Selection and neutral mutations drive pervasive mutability losses in long-lived B cell lineages

Marcos C. Vieira*, Daniel Zinder & Sarah Cobey,

Department of Ecology and Evolution, University of Chicago
* Corresponding author: mvieira@uchicago.edu

Supplementary information

Simulation of B cell receptor alignments

To assess the power of a random local clock model [1] combined with robust counting [2] to detect changes in synonymous and non-synonymous substitution rates during B cell evolution, we analyzed B cell alignments simulated under different underlying mutation rates. In each simulation, the mutation rate was either constant or variable over time, and either constant or variable across sites. In simulations with variable mutation rates across sites, site-specific rates either depended on the site’s S5F mutability [3] or on whether the site was a WRCH/DGYW hotspot [4]. In simulations with variable mutation rates across sites, mutation rates also changed over time as the S5F mutability of the sequences (or the number of WRCH/DGYW hotspots in them) changed. Simulations are summarized below.

**Scenario 1**: Mutation rate constant across time and across sites.

**Scenario 2**: Mutation rate constant across sites but decreasing over time.

2a: 10% decrease in rate over simulation period.
2b: 20% decrease in rate.
2c: 30% decrease in rate.
2d: 40% decrease in rate.
2e: 50% decrease in rate.

**Scenario 3**: Variable mutation rates over time and across sites.

3a: Site-specific mutation rates depend on S5F mutability.
3b: WRCH/DGYW hotspots are three times more likely to mutate.
3c: WRCH/DGYW hotspots are 30 times more likely to mutate.

For the simulations, we modified a simple forward-time Wright-Fisher model to impose a fitness cost on non-synonymous mutations. The status of a mutation (synonymous or non-synonymous) is defined relative to a fixed reference sequence. As a reference sequence, we used the heavy
chain sequence at the root node of lineage CH103, inferred under a logistic growth prior (98-100% identical to sequences inferred in four replicate chains under a constant population size prior). An initial population is generated consisting of a single copy of the reference sequence. At each subsequent generation \( t \), \( N_t \) sequences are produced by replicating sequences from the previous generation with the possibility of mutation. We assumed \( N \) grows logistically:

\[
N_t = N_{t-1} + rN_{t-1}\left(1 - \frac{N_{t-1}}{K}\right)
\]  

(1)

Logistic growth assumes the B cell population initially expands exponentially with intrinsic growth rate \( r \) and then saturates at the carrying capacity \( K \).

At each generation \( t \), we generate a number of new sequences equal to the nearest integer to \( N_t \). The probability that a newly generated sequence at generation \( t \) descends from sequence \( i \) in generation \( t-1 \) is equal to the fitness of sequence \( i \), \( w_i \), normalized by the sum of fitness values across the entire population at \( t-1 \). Any sequence \( i \) whose amino acid translation is the same as that of the reference sequence has fitness \( w_i = 1 \). Each non-synonymous mutation relative to the reference sequence adds \( s \) to the value of \( w_i \), with negative values of \( s \) representing a fitness cost (\( w_i \) cannot go below 0). Newly generated sequences undergo mutations at fixed or variable rates, depending on the scenario.

**Modeling relative and absolute mutation rates**

Different models have been proposed to describe variation in somatic hypermutation rates across different nucleotide motifs [3–5]. For site \( i \) in sequence \( j \), those models can be used to assign a relative mutation rate \( m_{i,j} \) to site \( i \), based on its local sequence context in sequence \( j \). However, the precise relationship between the relative mutability of a site and the site’s absolute mutation rate is unclear. To model absolute mutation rates as a function of the relative mutability of a sequence’s motifs, we assume that the average mutation rate per site per generation for sequence \( j \), \( \bar{\mu}_j \), is proportional to the sequence’s average relative mutability, \( \bar{m}_j \):

\[
\bar{\mu}_j = k \times \bar{m}_j
\]  

(2)

