

1
2
3
4 Phylogenetic analysis of migration, differentiation, and class switching in B cells
5

6
7 Kenneth B. Hoehn¹, Oliver G. Pybus² and Steven H. Kleinstein^{1,3,4,*}
8

9
10 ¹Department of Pathology, Yale University School of Medicine, New Haven, CT 06520, USA.
11

12 ²Department of Zoology, University of Oxford, Oxford OX1 3PS, UK.
13

14 ³Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New
15 Haven, CT 06511, USA

16
17 ⁴Department of Immunobiology, Yale School of Medicine, New Haven, CT 06520
18

19
20 * Corresponding author
21 Email: steven.kleinstein@yale.edu

22 **Abstract**

23

24 B cells undergo rapid mutation and selection for antibody binding affinity when producing
25 antibodies capable of neutralizing pathogens. This evolutionary process can be intermixed with
26 migration between tissues, differentiation between cellular subsets, and switching between
27 functional isotypes. B cell receptor (BCR) sequence data has the potential to elucidate important
28 information about these processes. However, there is currently no robust, generalizable
29 framework for making such inferences from BCR sequence data. To address this, we develop
30 three parsimony-based summary statistics to characterize migration, differentiation, and isotype
31 switching along B cell phylogenetic trees. We use simulations to demonstrate the effectiveness
32 of this approach. We then use this framework to infer patterns of cellular differentiation and
33 isotype switching from high throughput BCR sequence datasets obtained from patients in a study
34 of HIV infection and a study of food allergy. These methods are implemented in the R package
35 *dowser*, available at <https://bitbucket.org/kleinsteink/dowser>.

36

37 **Author summary**

38

39 B cells produce high affinity antibodies through an evolutionary process of mutation and
40 selection during adaptive immune responses. Migration between tissues, differentiation to
41 cellular subtypes, and switching between different antibody isotypes can be important factors in
42 shaping the role B cells play in response to infection, autoimmune disease, and allergies. B cell
43 receptor (BCR) sequence data has the potential to elucidate important information about these
44 processes. However, there is currently no robust, generalizable framework for making such
45 inferences from BCR sequence data. Here, we develop three parsimony-based summary statistics
46 to characterize migration, differentiation, and isotype switching along B cell phylogenetic trees.
47 Using simulations, we confirm the effectiveness of our approach, as well as identify some
48 caveats. We further use these summary statistics to investigate patterns of cellular differentiation
49 in three HIV patients, and patterns of isotype switching in an individual with food allergies. Our
50 methods are released in the R package *dowser*: <https://bitbucket.org/kleinsteink/dowser>.

51

52 **Introduction**

53

54 The adaptive immune system in humans depends on B cells to produce antibodies capable of
55 neutralizing a wide array of pathogens. Antibody structures are initially expressed as B cell
56 receptors (BCRs) on the surfaces of B cells. BCRs are generated through random V(D)J
57 recombination and then subjected to repeated rounds of somatic hypermutation (SHM), cell
58 proliferation, and selection for antigen binding [1]. This evolutionary process, called affinity
59 maturation, creates many lineages of B cells that each descend from a single naïve progenitor
60 cell. Cells within a clonal lineage differ predominately by point mutations. The genetic variation
61 within these clonal lineages has been long investigated using phylogenetic methods [2]. When
62 obtained through high throughput sequencing, BCR sequences have shown promise in
63 elucidating information about the adaptive immune response in humans, such as the sequence of
64 mutations that occur during antibody co-evolution with HIV [3], and the process of mutation and
65 selection during affinity maturation generally [4]. Other important biological processes may
66 occur as BCR sequences evolve, such as B-cell migration between tissues [5], differentiation into
67 cellular subsets [6], and antibody isotype class switching [7]. If these processes co-occur with
68 SHM, then in principle they can be investigated and inferred using phylogenetic techniques.

69

70 Migration and cellular differentiation in B cells can be viewed as analogous to geographic spread
71 of rapidly evolving viruses, the study of which – viral phylogeography – has advanced both in
72 theory and application in the past decade (e.g. Lemey *et al.* 2009). For example, phylogeographic
73 methods have been used to determine the origin of the HIV pandemic [9], factors influencing the
74 recent Ebola epidemic [10,11], and the epidemic spread of Zika virus [12,13]. Modern

75 phylogeographic analyses typically model phylogenetic sequence evolution [14], and changes in
76 geographic location within a unified framework [15,16]. Successfully developing a
77 phylogeographic framework for B cell lineages would enable the testing of new hypotheses
78 regarding the nature of evolution during affinity maturation.

79

80 There are serious challenges that must be addressed before using modern phylogeographic
81 methods on B cell repertoire datasets. Such techniques typically rely on molecular clock trees,
82 whose branch lengths represent elapsed time between nodes [15]. Accurately modelling
83 sequence change through time requires either data sampled at multiple time points, or prior
84 information about expected rate of sequence evolution. These are not frequently available for B
85 cell lineages. Data samples, particularly biopsies, are often only taken at a single time point
86 [5,17], and the variation of B cell mutation rate over time is largely unknown and likely
87 dependent on cell subset. Even using a Markov model to describe state changes along B cell
88 molecular phylogenies is not straightforward: B cell lineage trees frequently contain identical
89 sequences with different states. This results in state changes across zero-length branches, which
90 are not able to be fit within a Markov model framework. Further, modern phylogeographic
91 techniques often rely on Markov chain Monte Carlo sampling, which makes them
92 computationally intensive and impractical to apply to thousands of sequences. Unfortunately, B
93 cell bulk repertoires often contain millions of sequences, and individual lineages sometimes
94 contain thousands of unique sequences.

95

96 We propose that hypotheses about B cell migration and differentiation may be usefully
97 investigated using heuristic summary statistics that characterize the distribution of trait values

98 along phylogenetic trees. Indeed, such heuristic approaches which do not depend on branch
99 length have historically been a popular means of testing hypotheses about migration between
100 populations [18–20]. While tree based summary statistics have been previously used to assess B
101 cell migration [5], differentiation [6], and isotype switching [7], these approaches have not been
102 tested through simulations and their general accuracy is unclear. To address this methodological
103 gap, we develop a set of maximum parsimony-based statistics that summarize the relative
104 distribution of B cell states along lineage trees within repertoires and introduce a framework for
105 assessing the significance of their difference from randomized trees. We demonstrate through
106 simulations that these tests relate intuitively to different regimes of migration and differentiation.
107 To demonstrate its utility, we use this framework to test hypotheses regarding differentiation of
108 cell types in HIV infection, and sequential class switching to IgE and IgG4. We introduce a
109 statistically principled and scalable means of analyzing the evolution of discrete traits in B cell
110 repertoires. We release these methods in the R package *dowser*.

111

112

113 **Methods**

114

115 *Predicting states of internal tree nodes*

116

117 The goal of the discrete trait analysis framework presented here is to characterize the distribution
118 of predicted trait values along B cell lineage trees. Given an alignment of sequences inferred to
119 descend from the same naïve ancestor (i.e. the same clonal family), lineage tree topologies and
120 branch lengths were estimated using maximum parsimony using *dnapars* v3.967 [21].

121 Importantly, the statistics presented here are not limited to tree topologies inferred through
122 maximum parsimony.

123

124 Maximum parsimony is also used to infer the discrete character states (e.g. cell subtype, isotype,
125 tissue) of internal nodes, given a tree in which each tip is associated with a given character state.
126 Nodes with different states from their immediate ancestors are counted as state changes. More
127 specifically, internal node states were reconstructed using the Sankoff dynamic programming
128 maximum parsimony algorithm [22], which, given a weight matrix for each type of state change,
129 determines the minimum number of state changes that must be made along the tree given the
130 states at the tips. The backtrace step of this algorithm can be used to determine a set of most
131 parsimonious internal node states. Often there are multiple such maximum parsimony sets. To
132 represent state changes across ambiguous internal node sets, trajectories with equal parsimony
133 were randomly chosen in the backtrace step of the Sankoff algorithm, beginning at the root of the
134 tree and moving towards the tips. This process is performed 100 times for each tree, and the
135 mean of each type of state change was reported.

