Bioinformatics, 33(24), 2017, 3938–3946
doi: 10.1093/bioinformatics/btx533
Advance Access Publication Date: 31 August 2017
Original Paper

Phylogenetics

Comparison of methods for phylogenetic B-cell lineage inference using time-resolved antibody repertoire simulations (AbSim)

Alexander Yermanos[1], Victor Greiff[1], Nike Julia Krautler[2], Ulrike Menzel[1], Andreas Dounas[3], Enkelejda Miho[1], Annette Oxenius[2], Tanja Stadler[1] and Sai T. Reddy[1,*]

[1]Department of Biosystems Science and Engineering, ETH Zürich, 4058 Basel, Switzerland, [2]Institute of Microbiology and [3]Department of Chemistry and Applied Biosciences, ETH Zürich, 8093 Zürich, Switzerland

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on December 20, 2016; revised on August 14, 2017; editorial decision on August 18, 2017; accepted on August 30, 2017

Abstract

Motivation: The evolution of antibody repertoires represents a hallmark feature of adaptive B-cell immunity. Recent advancements in high-throughput sequencing have dramatically increased the resolution to which we can measure the molecular diversity of antibody repertoires, thereby offering for the first time the possibility to capture the antigen-driven evolution of B cells. However, there does not exist a repertoire simulation framework yet that enables the comparison of commonly utilized phylogenetic methods with regard to their accuracy in inferring antibody evolution.
Results: Here, we developed AbSim, a time-resolved antibody repertoire simulation framework, which we exploited for testing the accuracy of methods for the phylogenetic reconstruction of B-cell lineages and antibody molecular evolution. AbSim enables the (i) simulation of intermediate stages of antibody sequence evolution and (ii) the modeling of immunologically relevant parameters such as duration of repertoire evolution, and the method and frequency of mutations. First, we validated that our repertoire simulation framework recreates replicates topological similarities observed in experimental sequencing data. Second, we leveraged Absim to show that current methods fail to a certain extent to predict the true phylogenetic tree correctly. Finally, we formulated simulation-validated guidelines for antibody evolution, which in the future will enable the development of accurate phylogenetic methods.
Availability and implementation: https://cran.r-project.org/web/packages/AbSim/index.html
Contact: sai.reddy@ethz.ch
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Antibody repertoires, characterized by variable regions of the immunoglobulin B-cell receptor (Greiff *et al.*, 2015b; Tipton *et al.*, 2015), provide protective immunity by adaptation and mutation in response to antigen challenge. The variable region of the antibody heavy chain largely dictates specificity to antigen (Xu and Davis, 2000); it is produced from the somatic rearrangement of variable (V), diversity (D) and joining (J) germline genes during B-cell development (Tonegawa, 1983). In one of the most common laboratory mouse strains, C57BL/6, the fully annotated immunoglobulin locus consists of 164 V-genes, 27 D-genes and 4 J-genes (Johnston *et al.*, 2006), which allows for thousands of possible somatic recombinations, and thus a highly diverse antibody repertoire.

Quantitative investigation of antibody responses has become possible due to progress in high-throughput immunoglobulin repertoire sequencing (Ig-Seq) (Georgiou *et al.*, 2014; Yaari and Kleinstein,

2015). Many of these studies have focused extensively on antibody variable heavy chain repertoires, since they define B-cell clonality and often dictate antigen-specificity (Xu and Davis, 2000). Upon antigen challenge, B-cells expand and hypermutate their antibody variable regions, thus forming a B-cell lineage that spans from the naïve unmutated B-cells to isotype-switched, hypermutated memory B-cells (Janeway and Murphy, 2011) to terminally differentiated plasma cells (Manz *et al.*, 1997). Thus, antibody repertoire sequences represent both the immunological past and present of an individual.

Retracing antibody repertoire evolution across time would enable unprecedented insight into how vaccines (Jackson *et al.*, 2014) and pathogens shape the humoral immune response (Wang *et al.*, 2015; Zhu *et al.*, 2013). Several phylogenetic methods, such as Levenshtein distance (LD), neighbor-joining (NJ), maximum parsimony (MP), maximum-likelihood (ML) and Bayesian inference (BEAST) have been used for delineating the evolution of B-cell clonal lineages from antibody repertoire sequencing data (Andrews *et al.*, 2015; Barak *et al.*, 2008; Stern *et al.*, 2014; Wu *et al.*, 2015). The recent increase of antibody repertoire studies that sequence B-cell populations within the same individual at different time points (Ellebedy *et al.*, 2016; Wang *et al.*, 2015, 20) suggests that the utilization of methods that quantify repertoire evolution will continue to increase. However, the performance and comparison of the output of these methods has yet to be rigorously evaluated as antibody repertoire simulation suites do not yet incorporate time-dependent information (Safonova *et al.*, 2015).

Here, we have developed the R package AbSim, which enables the simulation of time-resolved antibody repertoires, thus producing a set of sequences that can be used to evaluate the performance and accuracy of both phylogenetic methods in reconstructing the evolution of a given B-cell clone over time and methods aiming to delineate independent B-cell lineages. The AbSim package allows the user to comprehensively control the vast biologically relevant parameter space of antibody repertoire selection and evolution. Specifically, the control of the following parameters is possible: (i) total time of evolution, (ii) rate and method of somatic hypermutation (SHM), (iii) number and rate of V-D-J recombination events, (iv) rate at which new sequences are produced, (v) baseline mutation rate, (vi) clonal frequency, (vii) and germline gene (V-D-J) usage distribution. By generating antibody repertoires with AbSim, we quantitatively compared reconstruction of B-cell lineage evolution with the following phylogenetic methods: LD, NJ, MP, ML and BEAST. Phylogenetic trees can be compartmentalized into clades, which contain temporal information as they are defined as sets of descendent sequences that all share a common ancestor. Thus, correctly inferring the clades of a phylogenetic tree is crucial for describing the evolutionary relationship between clonally selected and expanded B-cells (i.e. memory B cells) that belong to a given lineage (i.e. derived from a naïve B cell). We found that the clade prediction accuracy of all phylogenetic methods evaluated decreased from >88% to <50% with increasing inclusion of independent B-cell clonal lineages (separate V-D-J recombination events). However, under certain parameter conditions, simulated topologies could be predicted with near perfect accuracy (>95%)—this was especially the case when the mutations per time step increased. Furthermore, we confirmed that our simulation suite could reproduce the topological trends observed in time-resolved biological antibody repertoires [a mouse model of chronic viral infection, lymphocytic choriomeningitis virus (LCMV)] (Richter and Oxenius, 2013). Lastly, we validated that biologically relevant parameters could be accurately inferred from simulated phylogenies by showing that BEAST could accurately predict the mutation rate and duration of B-cell lineage evolution (tree height). In future work, we expect AbSim will enable the development and validation of more dedicated and highly predictive B-cell lineage inference methods, thereby fundamentally increasing our insight into antibody repertoire evolution.