Let \( \bar{m}_0 \) be the average relative mutability of the reference sequence. We choose \( k \) so that the reference sequence has an average mutation rate per site per generation, \( \bar{\mu}_0 \), equal to \( 1/4L \), where \( L \) is the sequence length (in number of nucleotides):

\[
k = \frac{1}{\bar{m}_04L}
\]  

(3)

In preliminary simulations under default values for other parameters (see below), this choice of \( \bar{\mu}_0 = 1/4L \) produced alignments visually similar to
those observed for real B cell lineages in terms of overall nucleotide diversity. For any sequence $j$, the average mutation rate per site per generation is then given by:

$$\mu_j = \frac{m_j}{m_04L}$$  \hspace{1cm} (4)

Thus, a sequence with half the average relative mutability of the reference sequence has half the average mutation rate per site per generation. Finally, site-specific mutation rates per generation for sequence $j$ are given by:

$$\mu_{i,j} = m_{i,j} \times k = \frac{m_{i,j}}{m_04L}$$  \hspace{1cm} (5)

Note that scaling site-specific mutation rates to $1/4L$ does not change the relative mutability of the sites. For example, for sites $a$ and $b$ in the same sequence $j$:

$$\frac{\mu_{a,j}}{\mu_{b,j}} = \frac{m_{a,j}}{m_{b,j}}$$

To models scenarios 1 and 2 (where mutation rates are independent of sequence context) we let $\mu_{i,j}(t) = c(t) \times 1/4L$ be the mutation rate per time per generation for all sites in all sequences at generation $t$, where $0 \leq c(t) \leq 1$. In scenario 1 we let $c(t) = 1$ for all $t$, whereas in scenario 2 we choose decreasing values of $c(t)$ for different time intervals.

In the S5F-based parameterization (scenario 3a), we set $m_{i,j}$ to the relative mutability scores from [3], based on a five-nucleotide window centered on site $i$. The first two sites and the last two sites, for which S5F is indeterminate since the neighbors are unknown, are assigned mutability zero. We used motif-specific transition probabilities between nucleotides inferred by the S5F model along with motif-specific mutation rates.

In the “hotspot” parameterization (scenarios 3b-c), we let $m_{i,j}$ be either 1, if site $i$ is not at the central position of a hotspot, or $h$, if it is. Site $i$ is at the central position of a hotspot if it is occupied by the underlined nucleotide in a $W\overline{RCH}$ or a $DGY\overline{W}$ motif, where $W = \{A/T\}$, $R = \{A/G\}$, $H = \{A/T/C\}$, $D = \{A/T/G\}$ and $Y = \{C/T\}$ [4]. A site is assigned mutability 1 if it cannot be determined whether or not that site is at the center of a $W\overline{RCH}$ or a $DGY\overline{W}$ motif (for example, the left neighbors of the first site in a sequence and the right neighbors of the last site in a sequence are unknown). We assumed uniform transition probabilities between nucleotides.

We ran simulations for 2000 generations, sampling 25 sequences every 250 generations starting at generation 500. We decreased $c(t)$ by 0.2 every 250 generations starting at generation $t = 1000$ for scenario 2a, by 0.4 every 250 generations starting at generation $t = 1000$ for scenario 2b, by 0.5 every 250 generations starting at generation $t = 750$ for scenario 2c, by 0.8 every
250 generations starting at generation $t = 1000$ for scenario 2d, and by 0.1 every 250 generations starting at generation $t = 1000$ for scenario 2d.

We ran all simulations with $s = 0.01$, $r = 0.7$ and $K = 1000$. For scenario 2, we repeated the simulations with $s = 0.005$, $r = 0.01$, $K = 2000$. Under the first set of parameters, the simulated population reached the carrying capacity after approximately 20 generations, before the start of sampling at generation 500. Under the second set of parameters, carrying capacity was reached after approximately 1500 generations, well into the sampling period.

We analyzed the simulated alignments using BEAST v.1.8.2 to test if declines in synonymous and non-synonymous substitution rates are correctly detected in scenarios 2 and 3.

