TRIBAL: Tree Inference of B cell Clonal Lineages

Leah L. Weber[1], Derek Reiman[2], Mrinmoy S. Roddur[1], Yuanyuan Qi[1], Mohammed El-Kebir[1,†], and Aly A. Khan[2,3,†]

[1]Department of Computer Science, University of Illinois at Urbana-Champaign, IL 61801, USA
[2]Toyota Technological Institute at Chicago, Chicago, IL 60637, USA
[3]Department of Pathology, University of Chicago, Chicago, IL 60637, USA
[†]Correspondence: melkebir@illinois.edu, aakhan@uchicago.edu

Abstract

B cells are a critical component of the adaptive immune system, responsible for producing antibodies that help protect the body from infections and foreign substances. Single cell RNA-sequencing (scRNA-seq) has allowed for both profiling of B cell receptor (BCR) sequences and gene expression. However, understanding the adaptive and evolutionary mechanisms of B cells in response to specific stimuli remains a significant challenge in the field of immunology. We introduce a new method, TRIBAL, which aims to infer the evolutionary history of clonally related B cells from scRNA-seq data. The key insight of TRIBAL is that inclusion of isotype data into the B cell lineage inference problem is valuable for reducing phylogenetic uncertainty that arises when only considering the receptor sequences. Consequently, the TRIBAL inferred B cell lineage trees jointly capture the somatic mutations introduced to the B cell receptor during affinity maturation and isotype transitions during class switch recombination. In addition, TRIBAL infers isotype transition probabilities that are valuable for gaining insight into the dynamics of class switching.

Via *in silico* experiments, we demonstrate that TRIBAL infers isotype transition probabilities with the ability to distinguish between direct versus sequential switching in a B cell population. This results in more accurate B cell lineage trees and corresponding ancestral sequence and class switch reconstruction compared to competing methods. Using real-world scRNA-seq datasets, we show that TRIBAL recapitulates expected biological trends in a model affinity maturation system. Furthermore, the B cell lineage trees inferred by TRIBAL were equally plausible for the BCR sequences as those inferred by competing methods but yielded lower entropic partitions for the isotypes of the sequenced B cell. Thus, our method holds the potential to further advance our understanding of vaccine responses, disease progression, and the identification of therapeutic antibodies.

Availability: TRIBAL is available at https://github.com/elkebir-group/tribal

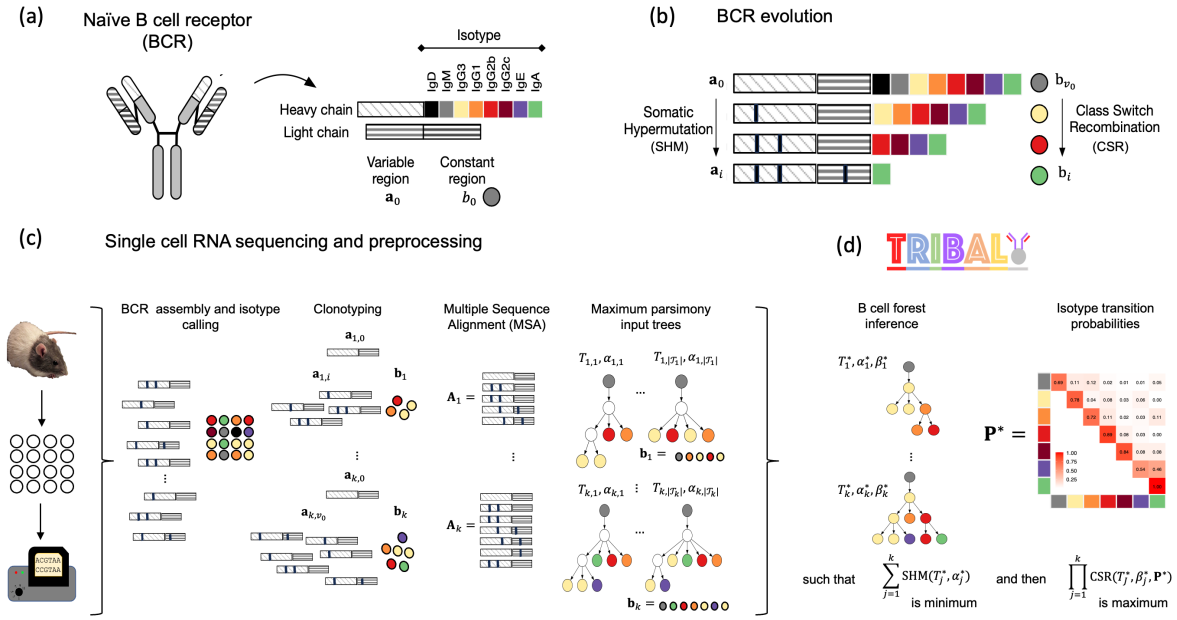


Figure 1: **TRIBAL infers B cell lineage trees and isotype transition probabilities for scRNA-seq data.** (a) A BCR consists of paired heavy and light immunoglobulin chains, each consisting of a variable and constant region. The isotype is the heavy chain constant locus that is transcribed. (b) The BCR of B cells undergo somatic hypermutation/affinity maturation, where point mutations are introduced into the variable region of the heavy and light chains, and class switch recombination, where the heavy chain constant locus undergoes recombination and beings transcribing a different isotype. (c) After scRNA-seq, the variable regions for the light and heavy immunoglobulin alleles are assembled, the isotypes are called and the B cells are clustered into k clonotypes. A multiple sequence alignment $\mathbf{A}_j$ is found for each clonotype j and used to infer a set of input trees with maximum parsimony. The leafs of each input tree are labeled by isotypes $\mathbf{b}$. (d) TRIBAL jointly infers a B cell lineage tree T_j^* for each clonotype j and population-specific isotype transition probabilities $\mathbf{P}^*$ with maximum parsimony for MSA $\mathbf{A}_j$ and maximum likelihood for isotypes $\mathbf{b}_j$.