2 Materials and methods

2.1 Simulation of time-resolved B-cell repertoires

The R package AbSim simulates in either a single-lineage or multi-lineage fashion B-cell antibody repertoires composed of the variable region of both heavy and light chains from either human or C57BL/6 mouse (in this study we focused on the variable heavy chain of C57BL/6 mice because most studies have analyzed IgH) (Andrews *et al.*, 2015; Greiff *et al.*, 2014). However, the package enables separate modeling of heavy and light chain evolution). Each simulation starts with a set of unmutated germline V-, D- and J-genes (Giudicelli *et al.*, 2004; Johnston *et al.*, 2006) to generate independently recombined V-D-J lineages. Each iteration through the algorithm represents one simulated time step (Supplementary Fig. S5). At each time step, there is a user-defined probability α, for which a new V-D-J recombination event occurs. Each new V-D-J recombinant starts a new branching event. At each time step, each tip in a phylogeny has a probability β for a SHM event to occur. This probability can either be equal for all clones or follow a different distribution defined by the user. Upon SHM, a branching event is produced in the phylogeny, with one descendant being the original lineage and the second descendant being the hypermutated one. Finally, each iteration can introduce baseline mutations with each site having a γ probability per time step to mutate (Supplementary Fig. S5). These mutations affect the current lineages of a phylogenetic tree without producing a new branching event in the phylogeny. Baseline mutations could represent cells that undergo mutations without giving rise to two distinct daughter cell lineages. The user can provide the frequency of these mutations as input. The simulations will run until either the maximum time limit (T_{max}) or the maximum number of sequences (N_{max}) has been reached. Below, each simulation step is described in detail.

2.1.1 Simulation of V-D-J recombination

V-D-J recombination was simulated based on previous knowledge of somatic diversification (Muramatsu *et al.*, 2000; Saada *et al.*, 2007). Each lineage is characterized by two independent recombination events. First, the D- and J-gene segments were combined, followed by the joining of the newly formed D-J-gene segment with the V-gene (Supplementary Fig. S5). Each independent joining event can include an insertion, deletion, or simply append the two segments. For an insertion, there was an equal probability of adding 2, 4, 6, 8 or 10 nucleotides at the given junction (Feeney, 1990; Mroczek *et al.*, 2014; Saada *et al.*, 2007), although alternative mouse or human-specific models based on experimental sequencing data (Collins *et al.*, 2015; Elhanati *et al.*, 2015) may be selected by the user. Nucleotides A, C, G and T had an equal probability of being inserted. For a deletion event, each end of the junction had a uniform probability to lose 0–5 nucleotides (Saada *et al.*, 2007). For the current study, potential introduction of frame-shift mutations were ignored, as the analysis was focused on sequence evolution rather than functionality of output sequences.

In addition to simulating clonal evolution from a single B-cell lineage (one set of V-, D- and J-genes), AbSim can simulate the evolution of comprehensive antibody repertoires (Supplementary Fig. S4). Specifically, germline gene usage can either be uniform for each

V-, D- and J-gene, or it can follow a custom distribution specified by the user, as previous work indicated non-uniform germline gene usage in humans and mice (Glanville *et al.*, 2011; Greiff *et al.*, 2017). Additionally, the user can add or delete certain germline genes to be incorporated into the simulation if desired. The clonal frequency follows a power law distribution by default, as this was shown in previous findings from antibody repertoire sequencing (Greiff *et al.*, 2015a; Mora *et al.*, 2010; Weinstein *et al.*, 2009). The user can however supply a custom distribution for this parameter.

2.1.2 Simulation of somatic hypermutation (SHM)

AbSim can currently simulate SHM by three different methods. (i) In the 'Poisson' mutation method, each nucleotide in a given sequence has a user-defined probability μ of being randomly mutated to any of the other nucleotides. (ii) A 'data-driven' method involves location-specific mutations targeting nucleotides in complementary determining regions (CDRs), where each nucleotide has a user-defined probability for mutation, defined by ν. Additionally, this method allows the use of different mutation rates for transitions ($p_{\text{transition}}$) and transversions ($p_{\text{transversion}}$), which is relevant for antibody evolution (Cui *et al.*, 2016). (iii) A third method uses substitution probabilities from 5-mer motifs determined from previous high-throughput sequencing studies (Yaari *et al.*, 2013). These motif-based substitution probabilities are given for the middle nucleotide in a sequence window of five nucleotides, and are unique for each possible combination. The user can define M_{max}, the maximum number of motif-based mutations per SHM event (Supplementary Fig. S1). Additionally, AbSim can simulate SHM by combining all three of these methods with user-defined weights for each method.

2.2 Phylogenetic methods

Phylogenetic trees were inferred using the five most common methods: (i) LD, (ii) NJ, (iii) ML, (iv) MP and (v) BEAST. All methods, with the exception of LD, were reconstructed by first aligning the sequences using Clustal Omega (Sievers *et al.*, 2014). For the classical NJ method, the output phylip files were subsequently read into R using phangorn's read.phyDat function (Schliep, 2011); distance matrices were constructed using ape's dna.bin function with the default settings, which use a kimura80 nucleotide substitution model to construct the distance matrix (Paradis *et al.*, 2004). These matrices were subsequently converted to trees using ape's NJ function, and the unmutated germline, composed of the appended simulation's input V-, D- and J-genes (with no insertions/deletions), was set as the outgroup using R-package phytools' reroot function (Revell, 2012). LD-based trees were produced by initially constructing a pairwise distance matrix using R package stringdist's stringdistmatrix function with the nucleotide sequences as input (Loo, 2014). This distance matrix was subsequently used as input into ape's nj function, and the unmutated germline was set as the outgroup (as performed with LD trees). ML trees were inferred using RAxML with a GTRgamma nucleotide-substitution model (Stamatakis, 2014) with the unmutated germline set as the outgroup. MP trees were generated using the R package Rphylip (Revell and Chamberlain, 2014), which is an R adaptation of Felsenstein's Phylip package (Felsenstein, 1989). The function Rdnapars was used to generate MP trees (the first tree was selected when multiple trees were produced). The unmutated germline was set as the outgroup for both ML and MP trees, as described previously by Stern and colleagues (Stern *et al.*, 2014).