**Changes in substitution rates over time**

For each tree in the posterior distributions inferred for observed and simulated alignments, we computed the estimated synonymous and non-synonymous substitution rates for each branch by dividing estimated counts of synonymous and non-synonymous substitutions (obtained by robust counting) by the branch’s length measured in time. For each tree in the posterior distributions of simulated alignments, we estimated pairwise linear regression coefficients between branch times (predictor variable, measured at each branch’s parent node) and total, synonymous and non-synonymous substitution rates (response variables).
Supplementary figures and tables

**Fig S1.** Mutability of the inferred ancestral sequences of long-lived B cell lineages (red squares) compared with the distribution of mutability values obtained by randomizing the ancestral codon sequence while keeping the amino acid sequence constant. Mutability was measured as the geometric mean of the S5F scores across sites. Results are shown for the whole sequences (WS) and separately for framework regions (FRs) and complementarity determining regions (CDRs).
Fig S2. Evolution of mutability in long-lived B cell lineages. Mutability was measured as the geometric mean of S5F scores across all sites in the sequence. Scatterplots show mutability over time for nodes from a sample of 100 trees from the posterior distribution. Blue points correspond to terminal nodes (observed sequences), and black points correspond to internal nodes whose sequences were inferred statistically. The solid red line represents an average of regression lines calculated for each tree in a sample of 1000 trees, with the 95% highest-posterior density interval for the slope of regression annotated on top of each panel. The dashed line is the regression line obtained from observed sequences alone, excluding internal nodes.
**Fig S3.** Evolution of mutability in the framework regions (FRs) of long-lived B cell lineages. Mutability was measured as the geometric mean of S5F scores across all sites in FRs. Scatterplots show mutability over time for nodes from a sample of 100 trees from the posterior distribution. Blue points correspond to terminal nodes (observed sequences), and black points correspond to internal nodes whose sequences were inferred statistically. The solid red line represents an average of regression lines calculated for each tree in a sample of 1000 trees, with the 95% highest-posterior density interval for the slope of regression annotated on top of each panel. The dashed line is the regression line obtained from observed sequences alone, excluding internal nodes.
Fig S4. Evolution of mutability in the complementarity determining regions (CDRs) of long-lived B cell lineages. Mutability was measured as the geometric mean of S5F scores across all sites in CDRs. Scatterplots show mutability over time for nodes from a sample of 100 trees from the posterior distribution. Blue points correspond to terminal nodes (observed sequences), and black points correspond to internal nodes whose sequences were inferred statistically. The solid red line represents an average of regression lines calculated for each tree in a sample of 1000 trees, with the 95% highest-posterior density interval for the slope of regression annotated on top of each panel. The dashed line is the regression line obtained from observed sequences alone, excluding internal nodes.
Fig S5. Evolution of the difference in mutability between complementarity determining regions (CDRs) and framework regions (FRs) in long-lived B cell lineages. The relative difference is calculated as the average log-S5F mutability of CDRs minus the average log-S5F mutability of FRs. Each point corresponds to a node in the maximum-clade-credibility tree of each lineage. Differences in CDR and FR mutability are plotted as a function of genetic distance from the root of the tree, measured as the expected number of substitutions per site since the root.
Fig S6. Mutability of B cell receptor sequences from different B cell lineages relative to the expected distribution of mutability values obtained by randomizing codon sequences while keeping the amino sequences constant. Mutability was measured as the geometric mean of S5F scores across sites. The distribution of mutability percentiles obtained for sequences sampled at the last sampling time point in each dataset is shown in gray. The mutability percentile of each lineage’s ancestor is shown in red. Results are shown separately for framework regions (FR) and complementarity determining regions (CDRs).
Fig S7. Changes in mean log-S5F mutability due to non-synonymous changes averaged across all branches of different anti-HIV B cell lineages. Gray points indicate values inferred from the data, and colored points indicate values obtained by simulation under different models. Red indicates an S5F-based model where different nucleotide motifs mutate with different rates, dark blue indicates a model with no mutation rate variation across sites, and light blue indicates a model with different mutation rates for each position of a codon. Simulations were performed independently for each branch on the MCC tree of different anti-HIV B cell lineages, starting from the inferred sequence of the parent node. Each simulated sequence was constrained to have the same number of non-synonymous and synonymous changes as observed in the branch. Vertical bars indicate the 95% range obtained from 100 simulations per model.