136

137 Strictly bifurcating B cell lineage trees frequently have clusters of nodes separated by zero-
138 length branches (soft polytomies), which represent a high degree of uncertainty in tree topology.
139 This uncertainty in the order of bifurcating nodes can result in a potentially large number of
140 uninformative state changes along the polytomy. Multiple steps were taken to minimize the
141 effects of random polytomy resolution (**Supplemental File S1**). Briefly, nodes within each
142 polytomy were first re-ordered to minimize the number of state changes along the tree. To
143 represent the uncertainty in the order of state changes, nodes within each polytomy were grouped
144 together into separate subtrees according to their predicted state. These state-specific subtrees
145 were then joined together in a balanced manner, ensuring that state changes could occur in any
146 direction among the states contained within the polytomy (**Supplemental File S1**).
147

148

Testing trait-phylogeny association

149

150 Analysis begins with a B-cell lineage tree topology with discrete character states (trait values)
151 associated with each tip, and internal node states reconstructed through maximum parsimony.
152 The goal of our discrete trait analysis framework is to determine how the distribution of discrete
153 character states along the internal nodes of a tree differs from its expectation if traits are
154 randomly distributed among the tips. The statistics introduced herein are shown graphically in
155 **Fig 1**. More formally, if there are m possible discrete character states, and o_{ij} is the number of
156 state changes from type i to type j , the three statistics investigated are defined as:
157

158

$$PS = \sum_i^m \sum_{j,j \neq i}^m o_{ij} \quad (1)$$

159

160 $SC_{ij} = o_{ij}$ (2)

161

162 $SP_{ij} = \frac{o_{ij}}{\sum_i^m \sum_{j,j \neq i}^m o_{ij}}$ (3)

163

164 The *PS* (parsimony score) statistic is the total number of state changes along a tree. The *SC* (switch count) statistic from *i* to *j* is the number of state changes from state *i* to *j*. The *SP* (switch proportion) statistic from *i* to *j* is the proportion of state changes from state *i* to *j*.

167

168 We calculate the significance of these three statistics using a permutation test. This is done by
169 randomizing traits at the tips of the lineage tree, re-calculating each statistic on the permuted
170 tree, and repeating for a specified number of replicates. For each replicate, we calculate δ , which
171 is the difference between the statistic calculated on the observed tree and the same statistic
172 calculated on the permuted tree. If mean $\delta > 0$ (hereafter mean δ is indicated by $\bar{\delta}$), this indicates
173 the statistic is on average higher in observed trees than in permuted trees. For a one-tailed test,
174 we calculate the *p* value that $\bar{\delta} > 0$ as the proportion of replicates in which $\delta \leq 0$. Similarly, we
175 calculate the *p* value that $\bar{\delta} < 0$ as the proportion of replicates in which $\delta \geq 0$. For a two-tailed
176 test, we calculate the *p* value that $\bar{\delta} > 0$ as the proportion of replicates in which $\delta < 0$, plus half of
177 the replicates in which $\delta = 0$. We refer to the calculation of these *p* values for the statistics in **Eq.**
178 **1-3** as the *PS* test, *SC* test, and *SP* test, respectively.

179

180 These three statistical tests capture different aspects of how the distribution of characters
181 observed along a lineage tree differs from random association between tree topology and trait
182 values. The *PS* test determines the extent to which trait values are clustered together within the
183 tree. A significantly low *PS* statistic (i.e. $\delta < 0$, $p < 0.05$) indicates identical trait values are more
184 closely clustered together within the tree than expected from random association between tree
185 topology and trait values. By contrast, a significantly high *PS* statistic indicates identical trait
186 values are less clustered together than expected by chance. Variations of this test were previously
187 developed in [18] and applied in [23] to study spread of influenza.

188

189 While the *PS* test only determines a general association between trait values and tree topology,
190 the *SC* and *SP* tests are both aimed at determining whether a particular trait value is more
191 ancestral to another in the tree. A significantly high *SC* statistic (Eq. 2) from state i to state j
192 indicates a greater number of switches from state i to state j than expected from random
193 association between tree topology and trait values. The *SC* test was used by [19] in the context of
194 virus phylogeography, and by [6] to decompose the phylogenetic relationships among B cell
195 subtypes within HIV infection. The *SC* test, however, is not purely a metric of whether state i
196 tends to be more immediately ancestral to state j than expected. This is because trees with
197 randomized tip states often have more state changes events in general, hence δ values tend to be
198 negative even when there is no polarity in the ancestor/descendant relationship among type i and
199 j (Fig 1B).

200

201 To normalize for changes in the total number of state changes between observed and permuted
202 trees, we introduce the *SP* test. A significantly high *SP* statistic (Eq. 2) from state i to state j

203 indicates a greater proportion of switches from state i to state j than expected from a random
204 distribution of trait values at the tips. In contrast to the SC test, a significant association between
205 the tree and trait may exist, but only associations in which one trait is more often ancestral to the
206 other will give rise to a significantly high or low SP statistic (**Fig 1C**). Further, the denominator
207 of the SP statistic can be altered to test other hypotheses. For instance, to test whether a greater
208 proportion of state changes to state j come immediately from state i than expected by chance, one
209 can restrict the analysis to consider only state changes towards j .

210

211 *Accounting for uncertainty in tree topology*

212

213 To account for uncertainty in tree topology, we bootstrap multiple sequence alignments within
214 each clone [24]. This is performed by random sampling with replacement of the columns of a
215 multiple sequence alignment. Lineage tree topology and branch length estimation then proceeds
216 as before. Test statistics are calculated for each bootstrap replicate tree, and then for a single
217 permutation of the traits at that tree's tips. Calculations for δ values proceed as before, but with
218 each o_{ij} value indexed for each replicate. This procedure is very similar to that proposed in [25]
219 and can in principle be extended to other sets of tree topologies, such the posterior distribution of
220 tree topologies generated by MCMC sampling under Bayesian phylogenetic inference.

221

222 *From trees to repertoires*

223

224 B cell repertoire datasets often consist of hundreds or thousands of B cell lineage trees. Often
225 hypotheses do not concern individual lineages, but instead the behavior of the collection of B cell

226 lineages as a group. To characterize multiple B cell lineages, the observed and permuted
227 summary statistics are summed across all lineages for each bootstrap replicate. Additionally,
228 traits may be permuted among trees, which may increase statistical power and detect nonrandom
229 association among trait types within trees.

230

231 *Simulations*

232

233 We tested the performance of the three proposed statistics using simulations based on B cell
234 lineage trees estimated from an empirical dataset. Sequences were obtained from peripheral
235 blood samples taken from one subject at ten time points, from eight days before to 28 days after
236 influenza vaccination [26] (subject 420IV). Sequence preprocessing and clonal clustering are
237 described in [4]. Sequences were down-sampled by 50%, and only clones with >10 unique
238 sequences were retained. A total of 399 clones containing 11 to 370 (mean=26.2) unique
239 sequences remained. Tree topologies and branch lengths were estimated for each clone using
240 *dnapars* v3.967 [21] via the R package Alakazam v0.3.0 [27]. We simulated state changing
241 down each tree using a Markov model parametrized by initial frequencies π for each state,
242 relative rate parameters r_{ij} for each pair of possible states i and j , and r , the average rate of state
243 changes per mutation per site. The mean value of this rate matrix was calculated as the sum of
244 the diagonal elements weighted by their initial state frequencies. All values of the matrix were
245 then divided by this mean and multiplied by r . This calibration was performed so that r^*l state
246 change events are expected to occur across a branch of length l mutations/site. For each tree, the
247 state at the germline node was randomly drawn based on each state's π value. For each node after
248 the germline, the rate matrix is multiplied by the node's ancestral branch length and

249 exponentiated to give the probability of each state at the descendant node, given the state at the
250 node's immediate ancestor. The state at the descendant node is randomly chosen based on these
251 probabilities. This process begins at the germline node and continues down the tree until each tip
252 node has a state. Because each tip corresponds to a sequence, this forms a dataset of sequences
253 paired with simulated discrete characters. Internal node states are not included in the final
254 simulated dataset used for analysis. Simulations were performed with two state models (*A* and *B*)
255 that explored a large parameter space ($\pi_a = 0.5, 1$; $r_{ab} = 0.1, 1, 10$; $r = 10, 25, 50, 100, 1000$), and
256 four state models (*A, B, C, D*) that explored more complex patterns of state change at low overall
257 rate ($r = 10$). Twenty simulation repetitions were performed for each parameter combination.
258 Statistical tests were performed as described in **Methods**; however, to improve computational
259 efficiency simulation analyses did not use bootstrapped multiple sequence alignments, and
260 instead performed 100 permutations on a fixed maximum parsimony tree for each clone. Only
261 clones with more than one state type were analyzed.

262

263 *Empirical datasets*

264

265 We demonstrate the utility of the proposed discrete trait framework by analyzing two empirical
266 datasets. The first was aimed and understanding B cell differentiation during HIV infection, and
267 consists of BCR mRNA sequences taken from sorted populations of unswitched memory B cell
268 (MBC), CD19^{hi} MBC, CD19^{lo} MBC, and germinal center B cells (GCBC) from three HIV
269 viremic subjects (subject 1-3; [6]) Each dataset was subsampled to a maximum of 50,000 total
270 sequences, and only clones with more than 10 sequences were retained. Unique sequences
271 associated with more than one cell type were kept distinct. This resulted in 128, 197, and 174

272 clones with a mean of 53, 38.6, and 31.8 unique sequences per clone, for subjects 1-3
273 respectively. State changes across all lineages for each subject were calculated over 100
274 bootstrap replicates.