1 Introduction

B cells play a pivotal role in the adaptive immune response, producing antibodies that neutralize foreign substances and infections [1, 2]. These antibodies, consisting of a heavy chain and a light chain, are initially formed as sequence-specific B cell receptors (BCRs) (Fig. 1a). The generation of specific BCR genes stems from a process known as V(D)J recombination, where DNA segments are rearranged to ensure a wide spectrum of antibodies to counter various pathogens [3]. To enhance their effectiveness, B cells undergo affinity maturation (Fig. 1b) [4], a a micro-evolutionary process involving repeated cycles of *somatic hypermutation* (SHM) and cellular divisions. SHM introduces mutations in the BCR genes, selecting for B cells expressing high-affinity BCRs, while eliminating those with low affinity. Concurrently, B cells have the ability for *class switch recombination* (CSR) (Fig.1b) [5], which diversifies their response by altering the antibody's functional class or *isotype*.

Understanding the evolutionary history of B cells during these adaptive processes is integral to better understanding B cell response to infection and vaccination. However, the selection pressures applied to B cells during the affinity maturation process necessitate more specialized analytical approaches than those utilized for species phylogeny inference [6–9]. Specifically, Hoehn et al. developed HLP17 [10] and HLP19 [11], which are specialized codon substitution models for use with maximum likelihood inference via IgPhyML. Another important difference between B cell and species evolution is the lower mutation rate and the relatively short length of the BCR sequence ($\approx$ 600 bp). These properties imply that maximum parsimony inference methods are viable [12–14] but typically result in a large solution space of many plausible phylogenies for the same data. In addition, these solutions exhibit additional tree uncertainty in the form of *polytomies*, i.e., multifurcating nodes with more than two children. This is counter to the underlying cell lineage tree, which is bifurcating as cell division results in exactly two daughter cells.

One approach to resolve phylogenetic uncertainty is the inclusion of additional data. This approach has proven effective in other areas, such as the use of physical location for studying cancer migration and metastasis [15], or geographical location for the inference of gene flow [16]. More related to B cell lineage inference, sequence abundance has been utilized by both GCTree [12] and ClonalTree [17] to discriminate between candidate solutions. However, both methods were originally developed for bulk RNA sequencing data and consequently were restricted to analysis of only the heavy chain sequences. With single cell RNA-sequencing (scRNA-seq), it is now possible to efficiently assemble BCR sequences that includes both heavy and light chain from a population of B cells [18] (Fig. 1c). As a result, the evolution of BCRs during affinity maturation can now be tracked with scRNA-seq with higher fidelity [19].

In addition, scRNA-seq yields another valuable data source in the form of the expressed isotype of the constant region of the heavy chain (Fig. 1c) [19]. Isotype data extracted from scRNA-seq has been helpful in related immunological domains, such as inferring dynamic cellular trajectories [20]. However, isotype expression is an especially useful marker of B cell evolution. When a B cell undergoes class switching from its current isotype to a new isotype, any heavy chain constant region locus between the current isotype and the new isotype in the genome is cut out or removed via a recombination process (Fig. 1b). Consequently, CSR is an irreversible process and the isotype state of a B cell offers a distinct milestone in its evolutionary history. Therefore, the inclusion of isotype information into the problem of B cell lineage inference has the potential to help minimize phylogenetic uncertainty by both reducing the size of the solution space and yielding more refined B cell lineage trees with fewer polytomies.

In this work, we present TRIBAL, which stands for Tree Inference of B cell Clonal Lineages (Fig. 1d). TRIBAL utilizes both the BCR sequence and isotype information from sequenced cells to infer a B cell lineage tree that jointly models the evolutionary processes of SHM and CSR. Additionally, TRIBAL infers the underlying isotype transition probabilities providing valuable insight into the dynamics of CSR (Fig. 1d). We demonstrate the accuracy of TRIBAL on simulated data and show that it is effective on experimental single cell data generated from the 5' 10x Genomics platform. TRIBAL is available open source and has the potential to improve understanding of vaccine responses, track disease progression, and identify therapeutic antibodies.

2 Problem Statement

Suppose that we have sequenced and subsequently aligned the variable regions of the heavy and light chain of the B cell receptor (BCR) of n B cells that descend from the same naive B cell post V(D)J recombination, resulting in a multiple sequence alignment (MSA) $\mathbf{A}$ composed of m columns — the typical alignment length is $m \approx 650$ (Fig. 1c). That is, for each B cell i, we are given the concatenated sequence $\mathbf{a}_i \in \Sigma^m$ where $\Sigma = \{A, G, C, T, -\}$. In addition, we are given the isotype $b_i \in [r] = \{1, \ldots, r\}$, determined using tools such as Cell Ranger [21]. For humans, there are $r = 8$ isotypes ordered as IgM/D, IgG3, IgG1, IgA1, IgG2, IgG4, IgE and IgA2, whereas for mice there are $r = 7$ isotypes ordered as IgM/D, IgG3, IgG1, IgG2b, IgG2c (2a), IgE, IgGA. Finally, we are given sequence $\mathbf{a}_0$ and isotype $b_0 = 1$ of the ancestral naive B cell of the n B cells — note that $\mathbf{a}_0$ and b_0 is post V(D)J recombination but prior to somatic hypermutation (SHM) and class switch recombination (CSR) and thus $b_0 = 1$. (Fig. 1a).

