Supporting Information for:

Quantifying selection in high-throughput Immunoglobulin sequencing datasets

Gur Yaari[1], Mohamed Uduman[2] and Steven H. Kleinstein[1,2]
gur.yaari@yale.edu, steven.kleinstein@yale.edu

[1] Department of Pathology, Yale University School of Medicine, New Haven, CT, USA
[2]Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA

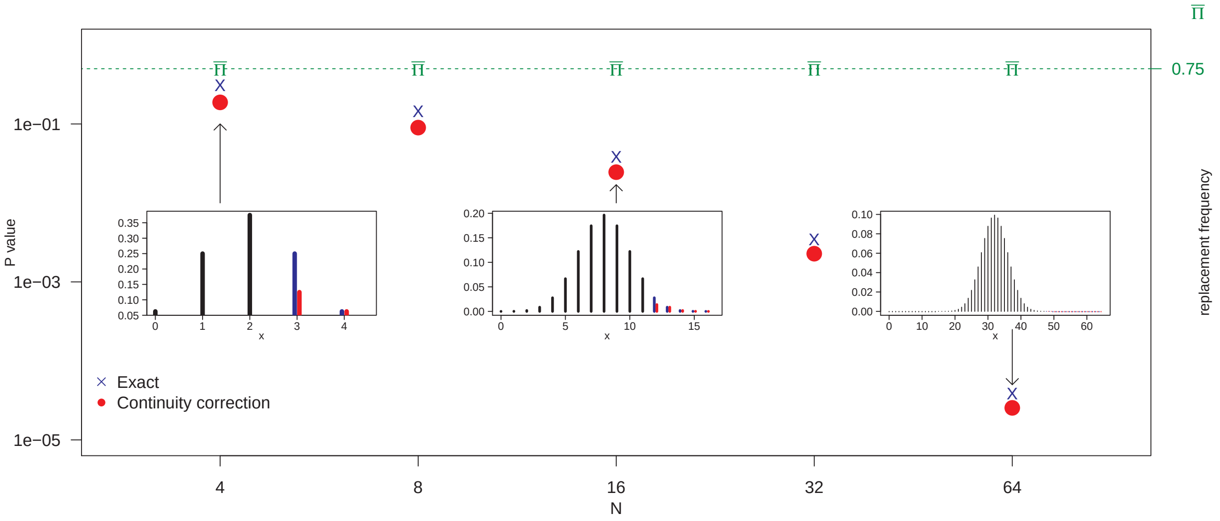

Figure S1: P values are not equivalent to selection strength. Increasing the number of mutations ($N = 4, 8, 16, 32, 64$) leads to decreasing P values from a Binomial test (points), even when the number of replacements (x) is set to maintain the same overall frequency ($\pi = 0.75$). P values were calculated either through an "exact" method (blue X's and bars in the subplots) or applying a continuity correction (red circles and bars in the subplots). In contrast, the maximum likelihood value for the probability of replacement mutations ($\bar{\Pi}$) remains the same (green symbols, right axis).