BEAST (version 2.4.2) was used as a Bayesian method for phylogenetic reconstruction (Bouckaert *et al.*, 2014). The GTR model with gamma-distributed site heterogeneity was used, with initial transition probabilities automatically estimated for all parameters. The gamma category count was set at 4, with the shape parameter of the gamma site model set to 1, and the substitution rate was automatically estimated. A relaxed lognormal clock model was used to estimate divergence times, with the number of discrete rates set at -1 and the clock rate parameter set to 1 (Drummond *et al.*, 2006). A coalescent Bayesian skyline was used as a tree prior (Drummond, 2005), with 'popSize' parameter following a gamma distribution. The chain length of the Markov chain Monte Carlo (MCMC) algorithm, which is employed by BEAST to sample from a posterior probability distribution, was set to 100 000 000 iterations to ensure convergence of the posterior distribution for all simulation runs. The maximum clade credibility (MCC) tree was extracted using the program TreeAnnotator (Bouckaert *et al.*, 2014), with a 10% burn-in to ensure that the analyzed trees had reached equilibrium in the posterior distribution.

2.3 Topology scoring

To quantify the accuracy of the phylogenetic methods tested, the estimated clades (set of descendent sequences at a given internal node) were compared to the known clades of the simulated trees. The percentage of correctly identified clades for each method, referred to as the clade accuracy, was used as a measure for topology reconstruction accuracy. At least ten simulations were run for each set of parameters to ensure that the stochastic nature of the simulation did not lead to spurious conclusions.

2.4 Treescape metric

In addition to topology scoring by quantification of clade overlap, the topologies of both experimental and simulated trees were compared using a recently described treescape metric for trees that are composed of the same set of tips (Kendall and Colijn, 2016). The treescape metric takes labeled, rooted trees as input and allows for a clear visualization of topological similarities based on shared edges between two sequences within the tree. The output trees from the five phylogenetic methods examined (LD, NJ, MP, ML tree, as well as MCC tree for BEAST) were compared with 1000 randomly selected posterior trees produced during BEAST's MCMC algorithm. These 1000 MCMC indicate whether the MCMC algorithm employed by BEAST has explored the regions occupied by the other four methods. The tree space was then visualized using the findGrooves function of the treescape R package (Jombart *et al.*, 2017). The lambda parameter in treescape was set to 0 to ensure that the metric was only analyzing the topology and not the branch lengths. Five principal components with six clusters were selected as demonstrated by the package creators (Jombart *et al.*, 2017). The distance from the known tree was extracted using the function refTreeDist from the treescape R package (Jombart *et al.*, 2017), again with the lambda parameter set to 0. The distance between a given phylogenetic method and the known, simulated tree can then be compared as an additional measure of topological accuracy. This metric, in addition to the clade accuracy (Section 2.4), is well-suited to compare different trees when using identical input sequences, which best reflects phylogenetic analysis on Ig-Seq data.

2.5 Parameter inference with BEAST

BEAST infers phylogenetic trees with branch lengths in calendar time, which in turn allows for the inference of parameters such as mutation rate or tree height. The tree height is the distance from the root to the farthest tip, thus providing a measure of how long the

evolutionary process has been occurring. The mutation rate describes the expected number of mutations each site will undergo per time step. Thus, the influence of different sampling schemes on both the inferred tree height and the inferred mutation rate was explored using simulated antibody lineages. Tree heights were set to 90 time steps to mirror a chronic LCMV infection, which persists for ~90 days. The probability for each site to mutate at each time step was set to 3.143×10^{-3} (Fig. 6) reflecting previously reported SHM rates (McKean *et al.*, 1984). Two sampling points were included throughout the 90 time steps, at time step 30 and 60, respectively. Four different sampling schemes were explored to examine how undersampling influences both the inferred tree height and the mutation rate.

1. All sequences from all time points were included in phylogenetic reconstruction.
2. Half of the sequences from time steps 30 and 60 were randomly selected, whereas all of the sequences from time step 90 were included.
3. All of the sequences from time steps 30 and 60 were included, whereas half of the sequences from time step 90 were included.
4. Half of the sequences were randomly sampled at each time point and included.

2.6 Mouse experiments

One female C57BL/6 mouse (8 weeks old) was chronically infected with 2×10^6 ffu LCMV Clone 13 intravenously. 200 µl of blood was taken from the leg vein 10 days before infection, along with 9, 17, 26 and 50 days post-infection. Blood was sampled in heparin coated tubes and frozen in 1.5 ml trizol at -80 °C (Fig. 1). Day 50 post-infection, the mouse was sacrificed and plasma cells isolated from the bone marrow and sort purified by flow cytometry (FACS Aria). Non-B cells were excluded by gating on NK1.1 (natural killer cells), CD4 (T-cells), CD8 (T-cells), Gr-1 (granulocytes), F4/80 (macrophages and monocytes); naïve B-cells were excluded by gating on the IgM/IgG negative fraction and plasma cells were isolated by gating on the $CD138^{bright}CD19^{low-int}$ population.

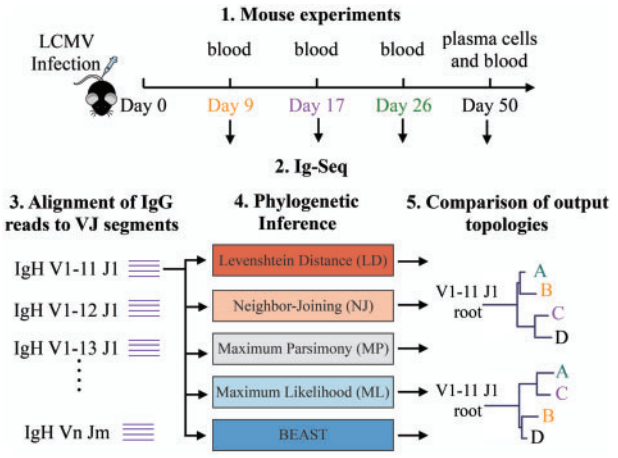

Fig. 1. Overview of experimental outline. Ig-Seq was performed on blood samples and bone marrow plasma cells from a mouse chronically infected with LCMV. The IgG sequencing reads were pooled together and separated by V- and J- germline elements. For each V-J combination, five trees were inferred using either Levenshtein distance, neighbor-joining, maximum parsimony, maximum likelihood, or BEAST. The output trees were compared to determine topological discrepancies between methods