To study the evolutionary history of these n cells, we aim to construct a B cell lineage tree T that jointly describes the evolution of the DNA sequences $\mathbf{A} = [\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_n]^\top$ and their isotypes $\mathbf{b} = [b_0, b_1, \ldots, b_n]^\top$. As such, each node v of T will be labeled by a sequence $\alpha(v) \in \Sigma^m$ and isotype $\beta(v) \in [r]$. In particular, the root v_0 will be labeled by $\alpha(v_0) = \mathbf{a}_0$ and $\beta(v_0) = b_0 = 1$ while the n leaves $L(T) = \{v_1, \ldots, v_n\}$ of T will be labeled by sequence $\alpha(v_i) = \mathbf{a}_i$ and isotype $\beta(b_i) = b_i$ for each $i \in [n]$. A key property of isotype evolution is that it is *irreversible*. As such, the isotype $\beta(u)$ of an ancestral cell u must be less than or equal to the isotype $\beta(v)$ of its descendants v. More formally, we have the following definition of a B cell lineage tree.

Definition 1. A rooted tree T whose nodes are labeled by sequences $\alpha : V(T) \to \Sigma^m$ and isotypes $\beta : V(T) \to [r]$ is a *B cell lineage tree* for MSA $\mathbf{A} = [\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_n]^\top$ and isotypes $\mathbf{b} = [b_0, b_1, \ldots, b_n]^\top$ provided (i) T has n leaves $L(T) = \{v_1, \ldots, v_n\}$ such that each leaf $v_i \in L(T)$ is labeled by sequence $\alpha(v_i) = \mathbf{a}_i$ and isotype $\beta(v_i) = b_i$, (ii) the root node v_0 of T is labeled by sequence $\alpha(v_0) = \mathbf{a}_0$ and isotype $\beta(v_0) = b_0$, and (iii) for all nodes $u, v \in V(T)$ such that u is ancestral to v it holds that $\beta(u) \leq \beta(v)$.

In the following, we will refer to B cell lineage trees as *lineage trees*. Lineage trees typically have shallow depth due to the limited number of mutations introduced during SHM, making parsimony a reasonable optimization

criterion [12–14]. Given a lineage tree T, the SHM parsimony score is computed as

$$\mathrm{SHM}(T, \alpha) = \sum_{(u,v)\in E(T)} D\left(\alpha(u), \alpha(v)\right), \tag{1}$$

where $D\left(\alpha(u), \alpha(v)\right)$ is the Hamming distance [22] between sequences $\alpha(u)$ and $\alpha(v)$. However, one common challenge of using parsimony to model SHM is that it often results in a large number of candidate lineage trees with equal optimal parsimony score. In addition, many inferred lineage trees contain *polytomies*, or internal nodes with out-degree greater than 2. To overcome these two challenges, we propose to infer lineage trees that optimize both sequence evolution (SHM) and isotype evolution (CSR).

Similarly to SHM, one could model the evolution of CSR using unweighted parsimony. That is, one would prefer lineage trees T with isotypes $\beta : V(T) \to [r]$ that minimize the number of isotype changes, i.e., $\sum_{(u,v)\in E(T)} D\left(\beta(u), \beta(v)\right)$. However, there are two issues with this approach. First, it does not appropriately penalize lineage trees that violate the irreversible property of isotype evolution [13]. Second, it does not account for the fact that given an isotype starting state the probability of transitioning to each of the possible isotype states is not necessarily equal. In fact, knowing these probability distributions is useful for researchers looking to gain basic insight into the patterns and casual factors of class switch recombination [23]. Therefore, we seek to develop an appropriate evolutionary model for CSR that captures the irreversible property of class switching and models preferential isotype class transitions.

We propose a state or tree dependence model [24, 25] evolutionary model for CSR, which models the joint probability distribution of a random variable vector under Markov-like assumptions on a given tree (Appendix A). Here, the random variables of interest in this state tree model are the isotypes $\beta(v)$ of each node v in lineage tree T. This model is parameterized by a probability distribution over the isotype of the root and isotype transition probabilities. As the root v_0 of a lineage tree T is a naive B cell post V(D)J recombination, the isotype $\beta(v_0)$ is always 1 (IgM) and the probability distribution of $\beta(v_0)$ is defined as $\Pr(\beta(v_0) = 1) = 1$ and 0 otherwise. Intuitively, isotype transition probabilities captures the conditional probability of a descendant isotype given the isotype of its parent subject to irreversible isotype evolution. Next, we give a formal definition of isotype transition probabilities.

Definition 2. An $r \times r$ matrix $\mathbf{P} = [p_{s,t}]$ is an *isotype transition probability matrix* provided for all isotypes $s, t \in [r]$ it holds that (i) $p_{s,t} \geq 0$, (ii) $p_{s,t} = 0$ if $s > t$, and (iii) $\sum_{t=1}^{r} p_{s,t} = 1$ for all isotypes $s \in [r]$.