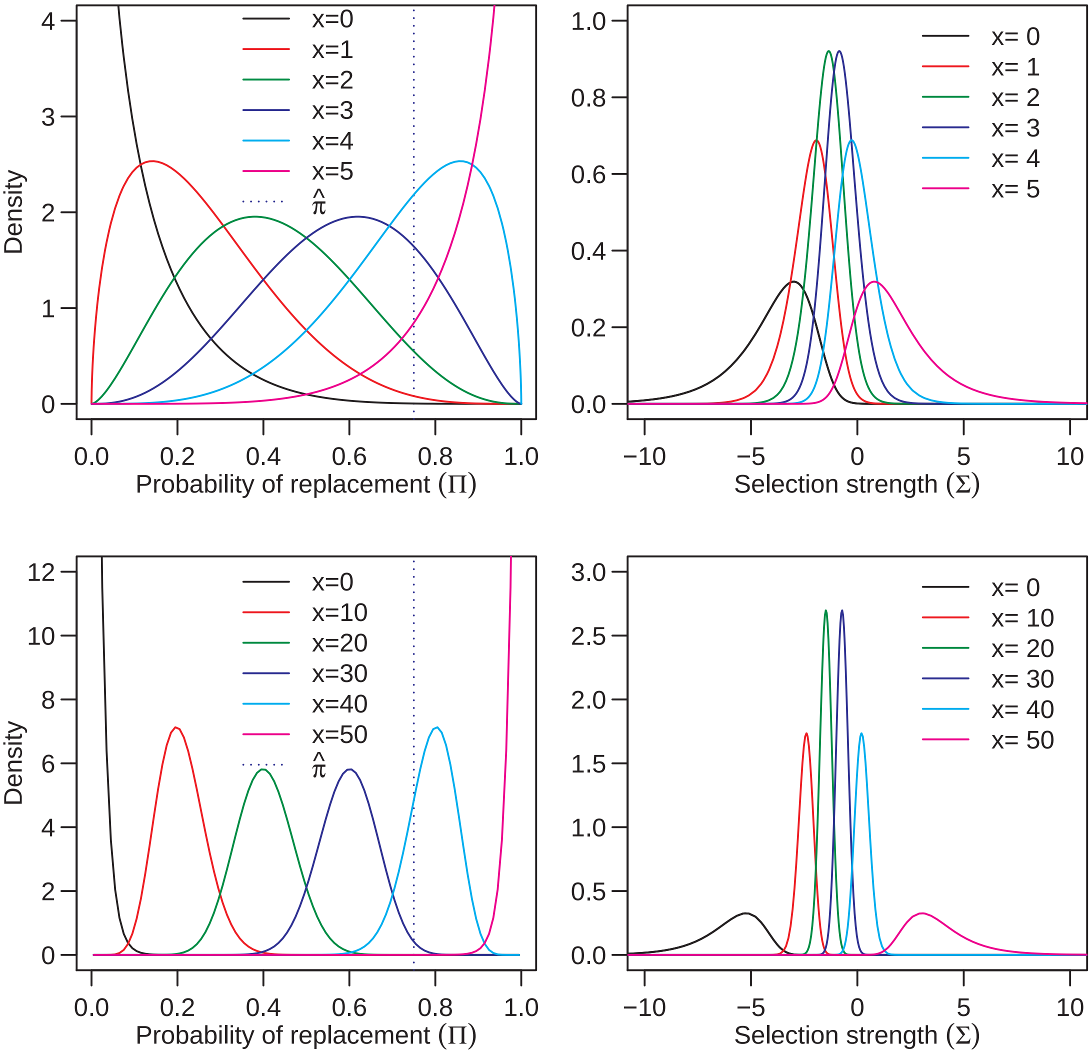

Figure S2: Moving from the frequency of replacement mutations (π) to selection strength (Σ). The Bayesian posterior distribution was calculated for different values of x (individual curves) and N ($N = 5$ for upper panels and 50 for lower panels). In all cases, the expected frequency $\hat{\pi} = 0.75$.

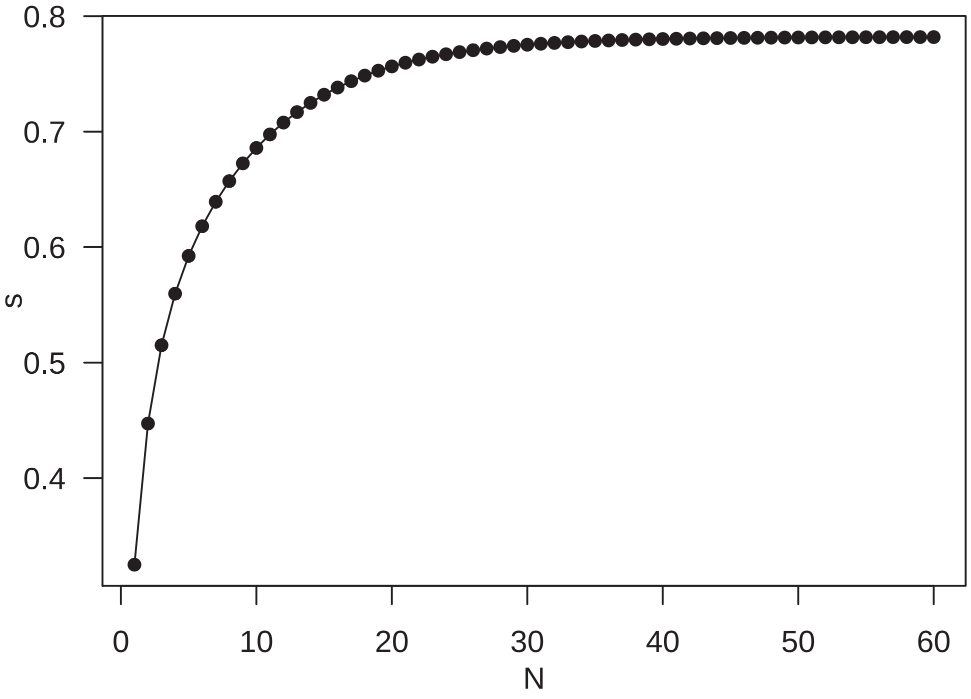

Figure S3: Estimated values of the hyperparameters ($a = b = s$) for the Beta prior. At each value of $N \in \{1...60\}$, fitting was carried out as in Figure 2b.