**Fig S8.** Frequency of amino acid transitions simulated under the S5F model compared to the frequencies in the MCC trees of B cell lineages inferred from the data. Amino acid transitions are colored according to their average effect on mean log-S5F mutability in 100 simulations.
Fig S9. Frequency of losses relative to the total number of changes in mean log-S5F mutability caused by synonymous substitutions during the evolution of anti-HIV B cell lineages. Blue indicates changes that on terminal branches, orange indicates changes along the trunk of the tree, and black indicates changes on the remaining internal branches. Results are shown separately for framework regions (FRs) and complementarity determining regions (CDRs). Each point denotes the frequency of changes in mutability that were losses, averaged across a sample of 1000 trees from the posterior distribution. Vertical lines indicate the 95% highest-posterior density interval.
Fig S10. Evolution of the number of WRCH/DGYW hotspots in long-lived B cell lineages. Scatterplots show the number of hotspots over time for nodes from a sample of 100 trees from the posterior distribution. Blue points correspond to terminal nodes (observed sequences), and black points correspond to internal nodes whose sequences were inferred statistically. The solid red line represents an average of regression lines calculated for each tree in a sample of 1000 trees, with the 95% highest-posterior density interval for the slope of regression annotated on top of each panel. The dashed line is the regression line obtained from observed sequences alone, excluding internal nodes.
Fig S11. Evolution of the number of overlapping hotspots in long-lived B cell lineages. Scatterplots show the number of overlapping hotspots over time for nodes from a sample of 100 trees from the posterior distribution. Blue points correspond to terminal nodes (observed sequences), and black points correspond to internal nodes whose sequences were inferred statistically. The solid red line represents an average of regression lines calculated for each tree in a sample of 1000 trees, with the 95% highest-posterior density interval for the slope of regression annotated on top of each panel. The dashed line is the regression line obtained from observed sequences alone, excluding internal nodes.
Fig S12. Evolution of 7-mer mutability in long-lived B cell lineages. Scatterplots show 7-mer mutability over time for nodes from a sample of 100 trees from the posterior distribution. Blue points correspond to terminal nodes (observed sequences), and black points correspond to internal nodes whose sequences were inferred statistically. The solid red line represents an average of regression lines calculated for each tree in a sample of 1000 trees, with the 95% highest-posterior density interval for the slope of regression annotated on top of each panel. The dashed line is the regression line obtained from observed sequences alone, excluding internal nodes.
Fig S13. Frequency of mutability losses relative to the total number of changes in mutability during the evolution of anti-HIV B cell lineages. Rows correspond to different mutability metrics, and column contain results obtained for the whole analyzed region of the BCR sequence, and separately for framework regions (FRs) and complementarity determining regions (CDRs). Each point denotes the frequency of changes in mutability that were losses, averaged across a sample of 1000 trees from the posterior distribution. Vertical red lines indicate the 95% highest-posterior density interval.
Fig S14. Total substitution rate inferred from the random local clock model, as a function of time for the observed lineages. Each plot shows the points corresponding to a sample of 100 trees from the posterior distribution inferred by BEAST. Terminal branches are shown in blue, and internal branches are shown in black.
Fig S15. Total substitution rate inferred from robust counting (summing synonymous and non-synonymous rate estimates), as a function of time for the observed lineages. Each plot shows the points corresponding to a sample of 100 trees from the posterior distribution inferred by BEAST. Terminal branches are shown in blue, and internal branches are shown in black.
Fig S16. Robust counting synonymous substitution rate as a function of time for the observed lineages. Each plot shows the points corresponding to a sample of 100 trees from the posterior distribution inferred by BEAST. Terminal branches are shown in blue, and internal branches are shown in black.
Fig S17. Robust counting non-synonymous substitution rate as a function of time for the observed lineages. Each plot shows the points corresponding to a sample of 100 trees from the posterior distribution inferred by BEAST. Terminal branches are shown in blue, and internal branches are shown in black.
Fig S18. Relationship between robust counting substitution rates and time for simulations performed under different levels of decline in the overall mutation rate. Each black line is the linear regression line between branch-specific rates and times for a single tree from the posterior distribution inferred for a simulated alignment using BEAST. Each plot shows a sample of 500 lines. The red lines are the “average” regression lines, with the average intercept and the average slope calculated from a larger sample of 1000 trees from each distribution. Parameter values for the simulations were fitness cost $s = -0.01$, intrinsic growth rate $r = 0.7$ and carrying-capacity $K = 1000$. 