We define the joint likelihood $\mathrm{CSR}(T, \beta, \mathbf{P})$ of the observed isotypes $\mathbf{b}$ for isotype transition probabilities $\mathbf{P}$ and any lineage tree T whose leaves have isotypes $\mathbf{b}$ as

$$\mathrm{CSR}(T, \beta, \mathbf{P}) = \Pr(\mathbf{b} \mid T, \alpha, \beta, \mathbf{P}) = \Pr(\mathbf{b} \mid T, \beta, \mathbf{P}) = \prod_{(u,v)\in E(T)} p_{\beta(u),\beta(v)}. \tag{2}$$

We consider the problem of simultaneously inferring a lineage tree T with nodes labeled by sequences $\alpha(v)$ and isotypes $\beta(v)$ given MSA $\mathbf{A}$ and isotypes $\mathbf{b}$ that optimizes both $\mathrm{SHM}(T, \alpha)$ and $\mathrm{CSR}(T, \beta, \mathbf{P})$. However, a significant barrier to solving this problem is that isotype transition probabilities $\mathbf{P}$ are unknown and need to be inferred. While there have been experimental studies that estimate these quantities under specific biological conditions [23], there currently exists no computational methods to infer these probabilities directly from a sequencing experiment. Here, we will leverage that typical single-cell sequencing experiments yield data from a set of diverse B cell lineages, where each lineage is commonly referred to as a *clonotype*. A common pre-processing step is to first cluster the n sequenced cells into k clonotypes, where each clonotype j contains n_j cells (Fig. 1c). We follow convention and group B cells based on shared V(D)J alleles in heavy and light chains into a clonotype. Moreover, we reason that under many experimental conditions, the transition probabilities between isotypes will be similar to those within clonotypes. Thus, inferring these isotype transition probabilities for k clonotypes will yield higher accuracy than inferring isotype transition probabilities for a single lineage in isolation. This leads to the following problem statement.

Problem 1 (B CELL LINEAGE FOREST INFERENCE (BLFI))**.** Given MSAs $\mathbf{A}_1, \ldots, \mathbf{A}_k$ and isotypes $\mathbf{b}_1, \ldots, \mathbf{b}_k$ for k clonotypes, find isotype transition probabilities $\mathbf{P}^*$ for r isotypes and lineage trees $T_1^*, \ldots, T_k^*$ for $(\mathbf{A}_1, \mathbf{b}_1), \ldots, (\mathbf{A}_k, \mathbf{b}_k)$ whose nodes are labeled by sequences $\alpha_1^*, \ldots, \alpha_k^*$ and isotypes $\beta_1^*, \ldots, \beta_k^*$, respectively, such that $\sum_{j=1}^{k} \mathrm{SHM}(T_j^*, \alpha_j^*)$ is minimum and then $\prod_{j=1}^{k} \mathrm{CSR}(T_j^*, \beta_j^*, \mathbf{P}^*)$ is maximum.

In other words, we prioritize the SHM objective and then among all solutions with minimum SHM score we prefer those that additionally maximize the CSR objective. It is easy to see that the BLFI problem is NP-hard via a simple reduction from the Large Parsimony problem — see Appendix B.1 for more details.