2.7 Ig-Seq library preparation and data preprocessing

Blood samples were resuspended in 2 ml additional trizol. RNA was extracted with 250 µl of chloroform, eluted in 25 µl water and subsequently stored at -80 °C. First-strand cDNA was synthesized using maxima reverse transcriptase (Fermentas) using 12.5 µl of RNA and oligo(dT) primers (Thermo Scientific) following the manufacturer's protocol. Antibody sequencing libraries were prepared by PCR on cDNA as previously described (Menzel *et al.*, 2014). Briefly, variable heavy chain regions were amplified by PCR using a set of 19 forward primers with the gene-specific regions annealing to the framework 1 of the V-D-J region (Krebber *et al.*, 1997) and reverse primers with the gene-specific region binding to the IgG constant region 1 (5′ CARKGGATRRRCHGATGGGG 3′) or IgM constant region 1 (5′ CG AGGGGGAAGACATTTGGG 3′). A first PCR step was performed using Taq polymerase (NEB) in a reaction volume of 50 µl with overhang-extended primers under the following conditions: 4 × 50°C, 4 × 55°C, 12 × 63°C. PCR clean up was performed to reduce the final sample volume to 20 µL, followed by gel-purification. A second PCR step was performed with overhang-specific primers that included Illumina sequencing adapter regions, with cycles of 2 × 40°C, 6 × 65°C. PCR clean-up and gel-purification was performed; quality of the libraries was assessed using a fragment analyzer (Bioanalyzer, Agilent). All libraries passing quality control (uniform product at expected size) were pooled and sequenced on the Illumina MiSeq platform using 2 × 300 base pair (bp) paired-end kit. Antibody sequences were clonotyped by entire V-D-J region and aligned to the C57BL/6 germline (Johnston *et al.*, 2006) using the MiXCR software platform (Bolotin *et al.*, 2015). Reads that did not map to a given V- or J-gene were excluded from the analysis. IgG clones were separated and analyzed separately.

2.8 Phylogenetic analysis of experimental Ig-Seq data

IgG sequences were separated by V- and J-gene alignments and subsequently used as input to the various phylogenetic inference methods. The topologies of resulting trees were compared in order to quantify the extent of clade overlap between the methods. We determined the extent of clade overlap between phylogenetic inference methods for each of the V-J combinations that had more than three sequences (Fig. 1). Following separation by V- and J-gene alignments, there were 184 V-J roots that had more than three sequences, with an average of 51 sequences per germline combination (Supplementary Fig. S2b). The D-gene alignment was ignored for rooting phylogenetic trees from experimental data due to low alignment accuracy (Bolotin *et al.*, 2015). Each V-J root was further separated by CDR3 length, in an attempt to reduce the number of independent V-D-J recombination events per phylogeny (Stern *et al.*, 2014). This additional preprocessing step led to 448 trees that had more than three sequences with an average of 19 sequences per V-J combination (Supplementary Fig. S2a). There were 307 and 7061 V-J roots and V-J roots with identical CDR3s, respectively, before excluding lineages with less than 4 sequences (Supplementary Fig. S2b). Thus, from 7061 clones detected (if defined by V-, D-, J- and CDR3 length) across all five time points, only ~6% of clones had sufficient sequences for phylogenetic analysis. This is may be due to the inherent challenges of obtaining sufficient biological sampling depth with the small number of B cells present in a non-terminal murine blood sample (Greiff *et al.*, 2015b).

2.9 Determination of statistical significance

Significance between groups was tested using the Wilcoxon rank-sum test if not indicated otherwise.

3 Results

3.1 Clade accuracy decreases substantially as the proportion of V-D-J recombination events increases

When performing phylogenetic antibody repertoire lineage reconstruction, the majority of Ig-Seq studies have not accounted for independent V-D-J recombination events that use identical germline genes (Andrews *et al.*, 2015; Green *et al.*, 2013; Seifert and Kuppers, 2009). These separate V-D-J recombination events are assumed to evolve independently, thus they represent different evolutionary topologies. Therefore, we investigated whether the inclusion of distinct V-D-J recombination events originating from the same germline elements (V1-47, D2-8, J1) influences the accuracy of phylogenetic B-cell lineage reconstruction (hereafter referred to as clade [prediction] accuracy). We tested this by simulating the evolution of 50 antibody sequences using the AbSim framework and found that as one increases the proportion of total sequences generated from V-D-J recombination events to SHM events (while keeping all other parameters constant, Supplementary Fig. S1), the clade prediction accuracy decreases from >88% (one V-D-J lineage) to <50% (20 V-D-J lineages); this trend was observed for all examined phylogenetic methods (Fig. 2). These findings indicate the need to minimize the number of V-D-J recombination events when reconstructing lineage trees (Ralph and Matsen, 2016; Stern *et al.*, 2014).

3.2 Altering the rate and method of simulated mutations can greatly improve clade accuracy

Given that the number of V-D-J recombination events significantly decreased the clade prediction accuracy, the following simulations were performed with one V-D-J recombination event per tree. Thus, a new V-D-J recombination event starts the tree, with SHM causing the subsequent branching events. All simulations were repeated 10 times per parameter set and used the randomly selected germline sequences V1-47, D2-8, J1. The parameter settings that dictated SHM simulation substantially influenced the accuracy of phylogenetic reconstruction (Fig. 3, Supplementary Fig. S1). As the number of mutations between branching events increased, the clade prediction accuracy with all of the phylogenetic methods increased (Fig. 3a). When there were no mutations present between branching events, trees produced using the LD (∼66%) and BEAST (∼68%) predicted the true topology with the highest accuracy, whereas MP performed poorest (∼55%) (Fig. 3a). As the number of mutations per time step increased ($\gamma = 0.01$), all methods predicted the true tree with >94% accuracy (Fig. 3a). Furthermore, we found that the model of SHM influenced the clade prediction accuracy (Fig. 3b). When the mutations followed a Poisson distribution, where each nucleotide had the same probability to mutate into any of the other bases, all methods performed equally well (Fig. 3b), and the mean clade accuracy for all methods was ∼87%. However, when mutations were localized in the complementarity determining regions (CDR) and transition mutations (e.g. A to G) were more common than transversions (e.g. A to T), ML and BEAST predicted the true clades with significantly higher accuracy ($P < 0.05$) than MP (Fig. 3b). This was due to the fact that parametric models allow for different substitution rates for transversions versus transitions. Thus, if the biological SHM mechanism preferentially incudes different rates of mutation for transitions or transversions (Cui *et al.*, 2016) it is important to select a phylogenetic method that can account for this. Lastly, we saw that when using motif-based transition probabilities, all methods performed worse in predicting simulated clades (Fig. 3b), with MP predicting the clades with significantly lower accuracy ($P < 0.05$). Given all parameter settings explored, BEAST and LD performed equally or better than the other methods; whereas MP often predicted the true topology with either equal or lower accuracy.