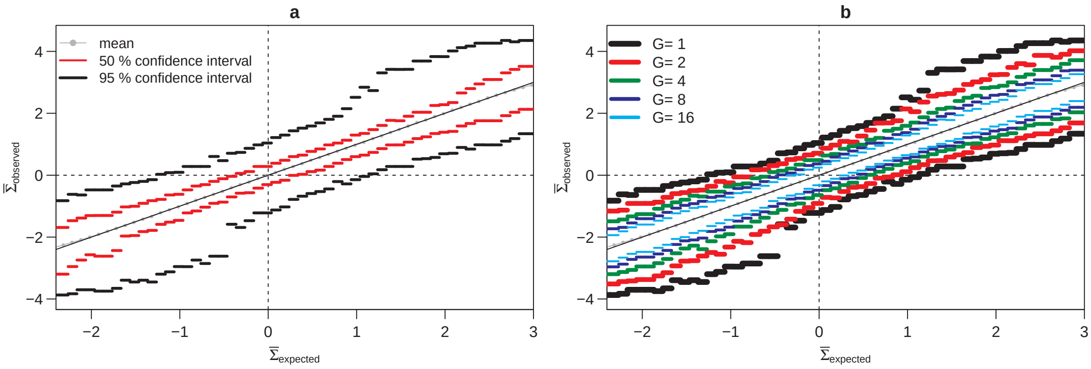

Figure S4: Validation of the Bayesian framework Analogous plot to main figure 2(d,e), but using a binomial-based simulation with an expected replacement frequency ($\hat{\pi}$) = 0.43, and the number of mutations drawn randomly between 5 and 25.