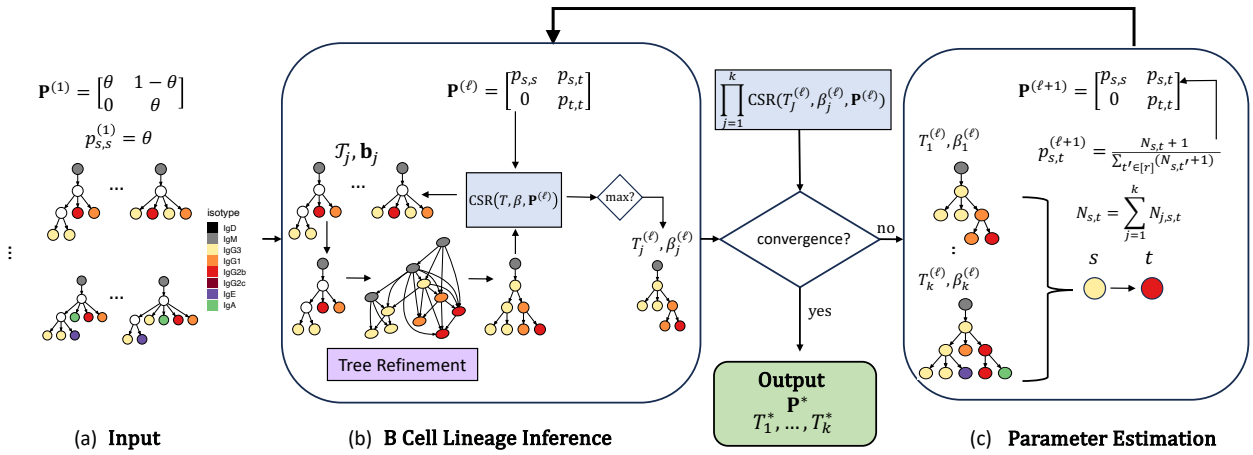


Figure 2: **TRIBAL infers B cell lineage forest $T_1^*, \ldots, T_k^*$ and isotype transitions $\mathbf{P}^*$ for k clonotypes utilizing coordinate ascent.** (a) The inputs to TRIBAL are isotype transition probabilities $\mathbf{P}^{(1)}$, which are initialized given a parameter $\theta \in [0.5, 1]$, and a tuples $(\mathcal{T}_1, \mathbf{b}_1) \ldots, (\mathcal{T}_k, \mathbf{b}_k)$, where set $\mathcal{T}_j$ are maximum parsimony trees for MSA $\mathbf{A}_j$ and $\mathbf{b}_j$ are the observed isotypes of the n_j cells of clonotype j. (b) Conditioning on isotype transition probabilities $\mathbf{P}^{(\ell)}$, a B cell lineage tree $T_j^{(\ell)}$ with nodes labeled by isotypes $\beta_j^{(\ell)}$ is inferred for each clonotype j by solving the MPTR problem for each tree in the input set $\mathcal{T}_j$. (c) Convergence between $\prod_{j=1}^{k} \mathrm{CSR}(T_j, \beta_j, \mathbf{P})$ for iterations ℓ and $\ell - 1$ is checked. If the difference has not converged, isotype transition probabilities $\mathbf{P}^{(\ell+1)}$ are updated using maximum likelihood estimation. If the difference has converged, the current inferred B cell lineage forest and isotype transition probabilities $\mathbf{P}$ are output. Multiple restarts may be performed for different values of θ.

3 Methods

We introduce Tree Inference of B cell Clonal Lineages, or TRIBAL, a method to solve the BLFI problem. This method is based on two key ideas that allow us to effectively solve the BLFI problem.

First, the lexicographical ordering of our two objectives — optimizing for SHM followed by CSR — enables one to use the following two-stage approach (Fig. 1c). In the first stage, we use existing maximum parsimony methods to generate a set $\mathcal{T}$ of input trees — also called a *maximum parsimony forest* — for each clonotype such that each tree $T \in \mathcal{T}$ minimizes the objective $\mathrm{SHM}(T, \alpha)$. To do so, we provide these methods only the sequence information $\mathbf{A}$ to enumerate a solution space $\mathcal{T}$ of trees whose nodes are labeled by sequences $\alpha_1, \ldots, \alpha_{|\mathcal{T}|}$. In the second stage, we incorporate isotype information $\mathbf{b}$ to further operate on the set $\mathcal{T}$ and additionally optimize $\mathrm{CSR}(T, \beta, \mathbf{P})$ in such a manner that maintains optimality of the SHM objective. We note that a lexicographically optimal lineage tree T^* does not necessarily need to be an element of $\mathcal{T}$, but instead it suffices that the evolutionary relationships in tree T^* are a refinement of the evolutionary relationships described by some tree T among the set $\mathcal{T}$ of input trees. More specifically, a *refinement* T' of tree T is obtained by a series of EXPAND operations such that an EXPAND operation of node v consists of splitting node v into v and v', joining them with an edge (v, v') and then redistributing the children of v to be descendants of either v or v'. We have the following key proposition and corollary.

Proposition 1. For any tree T labeled by sequences α and refinement T' of T, there exists a sequence labeling α' for T' such that $\mathrm{SHM}(T, \alpha) = \mathrm{SHM}(T', \alpha')$.

Proof. The sequencing labeling α' is found by setting $\alpha'(v) = \alpha(v)$ and $\alpha'(v') = \alpha(v)$ during each EXPAND operation. By construction, the new edge (v, v') has $D(\alpha'(v), \alpha'(v')) = 0$ and every original edge maintains its original Hamming distance in T'. Therefore, $\mathrm{SHM}(T, \alpha) = \mathrm{SHM}(T', \alpha')$. □

Corollary 1. Any lineage tree T' that lexicographically optimizes $\mathrm{SHM}(T', \alpha')$ and then $\mathrm{CSR}(T', \beta', \mathbf{P})$ must be a refinement of some tree T optimizing only $\mathrm{SHM}(T, \alpha)$.

Therefore, our sought lineage tree T^* that lexicographically optimizes both objectives must be a refinement of some tree T in the set $\mathcal{T}$.

The second key idea is that the inference of optimal lineage trees $T_1^*, \ldots, T_k^*$ is conditionally independent when given isotype transition probabilities $\mathbf{P}$. This motivates the use of a coordinate ascent algorithm where we randomly

initialize isotype transition probabilies $\mathbf{P}^{(1)}$ (Sec. 3.1 and Fig. 2a). Then, at each iteration ℓ, we use isotype transition probabilities $\mathbf{P}^{(\ell)}$ and the input set $\mathcal{T}_j$ of trees to independently infer an optimal lineage tree $T_j^{(\ell)}$ for each clonotype j (Sec. 3.2 and Fig. 2b). This is then followed by estimating updated isotype transition probabilities $\mathbf{P}^{(\ell+1)}$ given trees $T_1^{(\ell)}, \ldots, T_k^{(\ell)}$ (Sec. 3.3 and Fig. 2c). We terminate upon convergence of our CSR objective or when exceeding a specified number of maximum iterations.

TRIBAL is implemented in Python 3, is open source (BSD-3-Clause license), and is available at `https://github.com/elkebir-group/tribal`.