275

276 The second dataset was aimed at understanding isotype switching patterns in human children,
277 and consists of BCR mRNA sequences obtained from peripheral blood samples taken from a
278 human child each year from age 1 to 3 years old [28]. Preprocessing, including grouping of
279 sequences into clonal clusters, is detailed in **Supplemental File S2**. Only clones with at least 4
280 unique sequences and more than one isotype were retained. Unique sequences associated with
281 more than one isotype were kept distinct so each sequence was associated with one isotype. This
282 resulted in 768 clones with a mean of 9.3 unique sequences each. State changes across all
283 lineages were calculated over 100 bootstrap replicates.

284

285 **Results**

286

287 We outline three parsimony-based summary statistics to characterize the distribution of trait
288 values along B cell lineage trees (**Fig 1**). The significance of these statistics can be tested by
289 comparing observed values within the set of trees that comprise a repertoire to those obtained
290 from permuting trait values at the tree's tips. The first statistic, the parsimony score (*PS*), is the
291 total number of trait value state changes that occurred along a lineage tree. A *PS* test with $\delta < 0$
292 and $p < 0.05$ (*i.e.* a significantly low *PS* statistic) indicates the trait values cluster together in the
293 observed trees more often than expected by chance (**Fig 1**). We propose two other statistics
294 aimed at determining whether one state is more frequently the immediate ancestor to another
295 state than expected by chance. The switch count (*SC*) from state i to j is the number of state
296 changes that occurred from i to j [19], while the switch proportion (*SP*) from state i to j is the
297 proportion of state changes that occurred from i to j . An *SC* or *SP* test from i to j with $\delta > 0$ and p
298 < 0.05 indicates trait value i was more frequently immediately ancestral to state j than expected
299 by chance. We expect the *SP* test to be more sensitive to this relationship than the *SC* test
300 because it accounts for the increased number of state changes expected in randomized trees (**Fig**
301 **1B-C**). Similarly, an *SP* test from i to j with $\delta < 0$ and $p < 0.05$ indicates trait value i was less
302 ancestral to state j than expected (**Fig 1B**). All three of these tests may be used to characterize
303 individual lineages or entire B cell repertoires; in this paper we will focus exclusively on
304 repertoires.

305

306 *Differentiating state change patterns with two states*

307

308 We used simulations to test the performance of our proposed tests. We model B cell
309 migration/differentiation using a Markov model with two states, *A* and *B*, and empirically-
310 derived lineage tree topologies (**Methods**). Briefly, the pattern of state changes along a tree was
311 determined by the probability that the state at the root was *A* ($\pi_a = 0.5, 1$; $\pi_b = 1 - \pi_a$), the
312 average rate of state change ($r = 10, 25, 50, 100, 1000$ changes/mutation/site), and the relative
313 rate of change from *A* to *B* ($r_{ab} = 0.1, 1, 10$; $r_{ba} = 1/r_{ab}$). These parameters represent a range of
314 slow, fast, biased, and unbiased state change patterns along a B cell lineage. Each simulation
315 resulted in a dataset of BCR sequences, each associated with a single trait value (*A* or *B*)
316 resulting from the simulation process. The goal of our simulation analysis is to determine if the
317 summary statistics provide useful information about the mode and tempo of trait evolution.

318

319 We ran 20 simulation repetitions for each parameter combination, and tested the significance of
320 each of the proposed statistics to assess their statistical power. Our simulations are designed to
321 generate trees whose tip-states are more clustered together than if the tips states are randomly
322 distributed across the tree tips. Consistent with this expectation, 320/320 simulation repetitions in
323 which $r < 1000$ (*i.e.* overall rate of state change < 1000 changes/mutation/site) showed a
324 significantly low *PS* statistic regardless of other parameters ($\delta < 0$; one-tailed $p < 0.05$;
325 **Supplemental File S3**). This confirms the *PS* test's usability for detecting nonrandom
326 association between tree topology and trait values. However, at $r = 1000$, only 3/80 repetitions
327 showed a significantly low *PS* statistic (**Supplemental File S3**), indicating this relationship is
328 difficult to detect at high rates of state change.

329

330 We used the same simulations to test whether the *SC* statistic was capable of detecting the
331 direction of state changes in B cell repertoires. A total of 300 simulation repetitions were
332 performed using parameters expected to give biased (directed) state changes; namely, with
333 lineages always beginning in *A* ($\pi_a = 1$) and/or highly biased rates of state change from *A* to *B*
334 ($r_{ab} = 10$). Surprisingly, only 3/300 of these simulations showed a significantly high *SC* from *A*
335 to *B* ($\delta > 0$; one-tailed $p < 0.05$; **Supplemental File S4**). By contrast, 186/300 showed a
336 significantly low *SC* from *A* to *B* ($\delta < 0$; $p < 0.05$). This indicates that significantly high *SC*
337 statistics are highly conservative, while significantly low *SC* statistics are primarily driven by
338 overall phylogenetic association with a trait. This issue is likely exacerbated as dataset size
339 grows, hence the *SC* test is likely still useful for single lineages [19,20] or for detecting very
340 strong trends in large datasets [6]. However, given these results the *SC* test does not appear
341 appropriate as a general solution for detecting biased migration and differentiation in B cell
342 repertoire datasets.

343

344 We next tested whether biased state change patterns were detected by the *SP* test. To test this
345 method's false positive error rate, we first investigated simulations with totally unbiased state
346 changes; namely, in which lineage trees were equally likely to begin at state *A* as *B* ($\pi_a = 0.5$) and
347 had equal rates of state changes between *A* and *B* ($r_{ab} = r_{ba} = 1$). *SP* tests from *A* to *B* on these
348 datasets resulted in a roughly uniform distribution of *p* values at all tested migration rates ($\delta > 0$,
349 $p < 0.05$ in 5/100; **Fig 2A**). This indicates that completely unbiased state changing is consistent
350 with the null hypothesis of this test. Simulations in which lineages always had state *A* at the root
351 ($\pi_a = 1$) and/or the relative rate of state change was higher from *A* to *B* ($r_{ab} = 10$) were expected
352 to give high *SP* statistics. At low overall rates of state change ($r = 10$), 55/60 of these simulations

353 had significantly high *SP* statistics from *A* to *B* ($\delta > 0$; $p < 0.05$; **Fig 2B-D**). At higher rates of
354 state change ($r = 25, 50, 100$, or 1000), this relationship diminished in these simulations as the
355 distribution of trait values became less distinguishable from random association (**Fig 2B-D**).
356 These results indicate that, under this two state Markov model framework, a significantly high
357 *SP* statistic is associated with biased origination, biased rate of state change, or both depending
358 on the overall rate.

359
360 Finally, we used these simulated datasets to test whether the *SP* test is affected by biased data
361 sampling, as this potential bias is important for some other phylogeographic methods of trait
362 evolution e.g. [29]. We tested this by randomly discarding half of the sequences associated with
363 *A* in simulations with totally unbiased state change ($\pi_a = 0.5$, $r_{ab} = r_{ba} = 1$). Though *SP* tests from
364 *A* to *B* on these datasets gave a uniform *p* value distribution when all sequences were included
365 (**Fig 2A**), *SP* statistics became significantly high when half of *A* sequences were discarded (**Fig**
366 **2e**). This indicates that severely biased sampling may give a similar signature as biased
367 origination or state change for the *SP* test (**Figs 2B-D**). Biased sampling may be caused by a
368 variety of experimental factors, and applications of these statistics to empirical datasets will need
369 to carefully consider possible effects of biased data collection for each trait type.

370
371 *Differentiating complex relationships among trait values*
372
373 All the tests detailed above are extendable to data with more than two states; however, due to its
374 superior performance in two state simulations, we will focus in the rest of this study on the *SP*
375 test. The permutation step of the *SP* test usually permutes trait values within each tree separately

376 (Methods). However, when more than two states are present it may be advantageous to
377 randomize trait value assignments among trees rather than just within each tree. This changes the
378 null hypothesis, which is now that the proportion of state changes observed is the same as that
379 expected if trait values are randomly distributed among all trees. Deviations from this null
380 hypothesis may be due not only to biased ancestor/descendant relationships within individual
381 trees, but also co-occurrence of trait values within different trees. To demonstrate the difference
382 between these two mechanisms, we performed simulations with four trait values: *A*, *B*, *C*, and *D*.
383 To test the difference between simple association and biased ancestry, these simulations used
384 unbiased state change between *A* and *B*, and unidirectional state change from *C* to *D*. Trees
385 began with states *A*, *B*, or *C* in equal probability; state changes were allowed in both directions
386 from *A* to *B* and unidirectionally from *C* to *D*. For each repetition, the rate of state change (r) =
387 10, and relative rates were equal among allowed state changes. Performing the *SP* test on these
388 simulations while permuting among trees showed significantly higher *SP* statistics in both
389 directions between *A* and *B*, and between *C* and *D* than expected (20/20 for each; **Fig 3A**). This
390 indicates the *SP* test when permuting among trees detected the association between these trait
391 values but not the directionality of *C* to *D* state changes. In contrast, the *SP* test when permuting
392 only within trees correctly yielded a significantly high *SP* statistic from *C* to *D* in 19/20
393 simulations; further, no simulation yielded a significantly high *SP* statistic from *D* to *C*,
394 indicating a low false positive rate. No simulation using either permutation method showed a
395 significantly high *SP* statistic between unassociated trait values (e.g. *A* and *C*), indicating a low
396 false positive rate. These results indicate that permuting trait values within trees is a more
397 effective means of detecting biased ancestor/descendant relationships, while permuting between
398 trees is more appropriate for detecting associations among traits (**Fig 3B**).

