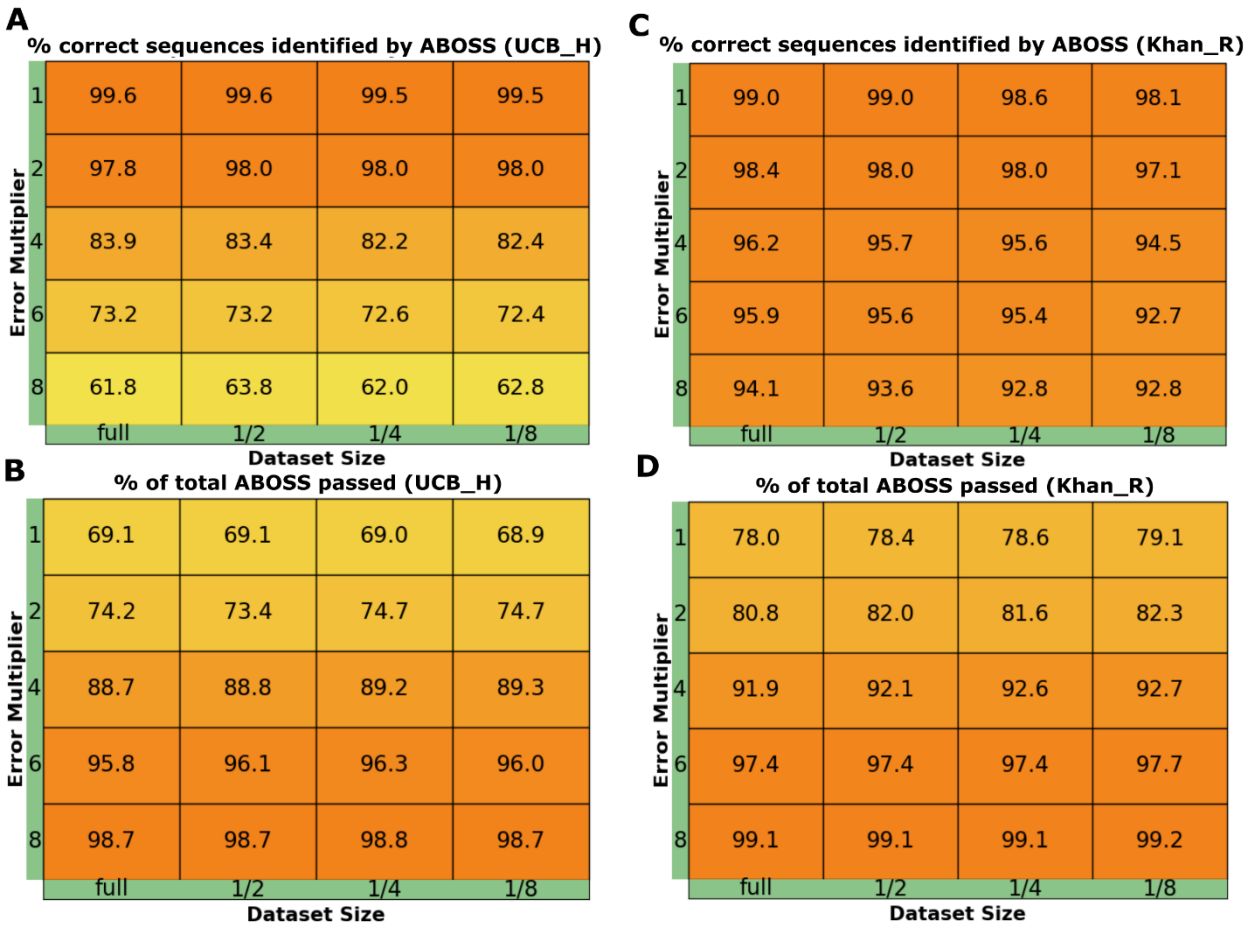


Figure S1. **Ig-seq error simulation to assess sequences volumes and error rate tolerance of ABOSS**. The percentage outputs from Ig-seq error simulations of UCB_H **(A-B)** and Khan_R **(C-D)** datasets. The X-axis corresponds to the proportions of UCB_H_Sim and Khan_R_Sim dataset sizes used for simulation relative to the size of the dataset that passed ANARCI. The Y-axis shows the multiplier of the original distribution of flagged residue/positions in the Ig-seq datasets (see Figure 3). Plots **(A,C)** indicate the percentage of correct sequences that were incorporated in the UCB_H_Sim and Khan_R_Sim dataset that passed the ABOSS analysis, while plots **(B,D)** present the percentage of these sequences relative to the total number of UCB_H_Sim and Khan_R_Sim sequences respectively that passed ABOSS.