3.1 Input

The input to TRIBAL is a set of k clonotypes with corresponding maximum parsimony forest $\mathcal{T}_j$ for each clonotype j. In addition, we are given isotypes $\mathbf{b}_j$ labeling the leaves of trees $\mathcal{T}_j$ for each clonotype j (Fig.1c, Fig. 2a). Obtaining this input requires a number of preprocessing steps of a scRNA-seq dataset (Fig. 1), including (i) BCR assembly and isotype calling of each sequenced cell, (ii) clonotyping or clustering the cells based on a shared germline alleles for both the heavy and light chains, (iii) obtaining an MSA for sequences within a clonotype and (iv) finding a parsimony forest for each MSA of a clonotype. There preprocessing steps are not part of TRIBAL.

For our coordinate ascent approach, we also require an initialization for the isotype transition probabilities (Fig. 2a). We set the initial transition probabilities to reflect the observation that under baseline conditions, the probability of a B cell undergoing class switching is lower than the probability of it maintaining its original antibody class. [4, 23]. Thus, we initialize $\mathbf{P}^{(1)}$ such that $p_{s,s} > p_{s,t}$ for all isotypes s and t. Let $\theta \in [0.5, 1]$ be the probability that a B cell does not class switch, i.e., $p_{s,s} = \theta$ for each isotype $s < r$ and $p_{s,s} = 1$ if $s = r$. We enforce irreversibility such that $p_{s,t} = 0$, if $s > t$. We then initialize the remaining parameters uniformly, i.e., $p_{s,t} = (1 - p_{s,s})/(r - s)$ where r is the total of number isotypes. We conduct multiple restarts, varying $\theta \in [0.5, 1]$ in each restart.

3.2 B cell lineage tree inference via refinement

As described above, the inference of optimal lineage trees $T_1^{(\ell)}, \ldots, T_k^{(\ell)}$ is conditionally independent given isotype transition probabilities $\mathbf{P}^{(\ell)}$. We therefore focus our discussion on how TRIBAL infers a B cell lineage tree $T_j^{(\ell)}$ for a single clonotype j during iteration ℓ given isotype transition probabilities $\mathbf{P}^{(\ell)}$. By Corollary 1, we solve this problem by finding an optimal refinement T' and corresponding isotype labeling β' for each tree T in the input set $\mathcal{T}_j^{(\ell)}$ and select the one that maximizes our CSR objective (Fig. 2b). Maximizing the log-likelihood of $\mathrm{CSR}(T, \beta, \mathbf{P})$ is equivalent to maximizing a weighted parsimony criterion. This leads to the following problem statement.

Problem 2 (MOST PARSIMONIOUS TREE REFINEMENT (MPTR))**.** Given a tree T on n leaves, isotypes $\mathbf{b} = [b_0, \ldots, b_n]$ and isotype transition probabilities $\mathbf{P}$, find a tree T' with root v_0' and isotype labels $\beta' : V(T') \to [r]$ such that (i) T' is a refinement of T, (ii) $\beta'(v_0') = b_0 = 1$, (iii) $\beta'(v_i') = b_i$ for each leaf $v_i' \in \{v_1', \ldots, v_n'\}$ and (iv) $\log \mathrm{CSR}(T', \beta', \mathbf{P})$ is maximum.

We prove in Appendix B.1 that the MPTR problem is NP-hard. Therefore, we convert this problem to a graph problem (Fig. S2). For each node u in tree T, multiple copies (u, s) are added to the expansion graph $G_{T,\mathbf{b}}$, one for each possible isotype $s \in [r]$ subject to additional constraints depending on the observed isotypes of the leaves descendant from u. The edges $((u, s), (v, t))$ in this graph have weight $\log p_{s,t}$. We then solve the MPTR problem by finding a valid, maximum weight subtree (T', β') in the expansion graph $G_{T,\mathbf{b}}$ such that T' with isotype labels β' is a most parsimonious refinement of T. This is achieved via a mixed-integer linear programming (MILP) formulation, similar to that used to solve the Steiner minimal tree problem [26]. See Appendix C.1 for more details on the construction of the expansion graph from tree T and leaf isotypes $\mathbf{b}$, proof of correctness for this algorithm and the MILP formulation.

To obtain outputs $(T_1^{(\ell)}, \beta_1^{(\ell)}), \ldots, (T_k^{(\ell)}, \beta_k^{(\ell)})$, we set $(T_j^{(\ell)}, \beta_j^{(\ell)})$ to the tree T' and corresponding isotype labeling β' that maximize $\mathrm{CSR}(T', \beta', \mathbf{P}^{(\ell)})$ among all input trees $\mathcal{T}_j^{(\ell)}$ for each clonotype $j \in [k]$.

3.3 Parameter estimation

In the next step, given tuples $(T_1^{(\ell)}, \beta_1^{(\ell)}), \ldots, (T_k^{(\ell)}, \beta_k^{(\ell)})$ of lineage trees and isotype labels for iteration ℓ, we seek updated isotype transition probabilities $\mathbf{P}^{(\ell+1)}$ for iteration $\ell + 1$ such that $\prod_{j=1}^{k} \mathrm{CSR}(T_j^{(\ell)}, \beta_j^{(\ell)}, \mathbf{P}^{(\ell+1)})$ is maximum (Fig. 2c). This is achieved via maximum likelihood estimation. For fixed lineage trees $T_1, \ldots T_k$ and isotypes