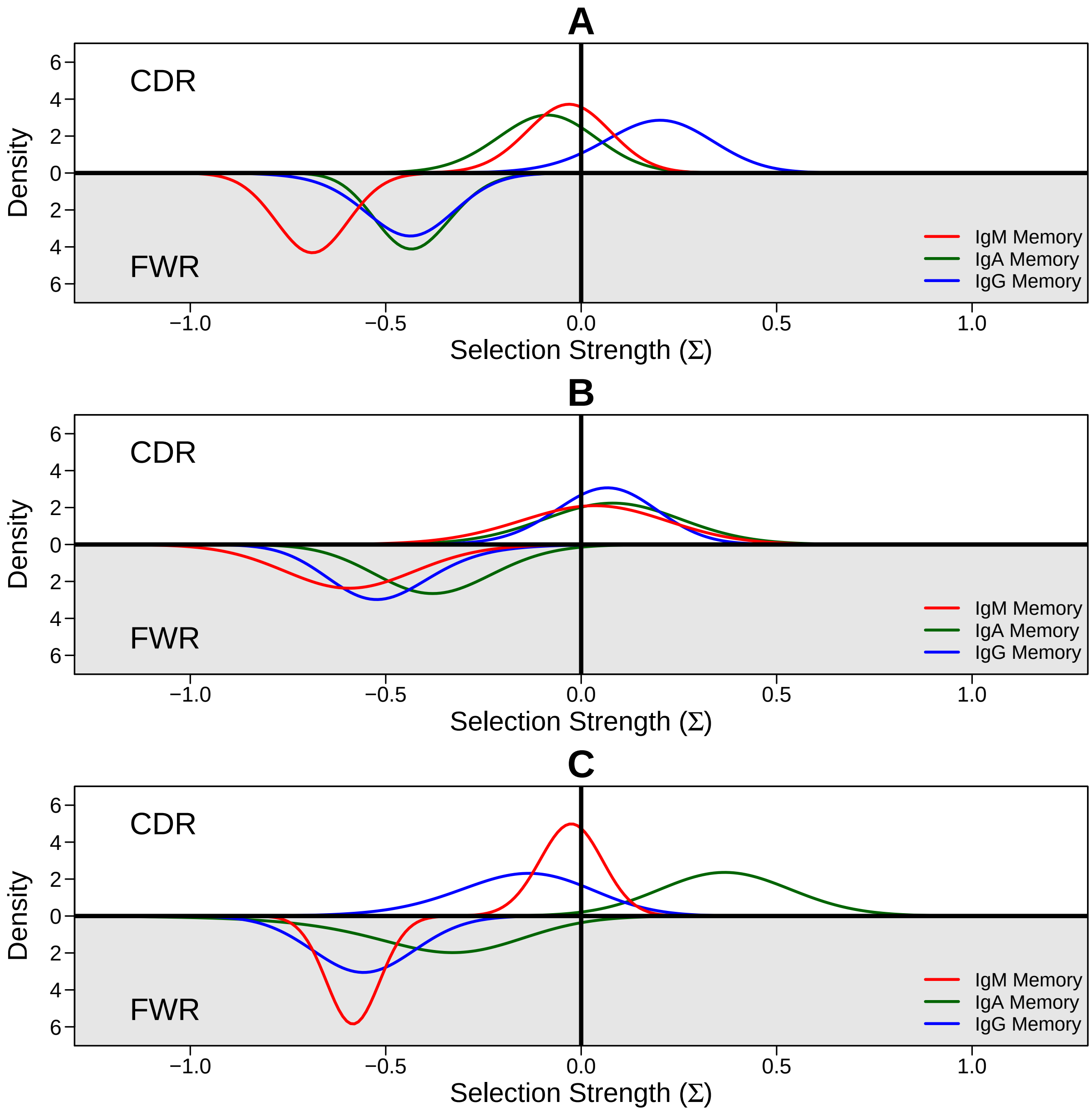

Figure S5: **Selection analysis from figure 3c, carried out seperately for the three individual subjects (A,B,C)**