![Graphs showing relationship between robust counting substitution rates and time for simulations performed under different levels of decline in the overall mutation rate.](image-url)
Fig S19. Relationship between robust counting substitution rates and time for simulations performed as in Fig. S18 but using a different set of parameters (fitness cost $s = -0.01$, intrinsic growth rate $r = 0.7$ and carrying-capacity $K = 1000$). Each black line is the linear regression line between branch-specific rates and times for a single tree from the posterior distribution inferred for a simulated alignment using BEAST. Each plot shows a sample of 500 lines. The red lines are the “average” regression lines, with the average intercept and the average slope calculated from a larger sample of 1000 trees from each distribution.
Fig S20. Relationship between robust counting substitution rates and time for simulations performed under models where the mutation rate at each site depends on its S5F mutability or on whether that site is at the center of a WRCH/DGYW hotspot (in which case it mutates either 3 or 30 times more frequently than non-hotspots sites). Each black line is the linear regression line between branch-specific rates and times for a single tree from the posterior distribution inferred for a simulated alignment using BEAST. Each plot shows a sample of 500 lines. The red lines are the “average” regression lines, with the average intercept and the average linear coefficient calculated from a larger sample of 1000 trees from each distribution.

Table S1. BEAST priors. Priors are left at the default choices expect where specified below.

| Parameter     | Description                          | Prior       | Initial value |
|---------------|--------------------------------------|-------------|---------------|
| CP1.mu        | Relative rate for codon position 1   | Γ [0.05, 10]| 1             |
| CP2.mu        | Relative rate for codon position 2   | Γ [0.05, 10]| 1             |
| CP3.mu        | Relative rate for codon position 3   | Γ [0.05, 10]| 1             |
| clock.rate    | Substitution rate                    | Uniform[0,10]| 0.0001        |
| logistic.popSize | Pop. size under logistic growth     | Default    | 100           |
| logistic.growRate | Logistic growth rate              | Default    | 0.5           |

References

1. Drummond AJ, Suchard MA. Bayesian random local clocks, or one rate to rule them al. BMC Biology. 2010;8(114). doi:10.1080/10635150590947041.
2. O’Brien JD, Minin VN, Suchard MA. Learning to count: robust estimates for labeled distances between molecular sequences. Molecular biology and evolution. 2009;26(4):801–14. doi:10.1093/molbev/msp003.

3. Yaari G, Vander Heiden JA, Uduman M, Gadala-Maria D, Gupta N, Stern JNH, et al. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. Frontiers in Immunology. 2013;4:358. doi:10.3389/fimmu.2013.00358.

4. Rogozin IB, Diaz M. Cutting Edge: DGYW/WRCH Is a Better Predictor of Mutability at G:C Bases in Ig Hypermutation Than the Widely Accepted RGYW/WRCY Motif and Probably Reflects a Two-Step Activation-Induced Cytidine Deaminase-Triggered Process. The Journal of Immunology. 2004;172(6):3382–3384. doi:10.4049/jimmunol.172.6.3382.

5. Elhanati Y, Sethna Z, Marcou Q, Callan CG, Mora T, Walczak AM. Inferring processes underlying B-cell repertoire diversity. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2015;370(1676):20140243. doi:10.1098/rstb.2014.0243.