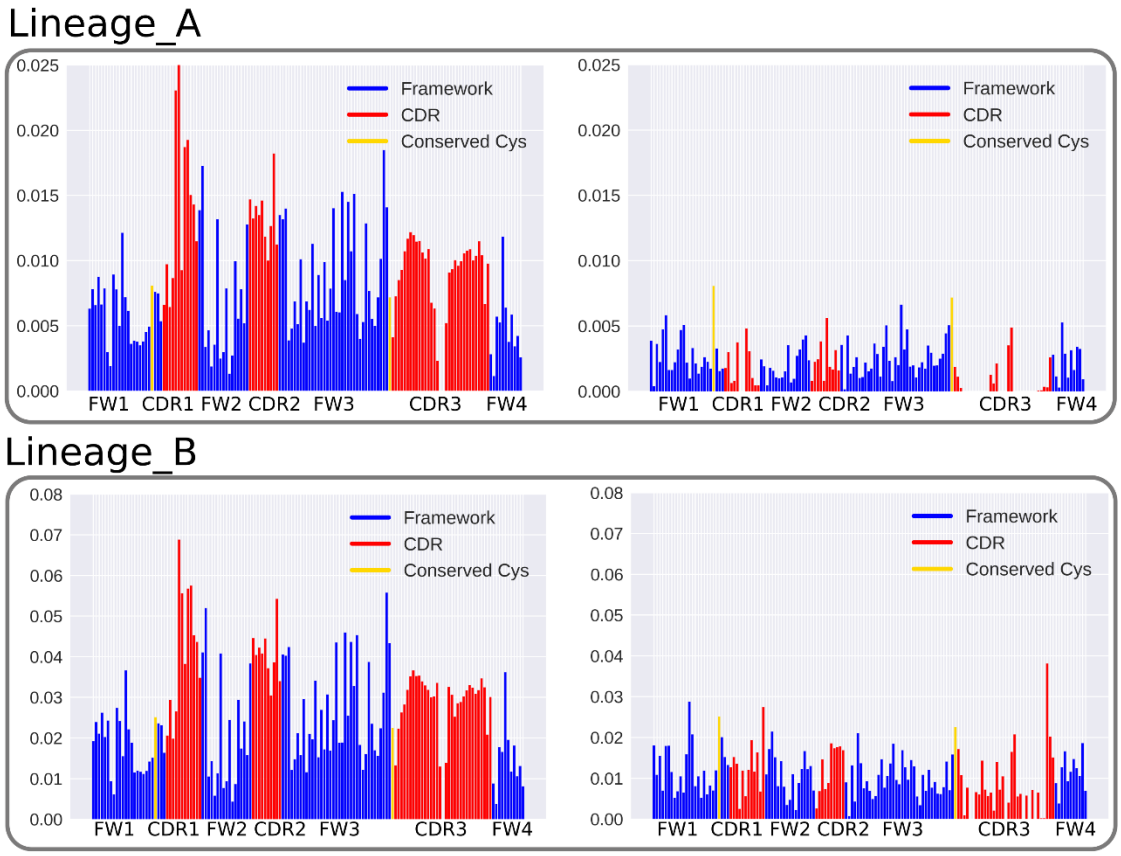


Figure S2. **ABOSS performance on SHM simulated Ig-seq data diversity**. Two antibody clonal lineage trees (Lineage_A and Lineage_B) were employed to provide the background mutational reference to introduce SHM substitutions into the ABOSS filtered Healthy_H dataset using the human HH_S5F targeting model (33). The x-axis shows positions along the VH chain, and the y-axis shows the proportions of residue/positions in the simulation datasets. The figures on the left depict the proportion of SHM substitutions introduced at positions in the VH chain. The figures on the right represent the proportions of ABOSS flagged residue/positions in the simulation datasets.

