The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.* **32**, 158–168. (doi:10.1038/nbt.2782)
10. Delker RK, Fugmann SD, Papavasiliou FN. 2009 A coming-of-age story: activation-induced cytidine deaminase turns 10. *Nat. Immunol.* **10**, 1147–1153. (doi:10.1038/ni.1799)
11. Yaari G *et al*. 2013 Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front. Immunol.* **4**, 358. (doi:10.3389/fimmu.2013.00358)
12. Kepler TB *et al*. 2014 Reconstructing a B-cell clonal lineage. II. Mutation, selection, and affinity maturation. *Front. Immunol.* **5**, 170. (doi:10.3389/fimmu.2014.00170)
13. 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. (doi:10.1046/j.1365-2567.1998.00607.x)

14. Lefranc MP *et al*. 2008 IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res*. **37**, D1006–D1012. (doi:10.1093/nar/gkn838)
15. Corcoran AE. 2005 Immunoglobulin locus silencing and allelic exclusion. *Semin. Immunol*. **17**, 141–154. (doi:10.1016/j.smim.2005.01.002)
16. 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. (doi:10.1093/bioinformatics/bts580)
17. Lanave C, Preparata G, Saccone C, Serio G. 1984 A new method for calculating evolutionary substitution rates. *J. Mol. Evol*. **20**, 86–93. (doi:10.1007/BF02101990)
18. Yang Z. 1994 Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol*. **39**, 306–314. (doi:10.1007/BF00160154)
19. Akaike H. 1974 A new look at the statistical model identification. *IEEE Trans. Automat. Contr*. **19**, 716–723. (doi:10.1109/TAC.1974.1100705)
20. Schwarz G. 1978 Estimating the dimension of a model. *Ann. Stat*. **6**, 461–464. (doi:10.1214/aos/1176344136)
21. Teng G, Papavasiliou FN. 2007 Immunoglobulin somatic hypermutation. *Annu. Rev. Genet*. **41**, 107–120. (doi:10.1146/annurev.genet.41.110306.130340)
22. Aitchison J. 1983 Principal component analysis of compositional data. *Biometrika* **70**, 57–65. (doi:10.1093/biomet/70.1.57)
23. McCoy CO, Bedford T, Minin VN, Bradley P, Robins H, Matsen IV FA. 2015 Selection plots for all genes. *Figshare* 1399200. (doi:10.6084/m9.figshare.1399200)
24. McCoy CO, Bedford T, Minin VN, Bradley P, Robins H, Matsen IV FA. 2015 Per-site per-gene estimates of selection posterior densities. *Figshare* 1399201. (doi:10.6084/m9.figshare.1399201)
25. Lefranc MP, Pommié C, Ruiz M, Giudicelli V, Foulquier E, Truong L, Thouvenin-Contet V, Lefranc G. 2003 IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Dev. Comp. Immunol*. **27**, 55–77. (doi:10.1016/S0145-305X(02)00039-3)
26. Gerhardt S *et al*. 2009 Structure of IL-17A in complex with a potent, fully human neutralizing antibody. *J. Mol. Biol*. **394**, 905–921. (doi:10.1016/j.jmb.2009.10.008)
27. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R *et al*. 2011 ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol*. **487**, 545–574. (doi:10.1016/B978-0-12-381270-4.00019-6)
28. Hollander M, Wolfe DA. 1973 *Nonparametric statistical methods*. New York, NY: Wiley.
29. Clark AG *et al*. 2007 Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**, 203–218. (doi:10.1038/nature06341)
30. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker *et al*. 2011 A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482. (doi:10.1038/nature10530)
31. Kosakovsky Pond SL, Scheffler K, Gravenor MB, Poon AFY, Frost SDW. 2010 Evolutionary fingerprinting of genes. *Mol. Biol. Evol*. **27**, 520–536. (doi:10.1093/molbev/msp260)
32. Mirsky A, Kazandjian L, Anisimova M. 2014 Antibody-specific model of amino acid substitution for immunological inferences from alignments of antibody sequences. *Mol. Biol. Evol*. **32**, 806–819. (doi:10.1093/molbev/msu340)
33. Clarke SH, Huppi K, Ruezinsky D, Staudt L, Gerhard W, Weigert M. 1985 Inter- and intraclonal diversity in the antibody response to influenza hemagglutinin. *J. Exp. Med*. **161**, 687–704. (doi:10.1084/jem.161.4.687)