399 *Differentiating constrained modes of state change*

400

401 In some instances, there are known constraints to the direction that state changes can occur, such
402 as in isotype switching. Isotype-determining constant regions in humans are ordered as IgM/IgD,
403 IgG3, IgG1, IgA1, IgG2, IgG4, IgE, IgA2. Human B cells begin with IgM/IgD, and because the
404 mechanism of class switching is irreversible, these events can only occur sequentially in the
405 order specified. For instance, IgA1 can switch to IgG4, but not to IgM or IgG1. This constraint
406 may be naturally incorporated into the Sankoff parsimony algorithm [22] by making impossible
407 isotype switches have an arbitrarily high weight. A frequent focus of isotype switching analysis
408 is whether a particular isotype (i.e. IgE) arises from direct switching from IgM or from
409 sequential switching from an intermediate isotype [30,31]. These types of hypotheses could be
410 investigated using the *SP* test.

411

412 To determine if the *SP* test can usefully distinguish between types of constrained relationships
413 among trait values, we simulated datasets to represent possible isotype switching patterns. As
414 above, datasets contained four trait values: *A*, *B*, *C*, and *D* under different modes of evolution.
415 Because questions often focus on the origin of a particular isotype [30] we only counted state
416 changes leading to *D* when calculating *SP* statistics. Further, because state changes can only
417 occur in a particular direction, we permute trait values among trees in these tests to increase
418 power. While we previously showed that permuting among trees confuses biased association
419 with biased ancestry (**Fig 3A**), switching between these states can only occur in one direction.
420 Because of this, association between two states implies a direction of switching and among tree
421 permutation is justifiable. We first simulate direct switching in which trees always had state *A* at

422 the root and only state changes from *A* to the other states were allowed (**Fig 3C**). We expected
423 these simulations to show a significantly high *SP* statistic only from *A* to *D*. Confirming this
424 expectation, all 20 of these simulations had a high *SP* statistic from *A* to *D* ($\delta > 0, p < 0.05$; **Fig**
425 **3C**). We next simulated sequential switching, where arriving at state *D* requires transitioning
426 through *B*. All trees began in *A* and state changes were allowed from *A* to *B* and *C*, but *D* arose
427 only from *B*. We expected these simulations to show a significantly high *SP* statistic only from *B*
428 to *D*. All 20 of these simulations showed a significantly high *SP* statistic from *B* to *D* ($\delta > 0, p <$
429 0.05; **Fig 3D**). These results demonstrate that the *SP* test using constrained parsimony can
430 discriminate between simple hypotheses of isotype switch patterns, such as direct versus
431 sequential switching.

432

433 We next investigated whether the *SP* test can distinguish between more complex types of
434 constrained switching. We simulated irreversible isotype switching in which trees begin with
435 state *A*, and only state changes moving alphabetically (*A* to *D*) were allowed. Naively, we may
436 expect these simulations should show similar *SP* test results from *A*, *B*, and *C* to *D*. However, all
437 20 of these simulation repetitions showed a significantly high *SP* statistic to *D* from *B* and *C*, but
438 not from *A* (**Fig 3E**). As a control, we simulated unconstrained switching in which trees begin
439 randomly at any state and may change between all states. Using a constrained parsimony model,
440 these simulations showed the same significantly high *SP* to *D* from *B* and *C*, but not from *A* (**Fig**
441 **3F**), indicating that this pattern is possibly an artifact of the constrained parsimony model. These
442 results demonstrate that, while the *SP* test outlined here can distinguish between simple types of
443 constrained state change, its relationship to more complex modes of constrained state change
444 such as irreversible evolution are difficult to predict, and should be interpreted cautiously.

445

446 *Differentiation of B cell subtypes during HIV infection*

447

448 Over the course of the immune response, B cells differentiate into multiple cellular subsets with
449 distinct properties. Recent studies have focused on the role of T-bet, a transcription factor usually
450 associated with differentiation of T cells, in shaping B cell responses during infection. For
451 example, [6] used data from three HIV+ patients to demonstrate that CD19^{hi} memory B cells
452 (CD19^{hi} MBCs, a surrogate for T-bet+ B cells) represented earlier states in the affinity
453 maturation process than germinal center B cells (GCBCs), and to define the relationships among
454 other B cell subtypes including CD19^{lo} MBCs and un-switched MBCs. More specifically, [6]
455 used the *SC* test with trait values permuted among trees. However, the simulation analyses
456 performed here demonstrated the *SC* test is highly conservative, and that permuting among trees
457 may only detect unstructured association among trait values (**Fig 3**). It is therefore not clear
458 whether the relationship from CD19^{hi} MBCs to GCBCs observed in [6] was driven by biased
459 ancestor/descendant relationships among these cell types within trees. Our results above suggests
460 that the *SP* test using within tree permutation would be a more appropriate test of this
461 relationship.

462

463 We characterized the relationships among B cell subsets with the *SP* test using within tree
464 permutation for each of the three subjects. These analyses showed a significantly high *SP*
465 statistic from CD19^{hi} MBCs to GCBCs and to CD19^{lo} MBCs in all three subjects ($\delta > 0, p <$
466 0.025; **Fig 4B-D**). These analyses confirm the conclusions in [6] that CD19^{hi} MBCs are
467 significantly closer, cladistically, to the predicted germline sequence than GCBC sequences.

468 Naively, one may interpret this as evidence that GCBC cells derive from CD19^{hi} MBCs.
469 However, because GCBCs are expected to have far higher mutation rates than MBCs, the
470 observed patterns are also consistent with early production of CD19^{hi} MBC from GCBCs,
471 followed by a near cessation of mutations in CD19^{hi} MBCs. This is consistent with the
472 conclusions of [6] that CD19^{hi} MBCs represent earlier stages in the GC reaction, rather than the
473 direction of differentiation. Overall, these *SP* tests confirm that the previously observed
474 relationships from CD19^{hi} MBCs and CD19^{lo} MBCs to GCBCs are driven by biased
475 ancestor/descendant relationships within trees rather than simply association in the same trees, as
476 may have been the case from the previously used *SC* tests with among tree permutations [6].

477

478 *Sequential isotype switching to IgE and IgG4*

479

480 Antibody isotypes are a major determinant of function. Of principle interest is characterizing
481 whether IgE antibodies, the primary antibody isotype associated with allergic response, arise
482 directly from IgM switching, or through sequential switching from another downstream isotype
483 [30,31]. Previous studies have shown evidence that IgE in mice and human adults arises from
484 sequential switching primarily from IgG [30,31], though a recent study in 27 humans in the first
485 three years of life found evidence of a greater association between IgA1 and IgE in children with
486 food allergy and eczema [28]. Specifically, [28] showed a higher number of shared clones
487 between IgE and IgA1 than between IgE and other isotypes in these subjects. A phylogenetic test
488 of this relationship would confirm that IgE and IgA1 sequences show a direct ancestor-
489 descendant relationship within these B cell trees rather than just being part of the same clone.

490

491 We applied our discrete trait framework to determine the origins of IgE in a single subject (id =
492 2442) from [28]. This subject was selected due to reported history eczema, food allergy, and B
493 cell clones containing IgE and other isotypes [28]. Using an *SP* test in which only state changes
494 leading to IgE were considered and trait values were permuted among trees, we found a
495 significantly high *SP* statistic from IgA1 to IgE (**Fig 5B**). No other isotype showed a
496 significantly high *SP* statistic to IgE. These results favor IgE arising from sequential switching
497 through IgA1 over direct switching from IgM in this subject. Performing a similar test using only
498 state changes leading to IgG4 revealed a significantly high *SP* statistic from IgG1 and IgG2 to
499 IgG4 (**Fig 5E**). This pattern is similar to irreversible switching within the IgG family (**Fig 3E**).
500 As shown in simulation analyses, this test is not suited to infer relative rates of switching from
501 different isotypes if all kinds of switches are considered. However, these results are most
502 consistent with origin of IgG4 through sequential switching with other IgG isotypes rather than
503 direct switching from IgM or sequential switching from IgA1. Overall, these results are
504 consistent with the conclusions of [28] that IgE arose preferentially through switching from IgA1
505 in this subject. Our results further suggest IgG4 arose preferentially via sequential switching
506 from other IgG subtypes in this subject.