$\beta_1, \ldots, \beta_k$, our CSR objective can be rewritten as

$$\prod_{j=1}^{k} \text{CSR}(T_j, \beta_j, \mathbf{P}) = \prod_{j=1}^{k} \prod_{(u,v)\in E(T_j)} p_{\beta_i(u),\beta_i(v)} = \prod_{(s,t)\in[r]\times[r]} p_{s,t}^{N_{s,t}} \tag{3}$$

where $N_{s,t} = \sum_{j=1}^{k} \sum_{(u,v)\in E(T_j)} \mathbf{1}(\beta_j(u) = s, \beta_j(v) = t)$ is the number of transition from s to t across all selected trees and clonotypes. Thus, we seek isotype transition probabilities $\mathbf{P}$ that maximize (3) subject to the constraints that $\sum_{t\in[r]} p_{s,t} = 1$ for every isotype s. This constrained optimization problem is easily solved with the aid of Lagrange multipliers (see Appendix C.2 for details). Additionally, we add pseudocounts of 1 to avoid overfitting the data in the event of unobserved transitions, yielding the following updated probabilities $p_{s,t}^{(\ell+1)}$ for iteration $\ell + 1$.

$$p_{s,t}^{(\ell+1)} = \frac{N_{s,t} + 1}{\sum_{t'\in[r]} (N_{s,t'} + 1)}. \tag{4}$$

After estimating $\mathbf{P}^{(\ell+1)}$, we then infer $(T_1^{(\ell+1)}, \beta_1^{(\ell+1)}), \ldots, (T_k^{(\ell+1)}, \beta_k^{(\ell+1)})$, as discussed in Section 3.2.

4 Results

We evaluated TRIBAL on both *in silico* experiments (Sec. 4.1) and two sets of experimental scRNA-seq data. The first assessed performance on a model immunological system for affinity maturation (Sec. 4.2) and the second on a study of the relationship between age-associated B cells and autoimmune disorders (Sec. 4.3).

4.1 *In silico* experiments

We designed *in silico* experiments to evaluate TRIBAL with known ground-truth isotype transition probabilities $\mathbf{P}$ and lineage trees T labeled by sequences α and isotypes β. Specifically, we used an existing BCR phylogenetic simulator [13] that models SHM but not CSR. We generated isotype transition probabilities $\mathbf{P}$ with $r = 7$ isotypes (as in mice) under two different models of CSR. Briefly, both CSR models assume the probability of not transitioning is higher than the probability of transitioning, but in the *sequential model* there is clear preference for transitions to the next contiguous isotype, while in the *direct model* the probabilities of contiguous and non-contiguous class are similar. Given $\mathbf{P}$, we evolved isotype characters down each ground truth lineage tree T. We generated 5 replications of each CSR model for $k = 75$ clonotypes and $n \in \{35, 65\}$ cells per clonotype, resulting in 20 *in silico* experiments, yielding a total of 1500 ground truth lineage trees. In addition to comparing TRIBAL to existing methods including dnapars [7], dnaml [7] and IgPhyML [10], we also compared to a version of TRIBAL without tree refinement, denoted as TRIBAL-NO REFINEMENT (TRIBAL-NR). To obtain the input set $\mathcal{T}_j$ of trees with maximum parsimony for each clonotype j, we utilized dnapars [7]. We refer to Appendix D for additional details on the simulations. In the following we focus our discussion on *in silico* experiments with $n = 35$ cells per clonotype (see Fig. S5 for $n = 65$).

To evaluate accuracy of isotype transition probability inference, we used *Kullback–Leibler (KL) divergence* [27] to compare the inferred transition probability distribution $\hat{\mathbf{p}}_s$ of each isotype s to the simulated ground truth distribution $\mathbf{p}_s$. KL divergence is defined as $D_{\text{KL}}(\hat{\mathbf{p}}_s || \mathbf{p}_s) = \sum_{q\in[r]} \hat{p}_{s,t} \log(\hat{p}_{s,t}/p_{s,t})$; the lower the KL divergence, the more similar the two distributions. Since no existing methods infer isotype transition probabilities, we restricted this analysis to TRIBAL and TRIBAL-NR. Overall, we observed good concordance between simulated and TRIBAL inferred isotype transition probabilities (Fig. 3a). Specifically, TRIBAL had lower median KL divergence than TRIBAL-NR for all isotype starting states, except IgA, which is trivially 0, under both direct and sequential CSR models (direct: median of 0.15 vs. 0.73; sequential: median of 0.099 vs. 0.55). We observed improved performance of TRIBAL (but not for TRIBAL-NR) for $n = 65$ cells per clonotype (Fig. S5a and Fig. S6). To assess the sensitivity of TRIBAL to infer isotype transition probabilities with fewer than $k = 75$ clonotypes, we downsampled the 75 clonotypes to 25 and 50 clonotypes per experiment. We observed similar trends for experiments with $k \in \{25, 50\}$ clonotypes, with TRIBAL continuing to outperform TRIBAL-NR while still achieving small KL divergences even as k decreases (Fig. S7). These findings demonstrate that tree refinement is key to accurately estimate isotype transition probabilities.