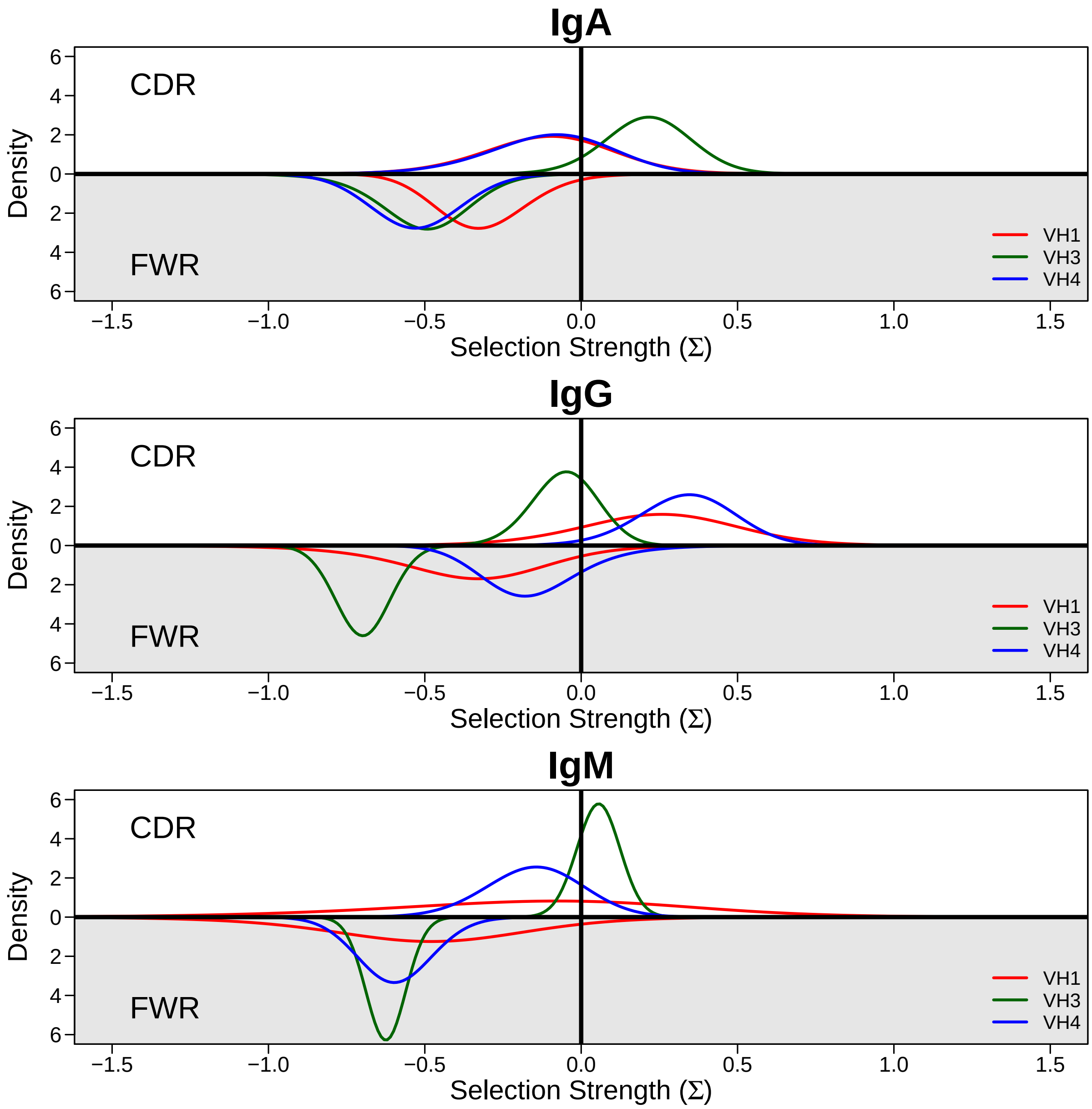

Figure S6: **Selection analysis from figure 3c, carried out seperately for the three cell isotype (IgA, IgG, IgM)**

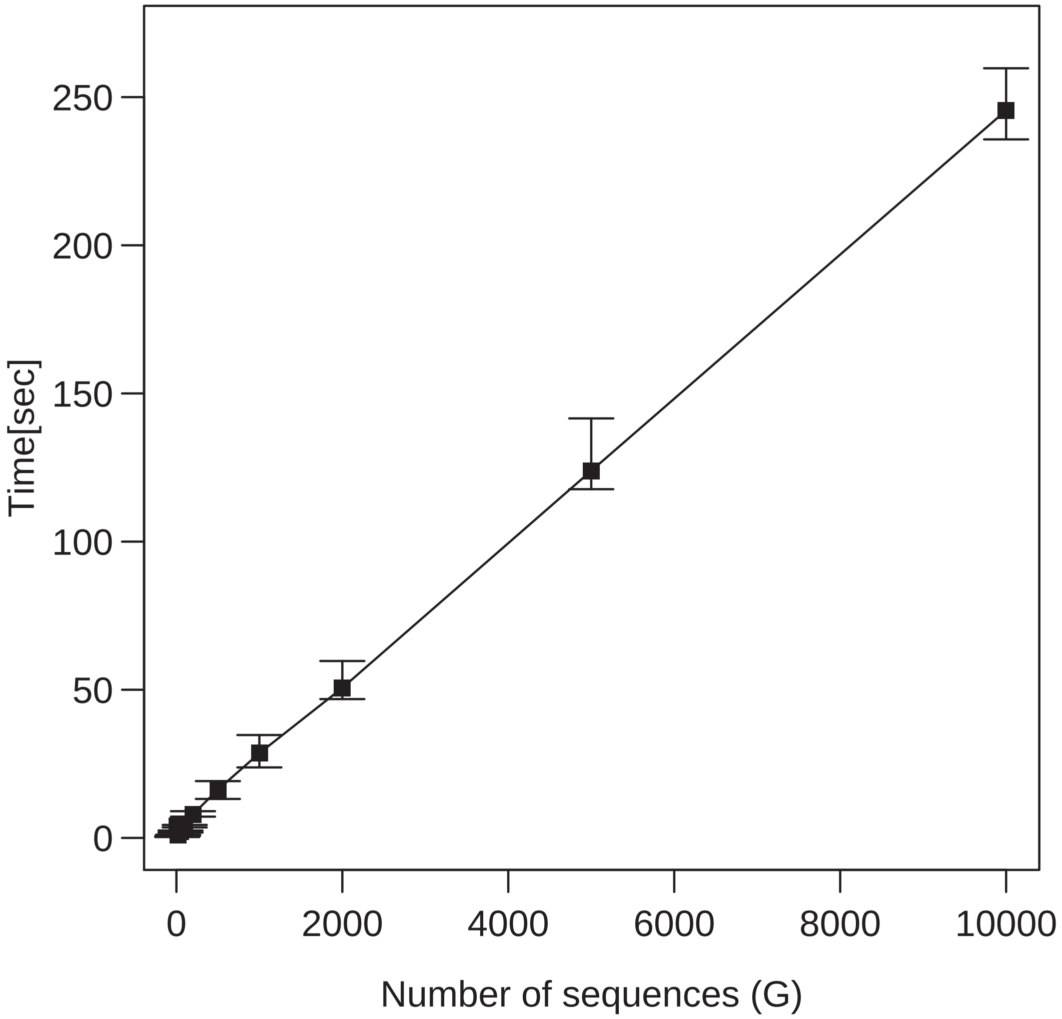

Figure S7: **Performance of the method on a single 1.73GHz processor** sequences were sampled from from high-throughput sequencing dataset with 46 different germline segments and an average of 23 mutations per sequence. Error bars represent 95% intervals of the 50 runs made for each value of G.