507

508 **Discussion**

509

510 Phylogenetic techniques have the potential to reveal important information about B cell
511 migration, class switching, and cellular differentiation. While great strides have been made using
512 phylogenetic models to study evolution and trait change generally, there are significant
513 challenges to translating these approach to B cells. As a step in this direction, hypotheses about
514 the ancestor/descendant relationships of B cell trait values may be usefully investigated using
515 heuristic approaches that are robust to uncertainties in branch length estimation. Here, we
516 introduce three maximum parsimony-based summary statistics to characterize the distribution of
517 trait values along phylogenetic trees. Significance of all of these statistics is tested by comparing
518 to statistics calculated on trees with permuted data. We demonstrate the efficacy of these tests
519 using simulations, and show that the *SP* test is the most useful for characterizing
520 ancestor/descendant relationships among trait values. We further demonstrate how these
521 statistics can test hypotheses about empirical B cell datasets by characterizing the relationship
522 between T-Bet+ memory B cells and germinal center B cells in three HIV+ patients, and the
523 class switching origins of IgE and IgG4 in a human subject over the first three years of life.

524

525 Simulations demonstrate that the *SP* test was uniquely able to determine the direction of biased
526 origination and state change among the approaches investigated. In simple simulations
527 containing two states (*A* and *B*) a significantly high *SP* statistic from *A* to *B* was associated with
528 origination in *A* and biased state change from *A*. This signal decreased as the overall rate of
529 switching increased. In more complex scenarios, the *SP* test was able to differentiate between
530 traits that were generated through biased state change in a particular direction versus traits that

531 were simply associated with each other. The *SP* test was also able to distinguish between simple
532 modes of constrained evolution such as direct and sequential switching. These results indicate
533 the *SP* test may have broad utility in characterizing ancestor/descendant relationships among B
534 cell discrete traits.

535

536 We next used two datasets to demonstrate that the *SP* test could be used to derive meaningful
537 biological conclusions. In the first, we confirm that T-Bet⁺ memory B cells tend to be the
538 predicted immediate ancestors of GC B cells within lineages trees obtained from three HIV+
539 subjects. Though this relationship may primarily be due to differences in mutation rate over time
540 between memory and GC B cell subsets, this does confirm prior findings and demonstrates that
541 the T-Bet⁺ memory B cell subset represents an earlier state in the affinity maturation process,
542 possibly contributing to an impaired immune response to HIV [6]. We next characterized the
543 isotype switching patterns of sequences obtained from a human over the first three years of life
544 [28]. In this analysis, we found evidence of sequential switching from IgA1 to IgE, as well as
545 evidence of sequential switching from IgG subtypes to IgG4. Sequential switching from IgA1 to
546 IgE is consistent with [28] but not other analyses performed on data taken from adults, which
547 favor sequential switching from IgG [30]. This possibly reflects differences in isotype switching
548 patterns between adults and children. Overall these results demonstrate that the discrete trait
549 analysis framework developed here can be used to test important hypotheses about B cell
550 differentiation and class switching.

551

552 There are a number of limitations with these methods. First, tree topologies were estimated using
553 maximum parsimony. While maximum parsimony is not a statistically consistent estimator of

554 tree topology and is known to give inaccurate predictions over long branch lengths [32], it has
555 been shown to be an accurate estimator of tree topology in certain B cell applications [33], and is
556 widely used in B cell phylogenetic analysis [5,17]. In any case, the statistics presented here are
557 not limited to tree topologies inferred through maximum parsimony. The three statistics
558 proposed here (**Eq. 1-3**) are also based on maximum parsimony, and may have similar
559 inaccuracies over long branch lengths. Further, the statistical tests assume that the process of
560 state change is independent of the tree shape, when the two may be coupled e.g. [34]. This
561 assumption of independence is commonly made in discrete trait analysis e.g. [8] to enable
562 computational tractability, and because the actual link between tree shape and state change is
563 unknown or cannot be modelled. A significantly high *SP* statistic could be potentially caused by
564 factors other than biased state change. For instance, because tree branch lengths represent genetic
565 distance rather than time, it is possible that cell types with low mutation rates over time will
566 spuriously appear ancestral to those with high mutation rates. This effect likely underlies our
567 analysis of B cell subtypes in HIV. Finally, it is possible that SHM is actually occurring at
568 another, un-sampled site which is seeding the sites that were sampled. Overall, it is important to
569 carefully consider alternative explanations when trying to determine the biological basis for a
570 significantly high *SP* statistic.

571

572 An important limitation of the *SP* test is that it, like many other phylogeographic approaches e.g.
573 [29] is affected by biased data sampling. This may arise due to experimental factors that are
574 difficult to control. For instance, under-sampling a trait value may cause a spurious, significantly
575 high *SP* from that trait value. Previous analyses of viral migration have dealt with potential
576 sampling bias by performing tests across multiple down-sampling repetitions [9]. In practice, it

577 can be difficult to know if B cells with certain trait values have been sampled proportionally to
578 their relative population sizes. However, if a type of B cell is known to be under-sampled in a
579 particular experiment, and is predicted to be the descendant of another B cell type, it can be
580 argued that this relationship is unlikely to be due to biased sampling (**Fig 2E**). Alternatively, if
581 multiple samples are tested it is possible that these samples will have a wide range of sequence
582 proportions belonging to different traits. If these differences in sequence proportions are
583 uncorrelated with *SP* test results, it could be argued that observed results are unlikely to be due to
584 consistent under-sampling of B cells with a particular trait value.

585

586 Our simulation analyses revealed that that the *SP* test is difficult to interpret when considering
587 complex constrained models such as irreversible isotype switching (**Fig 3C-E**). To recreate
588 isotype switching, we performed four state simulations in which only state changes proceeding in
589 the direction of state *A*, *B*, *C*, and *D* were allowed. Unexpectedly, these simulations tended to
590 show a significantly low *SP* statistic from *A* to *D*, but a significantly high *SP* statistic for *B* and *C*
591 to *D* (**Fig 3E**). This biased trend is likely driven by the fact, due to constraints in the direction of
592 state change, randomized trees tend to have more switches from *A* than expected based on the
593 relative frequency of *A*. This produces a significantly high *SP* for switches from *A* to *D*. An
594 alternative may be to use the *SP* statistic (**Eq. 3**) without comparing to a null distribution, which
595 is equivalent to comparing the relative frequency of each type of switch observed. However, the
596 observed switch frequency (*SP* statistic) is not proportional to the true relative rate of state
597 change in general. For instance, in the two state Markov model simulations presented here (**Fig**
598 **2**), the *SP* statistic alone is both positively and negatively related to the true relative rate of state
599 change, depending on other parameters (**Supplemental File S5**). Comparing *SP* statistics to

600 those obtained from randomized trees (i.e. the *SP* test) usefully corrects this relationship in
601 unconstrained models (**Fig 2**), but not always in constrained ones (**Fig 3C-E**). Ultimately,
602 isotype switching is a complex, constrained process, and our analyses suggest the relative rates
603 of isotype switching inferred from B cell trees should be interpreted cautiously. We suspect a
604 general method for accurately estimating these rates will require a model-based approach, such
605 as a non-reversible Markov model.

606

607 Future methods to differentiate migration, differentiation, and isotype switching patterns in B
608 cells might improve upon the approach developed here by explicitly modeling these processes
609 along a phylogeny, incorporating branch length information, and better accounting for
610 uncertainty in tree topology. The heuristic approach introduced here crucially does not use
611 branch lengths to help predict internal node states of the tree. Ignoring this source of information
612 likely lowers power, but is possibly advantageous because the relationship between mutation rate
613 and time is not currently well understood, and likely varies by cell type. While the approach
614 developed here uses phylogenetic bootstrap replicates to account for uncertainty in tree topology
615 [24], this may also be done using a posterior distribution of topologies generated by MCMC
616 sampling. This was recently done for naïve sequence inference in individual B cell lineages [35].
617 Phylogenetic bootstrapping has less desirable statistical properties than posterior distributions,
618 but is a widely used means of assessing reproducibility of tree topology and is more
619 computationally tractable for large datasets. Overall, though there is potential for improvement,
620 the approach introduced here effectively deals with important challenges such as incorporating
621 information across trees, accounting for uncertainty in tree topology, and scaling efficiently
622 when analyzing large datasets.