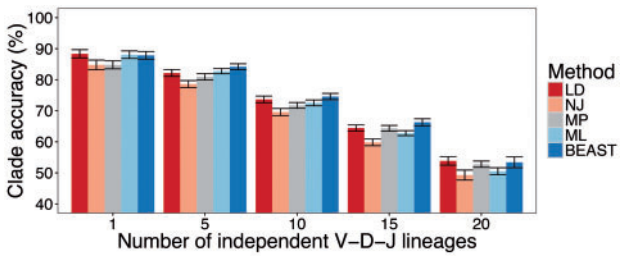

Fig. 2. Clade accuracy decreases as the number of simulated V-D-J events increases. The clades predicted by the five phylogenetic methods were compared to the clades of the known simulated tree. The percentage of the total 50 sequences generated from either V-D-J recombination events or SHM events was varied while keeping all other parameters constant. All bar plots depict mean ± s.e.m. 20 simulations for each parameter setting were run

3.3 Reproduction of topological discrepancies observed in Ig-Seq data

To confirm the correspondence of phylogenies simulated with AbSim with that of experimental data, we compared simulated phylogenies to those inferred from experimental Ig-Seq data, which was derived from B-cells of a mouse with chronic infection (LCMV). Samples were collected longitudinally at regular time intervals (see Section 2.6). Analysis of Ig-Seq data across all samples resulted in 184 lineage trees inferred using the full V-D-J sequence. When applying the phylogenetic methods to this data, the mean clade overlap among all five methods across the 184 lineage trees was only 28% (Fig. 4a). The clade overlap between any two methods ranged from 38 to 51% (Fig. 4a). Thus, the phylogenetic method chosen substantially influenced the output topology for Ig-Seq data. Using simulated data from AbSim, we confirmed that this immense discrepancy between predicted clades could be reproduced (Fig. 4). Similar to the experimental data, only 30% of the clades were identical across all five of the methods when up to 20 V-D-J lineages were allowed (Fig. 4a). Additionally, the mean overlap between any two methods ranged from 33 to 51% (Fig. 4b).

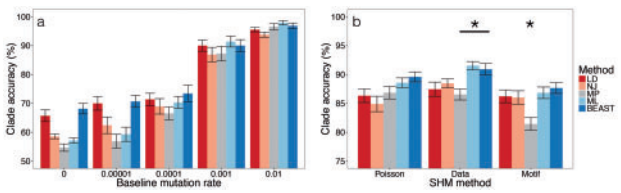

Fig. 3. The accuracy of phylogenetic methods varies as a function of AbSim's parameter settings. (**a**) Clade prediction accuracy was explored under different simulated baseline mutation rates. (**b**) Different models of somatic hypermutation (SHM) simulation can influence the clade prediction accuracy of phylogenetic methods. Each tree had a maximum of 30 sequences. BEAST and ML predicted the true clades with significantly higher accuracy than MP when data-driven mutations were incorporated ($P < 0.05$). With 'motif' driven mutations, MP predicted the true clades with significantly lower accuracy than all other methods ($P < 0.05$). All bar plots depict mean ± s.e.m. See methods (Section 2.1.2) for details on Poisson, Data and Motif driven SHM. Ten and twenty simulation runs per parameter setting were performed in (a) and (b), respectively

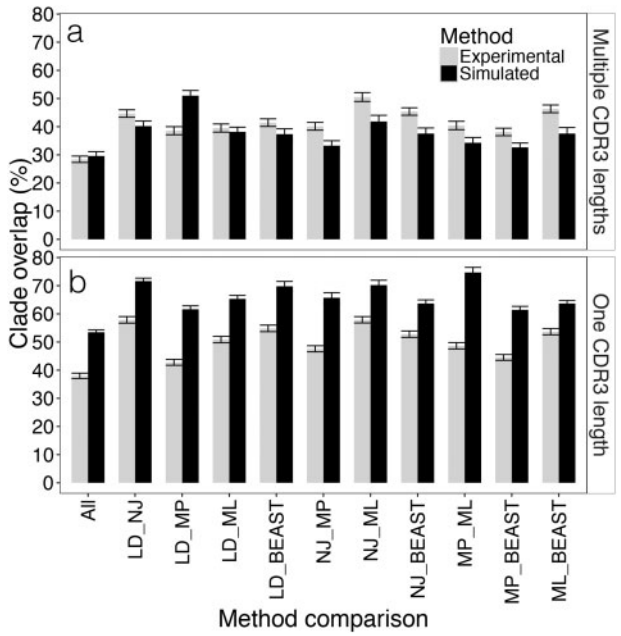

Fig. 4. AbSim reproduces discrepancies between clade overlap observed in experimental Ig-Seq data. (**a**) The clade overlap from both experimental and simulated lineages was compared when identical sequences were used as input for different phylogenetic methods. 20 V-D-J recombination events were allowed per simulation. (**b**) The clade overlap from experimental data was compared as in (a), in addition to separating lineage trees by CDR3 length. All bar plots depict mean ± s.e.m. Ten simulations per parameter setting were run. All simulations were performed using the V, D, J germline genes V1-47, D2-8 and J1, respectively

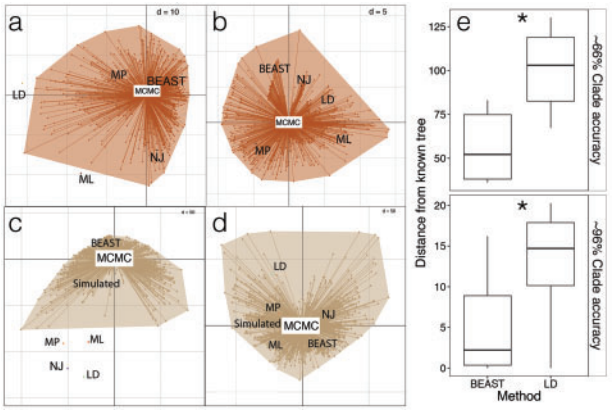

Fig. 5. AbSim can reproduce trends in tree space observed in Ig-Seq data. (**a**) and (**b**) Topologies from two V-J roots were projected into two-dimensional tree space. Both larger (a) and smaller (b) discrepancies could be observed between phylogenetic methods on Ig-Seq data. (**c**) and (**d**) Simulated trees could reproduce both larger (c) and smaller (d) discrepancies between phylogenetic methods, as observed in Ig-Seq data. The MCMC trees are displayed in brown. The differences in tree space seen in the Ig-Seq topologies can be recreated using simulated phylogenies. (**e**) The distance from the known tree was significantly less for BEAST than LD under both high accuracy and low accuracy conditions ($P < 0.05$). Ten simulations per parameter setting were run. To improve readability, both the label sizes were increased and the histograms describing clustering were removed (see Supplementary Fig. S3 for original plots)

Given that we have shown how including multiple V-D-J recombination events within one phylogenic tree substantially decreases clade accuracy (Fig. 2), we compared on experimental Ig-Seq data the clade overlap after separating by CDR3 length, which is one way to filter Ig-Seq data and reduce the number of unique lineages (Stern *et al.*, 2014). This led to an increase in clade overlap across all methods, reaching a maximum overlap of 58% (for LD-NJ and NJ-ML) (Fig. 4b). While for biological samples clade overlap does not necessarily translate to phylogenetic inference accuracy, it is reassuring that the impact of the phylogenetic method can be reduced by simple Ig-Seq filtering. Analogously, we could show using simulated phylogenies that the clade overlap could be increased by only simulating one V-D-J recombination event, as overlap across all five tested methods increased to 53% (Fig. 4). When comparing the methods directly to each other, the overlap ranged from 62 to 75% (Fig. 4b). Thus, when using clade overlap as a metric, AbSim is able to replicate the trends observed in experimental Ig-Seq data.