34. 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. (doi:10.1016/0167-5699(94)90175-9)
35. Lossos IS, Tibshirani R, Narasimhan B, Levy R. 2000 The inference of antigen selection on Ig genes. *J. Immunol*. **165**, 5122–5126. (doi:10.4049/jimmunol.165.9.5122)
36. 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. (doi:10.1111/j.1365-2567.2005.02208.x)
37. 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. (doi:10.1093/intimm/dxn026)
38. Yaari G, Uduman M, Kleinstein SH. 2012 Quantifying selection in high-throughput immunoglobulin sequencing data sets. *Nucleic Acids Res*. **40**, e134. (doi:10.1093/nar/gks457)
39. Steiman-Shimony A *et al*. 2006 Lineage tree analysis of immunoglobulin variable-region gene mutations in autoimmune diseases: chronic activation, normal selection. *Cell Immunol*. **244**, 130–136. (doi:10.1016/j.cellimm.2007.01.009)
40. Abraham RS *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. (doi:10.1007/s10875-006-9056-9)
41. Barak M, Zuckerman N, Edelman H, Unger R, Mehr R. 2008 IgTree (c): creating immunoglobulin variable region gene lineage trees. *J. Immunol. Methods* **338**, 67–74. (doi:10.1016/j.jim.2008.06.006)
42. Shahaf G, Barak M, Zuckerman NS, Swerdlin N, Gorfine M, 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**, 210–222. (doi:10.1016/j.jtbi.2008.08.005)
43. Liberman G, Benichou J, Tsaban L, Glanville J, Louzoun Y. 2013 Multi step selection in Ig H chains is initially focused on CDR3 and then on other CDR regions. *Front. Immunol*. **4**, 274. (doi:10.3389/fimmu.2013.00274)
44. 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. (doi:10.4049/jimmunol.1301551)
45. Mooers AO, Heard SB. 1997 Inferring evolutionary process from phylogenetic tree shape. *Q. Rev. Biol*. **72**, 31–54. (doi:10.2307/3036810)
46. Drummond AJ, Suchard MA. 2008 Fully Bayesian tests of neutrality using genealogical summary statistics. *BMC Genet*. **9**, 68. (doi:10.1186/1471-2156-9-68)
47. Li H, Wiehe T. 2013 Coalescent tree imbalance and a simple test for selective sweeps based on microsatellite variation. *PLoS Comput. Biol*. **9**, e1003060. (doi:10.1371/journal.pcbi.1003060)
48. Łuksza M, Lässig M. 2014 A predictive fitness model for influenza. *Nature* **507**, 57–61. (doi:10.1038/nature13087)
49. Neher RA, Russell CA, Shraiman BI. 2014 Predicting evolution from the shape of genealogical trees. *Elife* **3**, e03568. (doi:10.7554/eLife.03568)
50. Ralph DK, Matsen IV FA. In press. Consistency of VDJ rearrangement and substitution parameters enables accurate B cell receptor sequence annotation. *PLoS Comput. Biol.*
51. Liao HX *et al*. 2013 Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* **496**, 469–476. (doi:10.1038/nature12053)
52. Bashford-Rogers RJM, Palser AL, Huntly BJ, Rance R, Vassiliou GS, Follows GA, Kellam P. 2013 Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res*. **23**, 1874–1884. (doi:10.1101/gr.154815.113)
53. De Witt WS *et al*. In preparation. A public immunosequencing database of memory and naïve B cell receptors.
54. Carlson CS *et al*. 2013 Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun*. **4**, 2680. (doi:10.1038/ncomms3680)
55. Robins HS *et al*. 2009 Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood* **114**, 4099–4107. (doi:10.1182/blood-2009-04-217604)
56. Gotoh O. 1982 An improved algorithm for matching biological sequences. *J. Mol. Biol*. **162**, 705–708. (doi:10.1016/0022-2836(82)90398-9)
57. Xu JL, Davis MM. 2000 Diversity in the CDR3 region of VH is sufficient for most antibody specificities. *Immunity* **13**, 37–45. (doi:10.1016/S1074-7613(00)00006-6)
58. Zhu J *et al*. 2013 Mining the antibodyome for HIV-1-neutralizing antibodies with next-generation sequencing and phylogenetic pairing of heavy/light chains. *Proc. Natl Acad. Sci. USA* **110**, 6470–6475. (doi:10.1073/pnas.1219320110)