623

624 A phylogenetic discrete trait analysis framework fills an important gap in B cell sequence
625 analysis. The proposed framework provides a principled, flexible, and scalable approach for
626 characterizing migration, class switching, and differentiation in a wide array of contexts. This
627 differs from other phylogenetic tools we developed recently, which used model-based
628 approaches for characterizing somatic hypermutation and clonal selection [4,36]. The methods
629 developed in this paper are available in the R package *dowser*, available at
630 <https://bitbucket.org/kleinsteinkleinste/dowser> as part of the Immcantation suite
631 (<http://immcantation.org>). Scripts for performing analyses in this manuscript are available at:
632 <https://bitbucket.org/kleinsteinkleinste/projects>.

633

634 **Acknowledgements**

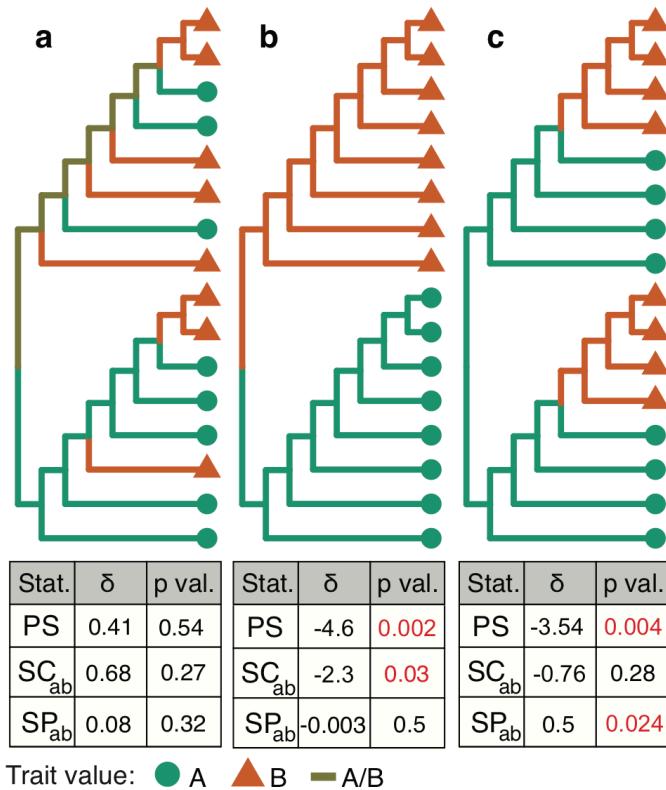
635 K.B.H was supported through a PhRMA Foundation post-doctoral fellowship in informatics.
636 This work was funded in part by the European Research Council under the European Union's
637 Seventh Framework Programme (FP7/2007-2013)/European Research Council grant agreement
638 number 614725-PATHPHYLODYN. This work was funded in part by National Institutes of
639 Health, National Institute of Allergy and Infectious Diseases grant R01 AI104739.

640

641 **Competing interests**

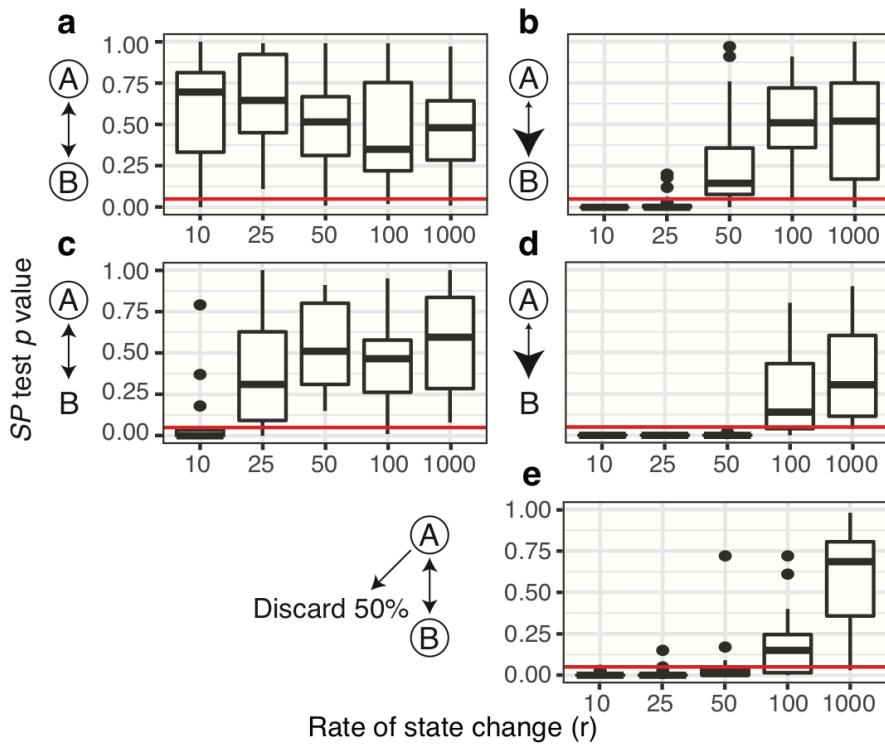
642 S.H.K. receives consulting fees from Northrop Grumman.

643



644

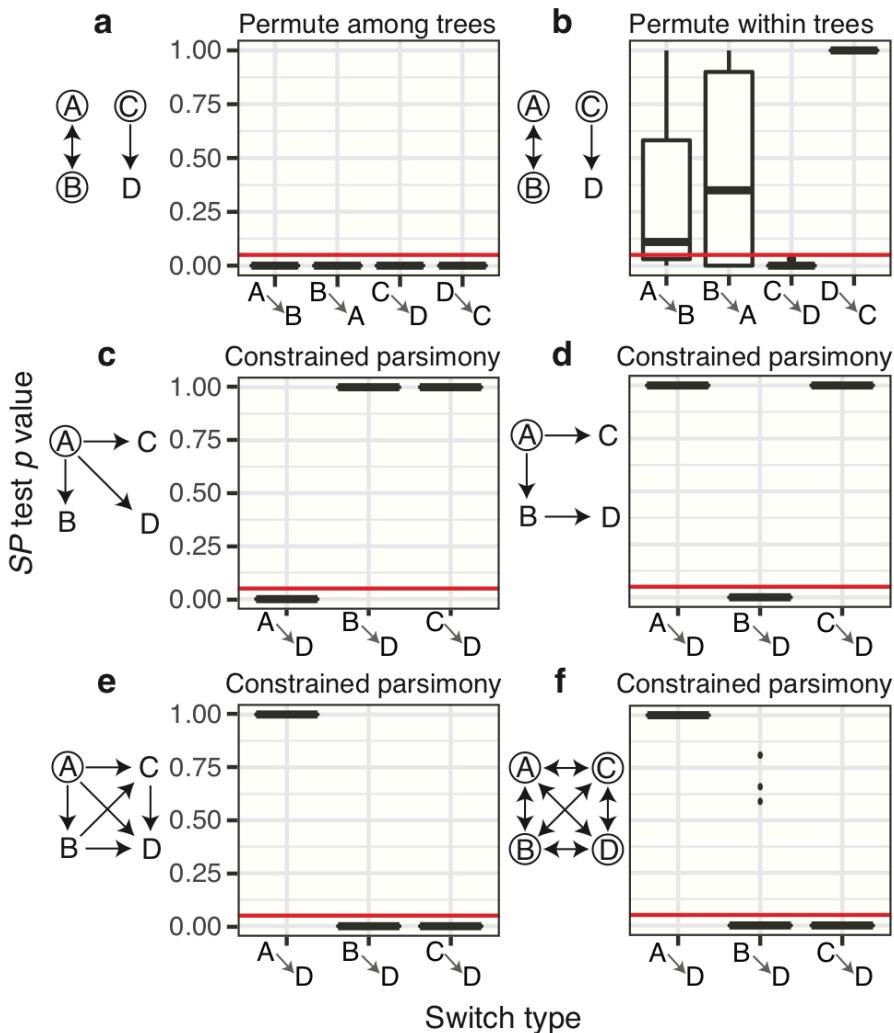
645 **Fig 1:** Hypothetical phylogenies used to illustrate tree/trait association statistics. Trait values at
646 internal nodes of the tree are predicted using maximum parsimony reconstruction given the trait
647 values at the tips, which are shown using different colors and shapes. Below each tree is a table of
648 δ and corresponding p values for PS , SC , and SP tests performed on each tree, calculated using
649 1000 permutations. Tests were performed on the tree topologies themselves - bootstrap replicates
650 were not performed. **(a) No association between tip-trait values and tree:** Distribution of traits
651 across this tree is indistinguishable from randomly distributed traits by any statistic used. **(b) Tip-
652 trait values clustered in tree:** Association between trait and tree structure revealed by significantly
653 low PS statistic. This tree also has a significantly low switch count statistic from A to B (SC_{ab}).
654 Further, this tree has an identical switch proportion statistic (1/2) from A to B as from B to A , which
655 is not significantly different from permuted data (SP_{ab}). **(c) Biased ancestor/descendant
656 relationships among trait values:** As in **b**, this tree also shows a significant relationship between
657 tree and trait distribution (PS). However, this tree also has a significantly high SP statistic from A
658 to B (SP_{ab}). Only the SP test captures the directionality of this relationship.
659