In addition to analyzing the percentage of clades shared between methods, we visualized the treespace of both simulated and experimental phylogenies. The treespace metric compares tip-labeled trees, i.e. trees containing the same sequences, and projects the topologies onto a two-dimensional space. Thus, if trees share similar topologies, they would be expected to cluster closely in the tree space (Kendall and Colijn, 2016). We can further use this method to compare the distance from a given phylogenetic method to the known, simulated tree. This can be used as a measure of phylogenetic accuracy based on the number of branching events separating two given sequences. The experimental phylogenies produced both trees that were sparsely distributed (Fig. 5a) and trees that clustered closely (Fig. 5b), based on whether the different methods produced trees within the perimeter of the sampled MCMC trees. These trees represent the topologies sampled during BEAST's MCMC algorithm. We could recreate this discrepancy observed in experimental tree space with simulated lineages by changing AbSim's parameter settings (Fig. 5c and d); indeed, we found conditions under which LD, NJ, MP and ML produced trees that were in a tree space that was not explored by BEAST's MCMC algorithm (Fig. 5c). This was primarily the case when the baseline mutation rate γ was set to zero, correspond

ing to low clade prediction accuracy for all methods. Additionally, the MCC trees produced by BEAST were significantly closer to the known, simulated tree both when γ was high (0.01) or zero ($P < 0.05$) (Fig. 5e). This is surprising given that at both γ conditions, BEAST and LD predicted the clades of the true tree with similar accuracy (68% and 65% when $\gamma = 0$, 96% and 97% when γ 0.01) (Fig. 3a). Therefore, even if the clade accuracy was similar between methods, the MCC trees inferred by BEAST were closer to the true phylogeny than the LD trees.

Conversely, we were also able to simulate topologies that cluster closely in tree space (Fig. 5d), as observed in experimental Ig-Seq data (Fig. 5b). We see that all five methods predict topologies that are explored in the range of BEAST's MCMC algorithm and are in similar tree space.

Thus, our results suggest that the topological diversity observed in experimental Ig-Seq data can be recreated with AbSim.

3.4 The simulated mutation rate and tree height can be inferred using BEAST with high accuracy

In addition to topology, there are several other key measurements related to antibody sequence evolution. For example, the mutation rate and duration of B-cell lineage evolution (tree height). We tested whether these two parameters could be inferred with high accuracy across a wide range of sequence sampling schemes. We performed simulations using AbSim and subsamples of sequences were selected

at defined sampling points during the course of evolution. This is similar to an experimental layout where a subpopulation of B-cells (obtained from blood) was sequentially drawn from a mouse (at given time points before isolating B-cells from lymphoid organs following mouse sacrifice). Given that only a fraction of blood can be sampled at each time step, the repertoire present in blood will not cover the entirety of the repertoire found at terminal time points. Thus, the influence of undersampling was assessed using simulations.

The parameter estimations for mutation rate per simulated time step and tree height were inferred across four different sampling methods (Fig. 6). Ten simulations were run for each sampling method, with the simulation ending after 90 time steps. The mutation rate was set to 3.143×10^{-3} [$\beta^*\mu+\gamma$] based on previous estimates of the rate of SHM (McKean *et al.*, 1984; Odegard and Schatz, 2006). BEAST could predict this mutation rate with up to 94% accuracy [inferred mutation rate/true mutation rate] (Fig. 6a). BEAST slightly underestimated the mutation rate for all four sampling cases, ranging from 89 to 94% (Fig. 6a). The simulated tree height was set to 90 and was predicted with up to 97% accuracy [inferred tree height/true tree height] (Fig. 6b). Similarly, undersampling played nearly no role in altering the inferred tree height (Fig. 6b). Thus, the ability of BEAST to accurately predict both the mutation rate (in meaningful units) and tree height offers a clear advantage in antibody repertoire phylogenetic analysis.

4 Discussion

Given that the analysis of antibody repertoire evolution is still in its infancy, there is no clear consensus as to which phylogenetic method is optimal for lineage inference (Greiff *et al.*, 2015b; Yaari *et al.*, 2013). Accurate phylogenetic analysis is important to further understand the context in which antibody repertoires change over time. The topology describing antibody evolution gives insight into intermediate sequences, which can be used for a variety of applications, such as vaccine development and immunotherapy (Jardine *et al.*, 2016; Wu *et al.*, 2015). Current phylogenetic methods rely on assumptions that may impact conclusions of antibody evolution more so than with species. One prominent example is the assumption that each site mutates independently of the neighboring nucleotides, which does not hold for antibody evolution (Yaari *et al.*, 2013). Additionally, antibodies evolve on dramatically smaller time scales than species. These two factors likely decrease both the clade accuracy and the clade overlap between methods, thereby, bias conclusions drawn from antibody evolution studies.

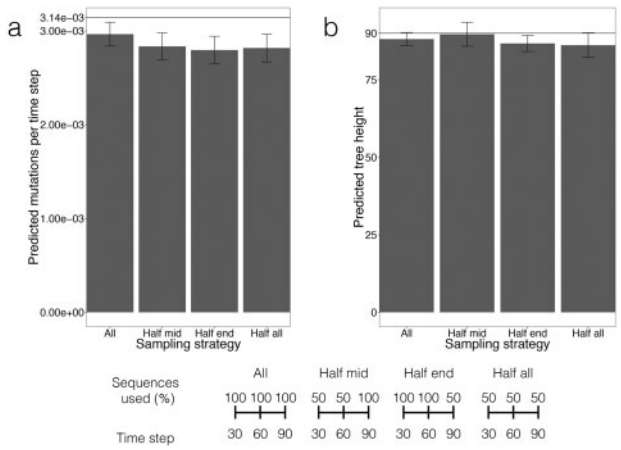

Fig. 6. The simulated mutation rate and tree height can be inferred using BEAST. Simulations were run for 90 time steps with a mutation rate of 3.143×10^{-3} mutations per nucleotide per time step. Four different sampling techniques were used to examine the influence of sampling on the predicted mutation rate (**a**) and the predicted tree height (**b**) using BEAST. There were two sampling points throughout the simulation, one at time step 30, and one at time step 60. All bar plots depict mean ± s.e.m. Ten simulations per sampling scheme were run