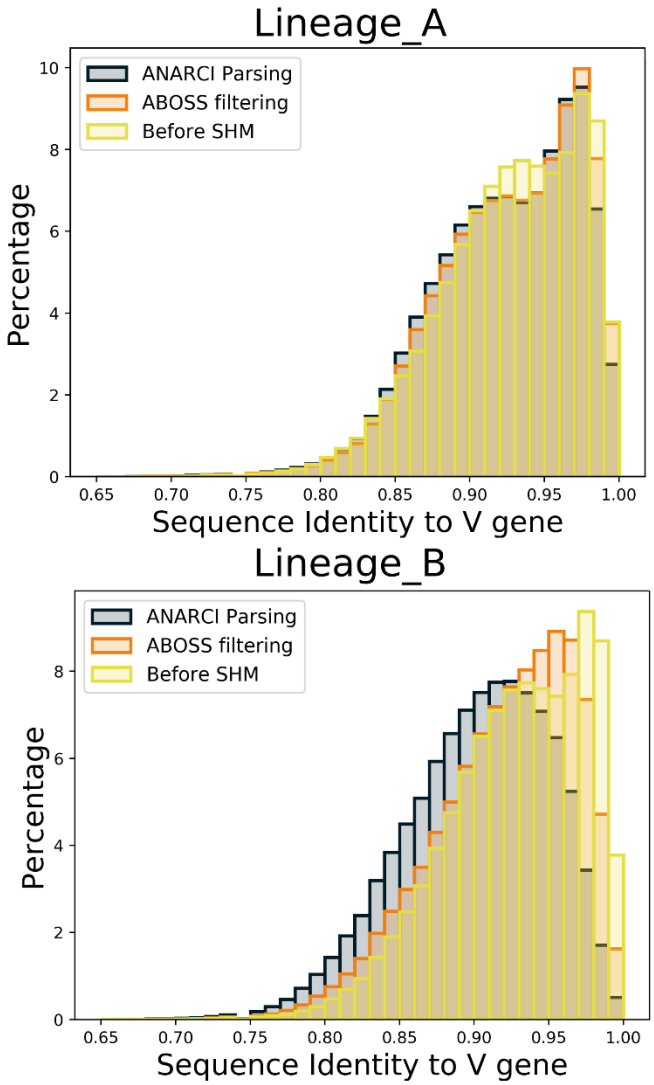


Figure S3. **Relationship between ABOSS filtering and V gene sequence identity.** The sequence identities to the closest V gene for ABOSS filtered UCB_H data (yellow), the SHM simulated (Lineage_A and Lineage_B) UCB_H data (grey) and ABOSS filtered SHM simulated UCB_H data (orange). The x-axis gives the sequence identity of the Ig-seq data to the closest V gene sequence, while the y-axis displays the percentage with that sequence identity. The majority of sequences in the ABOSS filtered data had germline mismatching residue/positions, less than 4% of the total sequences were identical to the V genes. Lineage_B had a higher substitution rate than Lineage_A. This produced a higher percentage of sequences in the Ig-seq data with lower

identities to the closest V gene as well as a higher residue error rate. As these sequences harbored an increased number of SHM substitutions, there was an increased probability for some of these substitutions to be found below the residue error rate in ABOSS analysis.

| Dataset | Dataset size | Found in ABOSS filtered out data | Found in ABOSS filtered out data (%) |
|---|---|---|---|
| IgReC-corrected UCB_H | 5,572,963 (4,069,318) | 1,693,246 (1,136,620) | 30% |
| IgReC-corrected Healthy_H | 1,303,128 (367,235) | 437,652 (166,842) | 33% |

Table S1. **Study of sequence overlaps between ABOSS filtered out data and IgReC-corrected data**. To investigate whether IgReC misses structurally incorrect sequences, ABOSS filtered out UCB_H and Healthy_H sequences were searched in the respective IgReC-corrected data. In both the IgReC-corrected UCB_H and IgReC-corrected Healthy_H datasets, roughly 30% of the sequences were found in the ABOSS filtered out data. The numbers of non-redundant sequences in the data are shown in parenthesis.