660
661

662 **Fig 2:** Distribution of *SP* test *p* values from *A* to *B* from two state simulation analyses in which
663 state change between state *A* and *B* was determined by the probability of starting in *A* (π_a), relative
664 rate of migrating from *A* to *B* (r_{ab}), and the average rate of state change (r). To the right of each
665 plot, possible starting states are circled, relative rates are shown by arrowhead size. (a) $\pi_a = 0.5$,
666 $r_{ab} = 1$, fully unbiased state change, shows roughly uniform distribution of *p* values at all tested
667 rates. (b) $\pi_a = 1$, $r_{ab} = 1$ shows low *p* values at low rates (10) but not at higher rates. (c) $\pi_a = 0.5$,
668 $r_{ab} = 10$ shows low *p* values at rates < 50 . (d) $\pi_a = 1$, and $r_{ab} = 10$ shows low *p* values at rates $<$
669 100. (e) $\pi_a = 1$, $r_{ab} = 1$ shows low *p* values at rate < 50 if 50% of *A* sequences are discarded.
670 Compared to (a), this shows that *p* values are sensitive to biased sampling of sequences. Red lines
671 show the cutoff of *p* value = 0.05.
672

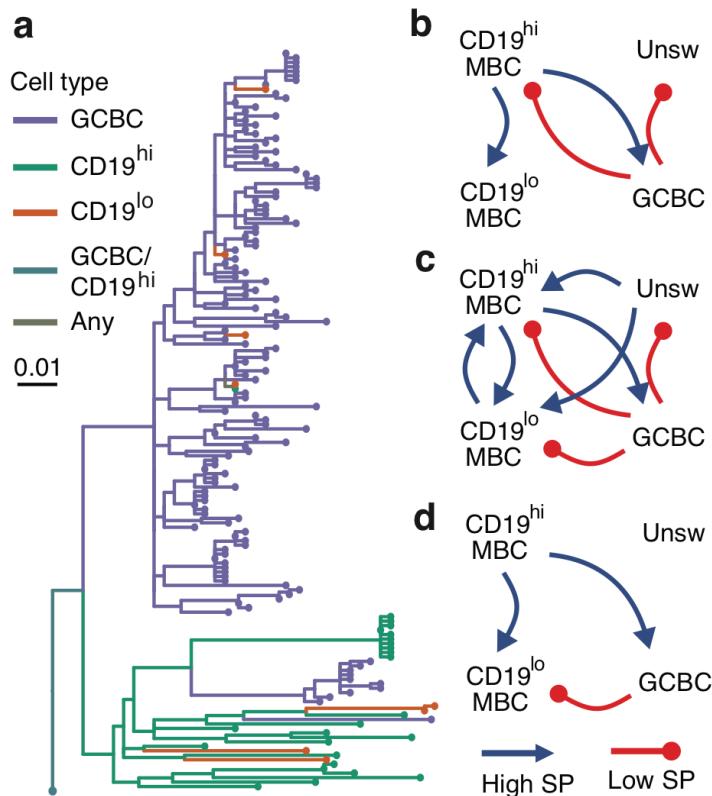
673



674

675

676 **Fig 3:** Distribution of *SP* test *p* values from four state simulation analyses under multiple modes
677 of evolution diagrammed to the left of each plot. Twenty repetitions were performed in each
678 scenario. In simulations, possible starting states are circled and possible state changes are shown
679 with arrows. All allowed state changes occurred at the same relative rate and the total rate of state
680 changing (*r*) was 10 changes/mutation/site (see **Fig 2**). **(a)** Permuting trait values among trees
681 reveals low *p* values for all state changes between *A* and *B*, and *C* and *D*. **(b)** Permuting within
682 each tree reveals low *p* values from *C* to *D*, but not between *A* and *B*. Both **a** and **b** imposed no
683 constraints on the types of state changes allowed in the maximum parsimony algorithm. **(c)** Direct
684 switching simulations result in low *p* values from *A* to *D*, but not from other states to *D*. **(d)**
685 Sequential switching simulations result in low *p* values from *B* to *D* but not from other states to *D*.
686 **(e)** Irreversible switching simulations result in low *p* values from *B* and *C* to *D*, but not from *A*. **(f)**
687 Unconstrained switching simulations also result in low *p* values from *B* and *C* to *D*, but not from
688 *A*. The strange results of **e** and **f** are likely artefacts of the constrained parsimony algorithm, which
689 forbids reverse alphabetical state changes (e.g. *D* to *C*), used to count state changes in simulations
690 **c-d**.
691



692
693

694 **Fig 4:** Analysis of B cell subtypes in three HIV+ subjects. **(a)** Example tree visualized using ggtree
695 [37,38] showing observed relationship between CD19^{hi} MBCs and GCBCs. Ambiguous node
696 states (“CD19^{hi}/GCBC” and “Any”) are also shown. **(b-d)** Direction of significant SP test δ values
697 for subjects 1 **(b)**, 2 **(c)**, and 3 **(d)**. Arrows within each diagram show the direction of significantly
698 high (blue) or significantly low (red) SP statistics between CD19^{hi} MBCs, CD19^{lo} MBCs,
699 unswitched MBCs (Unsw), and GCBCs in each subject.

700

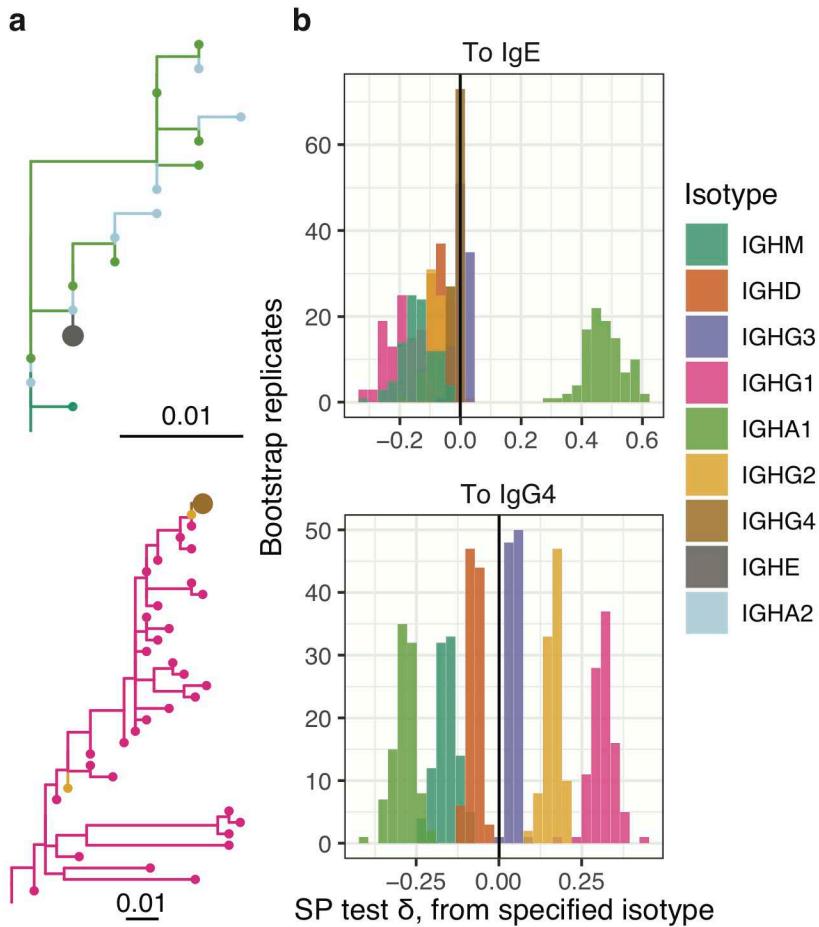


Fig 5: Analysis of antibody isotypes from a single subject. **(a)** Example trees visualized using ggtree [37,38] showing observed relationships between cells expressing BCRs with IgA1 and IgE, and IgG1 and IgG4 isotypes. IgE and IgG4 are indicated on each tree using larger tip circles. **(b)** Distribution of *SP* test δ values to IgE from each of the other isotypes (different colors). **(c)** Distribution of *SP* test δ values to IgG4 from each of the other isotypes (different colors). State changes in **b** and **c** were calculated using constrained parsimony which forbids state changes that violate the geometry of the Ig heavy chain locus, and *SP* tests were performed using permutation among trees.