By leveraging a time-resolved antibody repertoire simulation framework (AbSim), we found when aggregating individual V-D-J recombination events that the clade prediction accuracy of all phylogenetic reconstruction methods decreased substantially from >88% to <50% (Fig. 2). Grouping CDR3 sequences to filter and reduce V-D-J recombination events is one strategy to overcome this problem, but has seldom been used thus far (Stern *et al.*, 2014). Methods with more statistical rigor have recently been introduced (Ralph and Matsen, 2016) and should be evaluated as well to determine how they are influenced by V-D-J recombination and SHM-based evolution. In general, under the parameter range tested, we identified both BEAST and LD as the two methods that across all parameter settings performed better or equal to the other three methods tested (Figs 2 and 3). The benefits of BEAST and LD were particularly evident when the baseline mutation rate was set to 0, which represents when there are few mutations in between branching events (Fig. 3). Biologically, this could be the case when a response is dominated by only a few clones (oligoclonal response), and the mutations cause the predecessor cell to no longer be in the population after SHM occurs. Interestingly, when looking at the tree space for both BEAST and LD, the topologies of BEAST's MCC trees were consistently closer to the correct, simulated tree (Fig. 5e). This suggests that while the clade accuracy between the two methods may be similar, BEAST may be better at predicting other measures of topology (Fig. 5). Additionally, we showed that BEAST could also incorporate time-resolved information to accurately predict the mutation rate and tree height for simulated lineage trees (Fig. 6).

More generally, an advantage of our simulation framework is that all parameters as well as underlying models may be updated as novel insight on the mechanism of antibody V-D-J recombination is gained. For the sake of simplicity, we have made several assumptions regarding both V-D-J recombination and SHM events that may not represent the underlying immunobiology at the most precise level (Elhanati *et al.*, 2015). The most notable simplifications is that all V-D-J recombination events were simulated to allow for an arbitrary number of nucleotides inserted or deleted per event, regardless of the specific germline element used. Furthermore, the simulations performed here only represent a fraction of AbSim's potential parameter space as, for example, all simulations presented were performed using only one set of germline genes. Thus, these simplifications could influence the phylogenetic inference results when SHM incorporates motifs or different rates for transitions and transversions. Subsequent versions of AbSim will incorporate specific nucleotide information for sequence diversification during both V-D-J recombination and SHM, which will be made possible as increased biological insight emerges from future Ig-Seq studies.

Most of our analyses were performed using the Poisson distributed method of mutations due to the associated ease of calculating substitution rates. Trees produced using LD make a similar assumption that the mutational properties of each nucleotide are identical. This could explain why LD performed well when analyzing clade prediction accuracy across all simulations using the Poisson distributed mutations. We observed that MP often performed much worse than the other methods, regardless of which model of SHM was

simulated. Further investigation is warranted to identify if this decrease in clade prediction accuracy is due to the long-branch attraction problem, which is often mentioned as a downfall of MP (Felsenstein, 1978). Furthermore, we occasionally observe MP inferring trees with multifurcation branching-events (non-binary), which may also lead to spurious clade prediction. Uncovering which antibody-relevant parameters are responsible for these multifurcating events is a potential future study involving AbSim. ML and BEAST predicted the simulated clades with highest accuracy when the SHM model incorporated a different mutation rate for transitions and transversions (Fig. 3b). This is due to their parametric nature that allows for nucleotide specific mutation probabilities. A more detailed investigation into the specific clades differing between these two methods could be of future interest.

A further limitation of our study is that we exclusively compared phylogenetic characteristics among either simulated or experimental antibody repertoires, without investigating sequence similarity *between* them. Nevertheless, the AbSim framework allows the user to modify the weights of each of these mutational models in order to mirror their own experimental data. Importantly, our primary focus was to reproduce the discrepancies between phylogenetic methods using parameters based on experimental knowledge independent of the immune response arising from the LCMV infection of a single mouse.

Evolutionary models specific for antibody sequences have recently been developed (Hoehn *et al.*, 2017; Mirsky *et al.*, 2014) however they have yet to be incorporated into a Bayesian framework (BEAST). The general development of other BEAST-based methods would allow for a more complete picture of the evolutionary process that occurs in B-cells. For example, tools are already available that predict the root sequences from evolutionary trees (Bouckaert *et al.*, 2014). This could be useful when the exact germline gene sequences are unknown, as in the case of species without a completely sequenced and annotated immunoglobulin locus (Collins *et al.*, 2015; Greiff *et al.*, 2015b).

In summary, by exploiting a time-resolved antibody repertoire simulation framework, we determined precise guidelines under which current phylogenetic Ig-Seq methods can retrace and reconstruct antibody evolution with high accuracy. These guidelines are the following: (i) independent V-D-J events need to be separated (or filtered) to increase clade accuracy, (ii) parameter estimation of mutation rate and tree height can be inferred using a Bayesian framework and (iii) BEAST and LD should be used when the mutation rate between speciation events is low. We believe that the vast parameter-repertoire space simulated by AbSim will enable future rigorous testing and validation of a wide family of repertoire-focused bioinformatics tools, even those outside of phylogenetic analysis such as V-D-J annotation and error correction (Bolotin *et al.*, 2015; Callahan *et al.*, 2016; Khan *et al.*, 2016). Accurate estimation of parameters that describe evolution is crucial for comparing the evolution between pathogen and host immune responses. This information may help elucidate how escape variants avoid neutralization by the host antibody response (Liao *et al.*, 2013).

Acknowledgements

We thank Dr. Christian Beisel, Ina Nissen, Elodie Burcklen and Manuel Kohler from the Genomics Facility Basel at the ETH Zürich Department of Biosystems Science and Engineering for technical assistance with high-throughput sequencing.

Funding

This work has been supported by SystemsX.ch AntibodyX RTD project (to STR and AO), Swiss National Science Foundation (Project no: 31003A_143869, 31003A_170110 to STR), Swiss Vaccine Research Institute (to STR and AO). The professorship of STR is made possible by the generous endowment of the S. Leslie Misrock Foundation.

Conflict of Interest: none declared.

References

Andrews, S.F. *et al.* (2015) High preexisting serological antibody levels correlate with diversification of the influenza vaccine response. *J. Virol* **89**, 63308–63317.

Barak, M. *et al.* (2008) IgTree©: Creating Immunoglobulin variable region gene lineage trees. *J. Immunol. Methods*, **338**, 67–74.

Bolotin, D.A. *et al.* (2015) MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods*, **12**, 380–381.

Bouckaert, R. *et al.* (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.*, **10**, e1003537.

Callahan, B.J. *et al.* (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods*, **13**, 581–583.

Collins, A.M. *et al.* (2015) The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. *Philos. Trans. R. Soc. B*, **370**, 20140236.

Cui, A. *et al.* (2016) A model of somatic hypermutation targeting in mice based on high-throughput Ig sequencing data. *J. Immunol.*, **197**, 3566–3574.

Drummond, A.J. (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.*, **22**, 1185–1192.

Drummond, A.J. *et al.* (2006) Relaxed phylogenetics and dating with confidence. *PLoS Biol.*, **4**, e88.