59. Jiang N *et al*. 2013 Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci. Transl. Med.* **5**, 171ra19. (doi:10.1126/scitranslmed.3004794)

60. Kepler TB. 2013 Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors. *F1000Res* **2**, 103. (doi:10.12688/f1000research.2-103.v1)

61. Felsenstein J. 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376. (doi:10.1007/BF01734359)

62. Tavaré S. 1986 Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* **17**, 57–86.

63. Salemi M, Lemey P, Vandamme AM. 2009 *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge, UK: Cambridge University Press.

64. Felsenstein J. 2004 *Inferring phylogenies*. Sunderland, MA: Sinauer Associates.

65. Guéguen L *et al*. 2013 Bio++: efficient extensible libraries and tools for computational molecular evolution. *Mol. Biol. Evol.* **30**, 1745–1750. (doi:10.1093/molbev/mst097)

66. Ayres DL *et al*. 2011 BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst. Biol.* **61**, 170–173. (doi:10.1093/sysbio/syr100)

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

68. Johnson SG. 2010 The NLopt nonlinear-optimization package. See http://ab-initio.mit.edu/nlopt.

69. Brent RP. 1973 *Algorithms for minimization without derivatives*. Englewood Cliffs, NJ: Prentice Hall.

70. van den Boogaart KG, Tolosana-Delgado R. 2008 'Compositions': a unified R package to analyze compositional data. *Comput. Geosci.* **34**, 320–338. (doi:10.1016/j.cageo.2006.11.017)

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

72. 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.

73. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K. 2013 FUBAR: a fast, unconstrained Bayesian approximation for inferring selection. *Mol. Biol. Evol.* **30**, 1196–1205. (doi:10.1093/molbev/mst030)

74. Robbins H. 1956 An empirical Bayes approach to statistics. In *Proc. Third Berkeley Symp. on Mathematical Statistics and Probability, vol. 1: Contributions to the Theory of Statistics, Berkeley, CA, December 1954 and July–August 1955*. Oakland, CA: University of California Press. See https://projecteuclid.org/euclid.bsmsp/1200501640.

75. 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. (doi:10.1007/BF02101694)

76. Bielejec F, Lemey P, Carvalho LM, Baele G, Rambaut A, Suchard MA. 2014 piBUSS: a parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios. *BMC Bioinformatics* **15,** 133. (doi:10.1186/1471-2105-15-133)

77. Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012 Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* **29**, 1969–1973. (doi:10.1093/molbev/mss075)

78. Zaharia M, Chowdhury M, Franklin M, Shenker S, Stoica I. 2010 Spark: cluster computing with working sets. In *2nd USENIX Conf. on Hot Topics in Cloud Computing* (eds E Nahum, D Xu), p. 10. See http://static.usenix.org/legacy/events/hotcloud10/tech/full_papers/Zaharia.pdf.

79. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bournel PE. 2000 The protein data bank. *Nucleic Acids Res.* **28**, 235–242. (doi:10.1093/nar/28.1.235)

80. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402. (doi:10.1093/nar/25.17.3389)

81. Delano WL. 2002 *The PyMOL molecular graphics system*. Palo Alto, CA: DeLano Scientific. See http://www.pymol.org.

82. Holm S. 1979 A simple sequentially rejective multiple test procedure. *Scand. Stat. Theory Appl.* **6**, 65–70.

83. Amazon Web Services 2014 *Amazon EC2 FAQs*. See http://aws.amazon.com/ec2/faqs/#What_is_Amazon_Elastic_Compute_Cloud_Amazon_EC2.