712 **References**

- 713
- 714 1. Murphy K, Travers P, Walport M, Janeway C. Janeway's immunobiology. New York:
715 Garland Science; 2012.
- 716 2. Shlomchik MJ, Marshak-Rothstein A, Wolfowicz CB, Rothstein TL, Weigert MG. The role
717 of clonal selection and somatic mutation in autoimmunity. *Nature*. 1987;328: 805–811.
718 doi:10.1038/328805a0
- 719 3. Liao H-X, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD, et al. Co-evolution of a broadly
720 neutralizing HIV-1 antibody and founder virus. *Nature*. 2013;496: 469–476.
721 doi:10.1038/nature12053
- 722 4. Hoehn KB, Heiden JAV, Zhou JQ, Lunter G, Pybus OG, Kleinstein SH. Repertoire-wide
723 phylogenetic models of B cell molecular evolution reveal evolutionary signatures of aging
724 and vaccination. *PNAS*. 2019; 201906020. doi:10.1073/pnas.1906020116
- 725 5. Stern JNH, Yaari G, Heiden JAV, Church G, Donahue WF, Hintzen RQ, et al. B cells
726 populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci
727 Transl Med*. 2014;6: 248ra107-248ra107. doi:10.1126/scitranslmed.3008879
- 728 6. Austin JW, Buckner CM, Kardava L, Wang W, Zhang X, Melson VA, et al.
729 Overexpression of T-bet in HIV infection is associated with accumulation of B cells outside
730 germinal centers and poor affinity maturation. *Science Translational Medicine*. 2019;11.
731 doi:10.1126/scitranslmed.aax0904
- 732 7. Horns F, Vollmers C, Croote D, Mackey SF, Swan GE, Dekker CL, et al. Lineage tracing
733 of human B cells reveals the in vivo landscape of human antibody class switching. *eLife*.
734 2016;5: e16578.
- 735 8. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian Phylogeography Finds Its
736 Roots. *PLOS Computational Biology*. 2009;5: e1000520. doi:10.1371/journal.pcbi.1000520
- 737 9. Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, et al. The early spread
738 and epidemic ignition of HIV-1 in human populations. *Science*. 2014;346: 56–61.
739 doi:10.1126/science.1256739
- 740 10. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. Virus genomes
741 reveal factors that spread and sustained the Ebola epidemic. *Nature*. 2017;544: 309–315.
742 doi:10.1038/nature22040
- 743 11. Dellicour S, Baele G, Dudas G, Faria NR, Pybus OG, Suchard MA, et al. Phylodynamic
744 assessment of intervention strategies for the West African Ebola virus outbreak. *Nat
745 Commun*. 2018;9: 1–9. doi:10.1038/s41467-018-03763-2
- 746 12. Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, et al. Establishment and
747 cryptic transmission of Zika virus in Brazil and the Americas. *Nature*. 2017;546: 406–410.
748 doi:10.1038/nature22401

- 749 13. Hill SC, Vasconcelos J, Neto Z, Jandondo D, Zé-Zé L, Aguiar RS, et al. Emergence of the
750 Asian lineage of Zika virus in Angola: an outbreak investigation. *The Lancet Infectious
751 Diseases*. 2019;19: 1138–1147. doi:10.1016/S1473-3099(19)30293-2
- 752 14. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J
753 Mol Evol*. 1981;17: 368–376. doi:10.1007/BF01734359
- 754 15. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian
755 phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol*. 2018;4.
756 doi:10.1093/ve/vey016
- 757 16. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et
758 al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS
759 Computational Biology*. 2019;15: e1006650. doi:10.1371/journal.pcbi.1006650
- 760 17. Ohm-Laursen L, Meng H, Chen J, Zhou JQ, Corrigan CJ, Gould HJ, et al. Local Clonal
761 Diversification and Dissemination of B Lymphocytes in the Human Bronchial Mucosa.
762 *Front Immunol*. 2018;9: 1976. doi:10.3389/fimmu.2018.01976
- 763 18. Slatkin M, Maddison WP. A cladistic measure of gene flow inferred from the phylogenies
764 of alleles. *Genetics*. 1989;123: 603–613.
- 765 19. Nakano T, Lu L, Liu P, Pybus OG. Viral Gene Sequences Reveal the Variable History of
766 Hepatitis C Virus Infection among Countries. *The Journal of Infectious Diseases*. 2004;190:
767 1098–1108. doi:10.1086/422606
- 768 20. Wallace RG, HoDac H, Lathrop RH, Fitch WM. A statistical phylogeography of influenza
769 A H5N1. *PNAS*. 2007;104: 4473–4478. doi:10.1073/pnas.0700435104
- 770 21. Felsenstein J. {PHYLIP} (Phylogeny Inference Package) version 3.6a3. 2002.
- 771 22. Sankoff D. Minimal Mutation Trees of Sequences. *SIAM J Appl Math*. 1975;28: 35–42.
772 doi:10.1137/0128004
- 773 23. Chen R, Holmes EC. Frequent inter-species transmission and geographic subdivision in
774 avian influenza viruses from wild birds. *Virology*. 2009;383: 156–161.
775 doi:10.1016/j.virol.2008.10.015
- 776 24. Felsenstein J. Confidence Limits on Phylogenies: An Approach Using the Bootstrap.
777 *Evolution*. 1985;39: 783. doi:10.2307/2408678
- 778 25. Parker J, Rambaut A, Pybus OG. Correlating viral phenotypes with phylogeny: Accounting
779 for phylogenetic uncertainty. *Infection, Genetics and Evolution*. 2008;8: 239–246.
780 doi:10.1016/j.meegid.2007.08.001
- 781 26. Laserson U, Vigneault F, Gadala-Maria D, Yaari G, Uduman M, Heiden JAV, et al. High-
782 resolution antibody dynamics of vaccine-induced immune responses. *PNAS*. 2014;111:
783 4928–4933. doi:10.1073/pnas.1323862111

- 784 27. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinsteine SH.
785 Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing
786 data. *Bioinformatics*. 2015;31: 3356–3358. doi:10.1093/bioinformatics/btv359
- 787 28. Nielsen SCA, Roskin KM, Jackson KJL, Joshi SA, Nejad P, Lee J-Y, et al. Shaping of
788 infant B cell receptor repertoires by environmental factors and infectious disease. *Science
789 Translational Medicine*. 2019;11: eaat2004. doi:10.1126/scitranslmed.aat2004
- 790 29. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, et al. Unifying Viral
791 Genetics and Human Transportation Data to Predict the Global Transmission Dynamics of
792 Human Influenza H3N2. *PLOS Pathogens*. 2014;10: e1003932.
793 doi:10.1371/journal.ppat.1003932
- 794 30. Looney TJ, Lee J-Y, Roskin KM, Hoh RA, King J, Glanville J, et al. Human B-cell isotype
795 switching origins of IgE. *Journal of Allergy and Clinical Immunology*. 2016;137: 579–
796 586.e7. doi:10.1016/j.jaci.2015.07.014
- 797 31. He J-S, Subramaniam S, Narang V, Srinivasan K, Saunders SP, Carbajo D, et al. IgG1
798 memory B cells keep the memory of IgE responses. *Nature Communications*. 2017;8: 641.
799 doi:10.1038/s41467-017-00723-0
- 800 32. Felsenstein J. Cases in which Parsimony or Compatibility Methods Will be Positively
801 Misleading. *Systematic Zoology*. 1978;27: 401–410. doi:10.2307/2412923
- 802 33. Davidsen K, Matsen FAI. Benchmarking Tree and Ancestral Sequence Inference for B Cell
803 Receptor Sequences. *Front Immunol*. 2018;9. doi:10.3389/fimmu.2018.02451
- 804 34. Hudson RR. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary
805 Biology*. 1991;7: 1–44.
- 806 35. Dhar A, Ralph DK, Minin VN, Matsen IV FA. A Bayesian Phylogenetic Hidden Markov
807 Model for B Cell Receptor Sequence Analysis. *arXiv:190611982 [q-bio, stat]*. 2019 [cited 5
808 Jul 2019]. Available: <http://arxiv.org/abs/1906.11982>
- 809 36. Hoehn KB, Lunter G, Pybus OG. A Phylogenetic Codon Substitution Model for Antibody
810 Lineages. *Genetics*. 2017;206: 417–427. doi:10.1534/genetics.116.196303
- 811 37. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. *ggtree*: an r package for visualization and
812 annotation of phylogenetic trees with their covariates and other associated data. *Methods in
813 Ecology and Evolution*. 2017;8: 28–36. doi:10.1111/2041-210X.12628
- 814 38. Yu G, Lam TT-Y, Zhu H, Guan Y. Two Methods for Mapping and Visualizing Associated
815 Data on Phylogeny Using Ggtree. *Mol Biol Evol*. 2018;35: 3041–3043.
816 doi:10.1093/molbev/msy194

817