Elhanati, Y. *et al.* (2015) Inferring processes underlying B-cell repertoire diversity. *Phil. Trans. R. Soc. B*, **370**, 20140243.

Ellebedy, A.H. *et al.* (2016) Defining antigen-specific plasmablast and memory B cell subsets in human blood after viral infection or vaccination. *Nat. Immunol.*, **17**, 1226–1234.

Feeney, A.J. (1990) Lack of N regions in fetal and neonatal mouse immunoglobulin V-D-J junctional sequences. *J. Exp. Med.*, **172**, 1377–1390.

Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.*, **27**, 401.

Felsenstein, J. (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.

Georgiou, G. *et al.* (2014) The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nat. Biotechnol.*, **32**, 158–168.

Giudicelli, V. *et al.* (2004) IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucleic Acids Res.*, **32**, W435–W440.

Glanville, J. *et al.* (2011) Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl. Acad. Sci. USA*, **108**, 20066–20071.

Green, M.R. *et al.* (2013) Hierarchy in somatic mutations arising during genomic evolution and progression of follicular lymphoma. *Blood*, **121**, 1604–1611.

Greiff, V. *et al.* (2015a) A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med.*, **7**, 49.

Greiff, V. *et al.* (2015b) Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol.*, **36**, 738–749.

Greiff, V. *et al.* (2014) Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC Immunol.*, **15**, 40.

Greiff, V. *et al.* (2017) Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B-cell development. *Cell Rep*, **19**, 1467–1478.

Hoehn, K.B. *et al.* (2017) A phylogenetic codon substitution model for antibody lineages. *Genetics*, **206**, 417–427.

Jackson, K.J.L. *et al.* (2014) Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell Host Microbe*, **16**, 105–114.

Janeway, C.A. and Murphy, K. (2011) *Janeway's Immunobiology 8th Revised Edition*. Taylor & Francis, New York, NY.

Jardine, J.G. *et al.* (2016) HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen. *Science*, **351**, 1458–1463.

Johnston, C.M. *et al.* (2006) Complete sequence assembly and characterization of the C57BL/6 mouse Ig heavy chain V region. *J. Immunol.*, **176**, 4221–4234.

Jombart, T. *et al.* (2017) TREESPACE: Statistical exploration of landscapes of phylogenetic trees. *Mol. Ecol. Resour.*, https://doi.org/10.1111/1755-0998.12676.

Kendall, M. and Colijn, C. (2016) Mapping phylogenetic trees to reveal distinct patterns of evolution. *Mol. Biol. Evol.*, **33**, 2735–2743.

Khan, T.A. *et al.* (2016) Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci. Adv.*, **2**, e1501371.

Krebber, A. *et al.* (1997) Reliable cloning of functional antibody variable domains from hybridomas and spleen cell repertoires employing a reengineered phage display system. *J. Immunol. Methods*, **201**, 35–55.

Liao, H.-X. *et al.* (2013) Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature*, **496**, 496–476.

Loo, MPJ. v d. (2014) The stringdist package for approximate string matching. *R. J.*, **6**, 111–122.

Manz, R.A. *et al.* (1997) Lifetime of plasma cells in the bone marrow. *Nature*, **388**, 133–134.

McKean, D. *et al.* (1984) Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. *Proc. Natl. Acad. Sci, USA*, **81**, 3180–3184.

Menzel, U. *et al.* (2014) Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. *PLoS ONE*, **9**, e96727.

Mirsky, A. *et al.* (2014) Antibody-specific model of amino acid substitution for immunological inferences from alignments of antibody sequences. *Mol. Biol. Evol.*, msu340.

Mora, T. *et al.* (2010) Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. USA*, **107**, 5405–5410.

Mroczek, E.S. *et al.* (2014) Differences in the composition of the human antibody repertoire by B cell subsets in the blood. *B Cell Biol.*, **5**, 96.

Muramatsu, M. *et al.* (2000) Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell*, **102**, 553–563.

Odegard, V.H. and Schatz, D.G. (2006) Targeting of somatic hypermutation. *Nat. Rev. Immunol.*, **6**, 573–583.

Paradis, E. *et al.* (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**, 289–290.

Ralph, D.K. and Matsen, F.A.IV (2016) Likelihood-based inference of B cell clonal families. *PLOS Comput. Biol.*, **12**, e1005086.

Revell, L.J. (2012) phytools: an R package for phylogenetic comparative biology (and other things): phytools: R package. *Methods Ecol. Evol.*, **3**, 217–223.

Revell, L.J. and Chamberlain, S.A. (2014) Rphylip: an R interface for PHYLIP. *Methods Ecol. Evol.*, **5**, 976–981.

Richter, K. and Oxenius, A. (2013) Non-neutralizing antibodies protect from chronic LCMV infection independently of activating FcγR or complement: Immunity to infection. *Eur. J. Immunol.*, **43**, 2349–2360.

Saada, R. *et al.* (2007) Models for antigen receptor gene rearrangement: CDR3 length. *Immunol. Cell Biol.*, **85**, 323–332.

Safonova, Y. *et al.* (2015) IgRepertoireConstructor: a novel algorithm for antibody repertoire construction and immunoproteogenomics analysis. *Bioinformatics*, **31**, i53–i61.

Schliep, K.P. (2011) phangorn: phylogenetic analysis in R. *Bioinformatics*, **27**, 592–593.

Seifert, M. and Kuppers, R. (2009) Molecular footprints of a germinal center derivation of human IgM+(IgD+)CD27+ B cells and the dynamics of memory B cell generation. *J Exp Med.*, **206**, 2659–2669.

Sievers, F. *et al.* (2014) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539–539.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Stern, J.N.H. *et al.* (2014) B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci. Transl. Med*, **6**, 248ra107–248ra107.

Tipton, C.M. *et al.* (2015) Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. *Nat. Immunol.*, **16**, 755–765.

Tonegawa, S. (1983) Somatic generation of antibody diversity. *Nature*, **302**, 575–581.

Wang, C. *et al.* (2015) B-cell repertoire responses to varicella-zoster vaccination in human identical twins. *Proc. Natl. Acad. Sci. USA*, **112**, 500–505.

Weinstein, J.A. *et al.* (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science*, **324**, 807–810.

Wu, X. *et al.* (2015) Maturation and diversity of the VRC01-antibody lineage over 15 years of chronic HIV-1 infection. *Cell*, **161**, 470–485.

Xu, J.L. and Davis, M.M. (2000) Diversity in the CDR3 region of VH is sufficient for most antibody specificities. *Immunity*, **13**, 37–45.

Yaari, G. *et al.* (2013) Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput Immunoglobulin sequencing data. *Front. B Cell Biol.*, **4**, 358.

Yaari, G. and Kleinstein, S.H. (2015) Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med.*, **7**, 121.

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.