Next, we assessed the accuracy of lineage tree inference using the *Robinson-Foulds (RF) distance* [28] normalized by the total number of bipartitions in the ground truth T and inferred lineage tree $\hat{T}$ (25). Since TRIBAL, TRIBAL-NR and dnapars return multiple optimal solutions, we report the mean of the lineage tree inference metrics over all

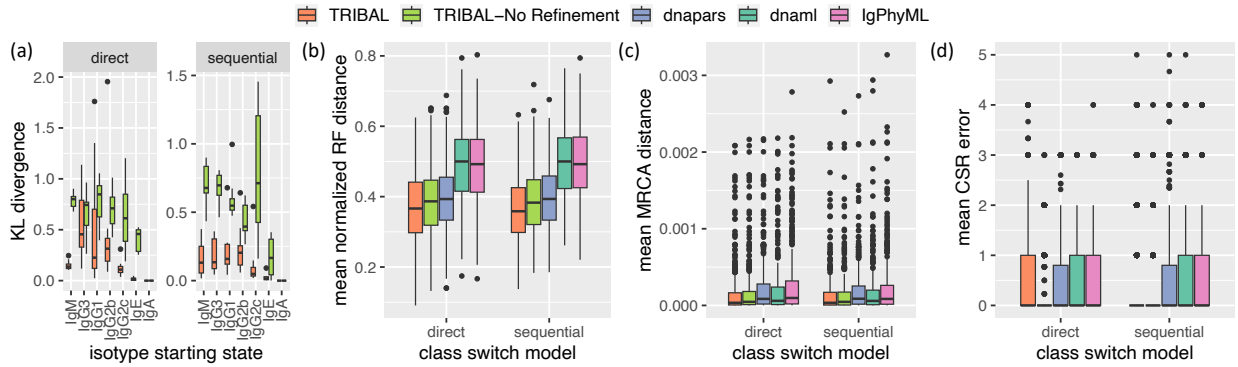


Figure 3: **TRIBAL accurately infers isotype transition probabilities on simulated data while outperforming existing methods on lineage tree inference.** Simulation results shown are for 5 replications with $k = 75$ clonotypes per replication and $n = 35$ cells per clonotype. (a) KL divergence between inferred isotype transition probabilities and the reference ground truth distribution. (b) mean Robinson-Foulds distance between ground truth and inferred lineages tree per clonotype. (c) mean MRCA distance (26) between ground truth and inferred lineage trees per clonotype. (d) mean CSR error between ground truth and inferred B cells.

optimal solutions. We found TRIBAL had the lowest mean normalized RF distance for both direct and sequential CSR models (Fig. 3b). Overall, dnaml (median: 0.5) and IgPhyML (median: 0.49) had the worst performance on normalized RF. Interestingly, even though the starting trees of dnapars are used by TRIBAL, both TRIBAL (median: 0.36) and TRIBAL-NR (median: 0.38) outperformed dnapars (median: 0.39), showing the importance of using isotype information to resolve phylogenetic uncertainty.

While normalized RF distance only assesses the accuracy of the tree topology, it is important to also assess the accuracy of the ancestral sequence reconstruction. To that end, we used a metric called *Most Recent Common Ancestor (MRCA) distance* (26) introduced by Davidsen and Matsen [13]. For any two simulated B cells (leaves), the MRCA distance is the Hamming distance between the MRCA sequences of these two B cells in both the ground truth and inferred lineage trees. This distance is then averaged over all pairs of simulated B cells — see Appendix D.4 and Fig. S4a for additional details. Again, we report the mean of over all optimal solutions for TRIBAL, TRIBAL-NR and dnapars. We found TRIBAL outperformed all other methods (Fig. 3c), achieving the lowest overall median MRCA distance (3.46×10^{-5}), followed by TRIBAL-NR (4.63×10^{-5}). IgPhyML had the worst performance with a median of 8.78×10^{-5}. Performance trends were consistent between methods across both CSR models.

Lastly, we assessed the accuracy of isotype inference by a new metric called *CSR error*, which is computed for each B cell i and clonotype j and is the absolute difference between the number of ground-truth class switches and inferred number of class switches that occurred along its evolutionary path from the root — see Appendix D and Fig S4b for additional details. Since dnaml, dnapars and IgPhyML do not infer isotypes for internal nodes, we pair these methods with the Sankoff algorithm [29] using $w_{s,t}$ equals 1 if $s = t$, 0 if $s < t$ and ∞ otherwise. We account for the presence of multiple solutions by taking the mean across solutions. All methods had a median CSR error of 0 for both the direct and sequential models (Fig. 3d). Therefore, we utilized the third quartile for a more robust comparison. We found that under the direct model TRIBAL-NR (third quartile: 0) was the best performing method, while dnapars was second best (third quartile 0.8) and all other methods, including TRIBAL had a third quartile of 1. We observed a slight tendency of TRIBAL to overestimate the number of transitions due to the tree refinement step, while other methods tended to underestimate the number of transitions. However, under a sequential model, where refinement is helpful in accurately capturing sequential state transitions, we found that TRIBAL was tied with TRIBAL-NR for the best performance (third quartile: 0). All other methods had similar performance between both CSR models for this metric.

We observed similar trends on these metrics for *in silico* experiments containing $n = 65$ cells per clonotype for $k = 75$ clonotypes (Fig. S5). In summary, these results suggests that the incorporation of isotype data and the use of tree refinement are beneficial for both lineage tree inference and ancestral sequence and class switch reconstruction.

4.2 Keyhole limpet haemocyanin (NP-KLH) antigen immune response studies

We applied TRIBAL as well as IgPhyML to $10\times$ $5'$ scRNA-seq data of B cells extracted from mice immunized with nucleoprotein keyhole limpet haemocyanin (NP-KLH), a commonly used antigen in the study of antibody affinity maturation [30]. Our goal was to determine whether these methods recapitulate known patterns of B cell lineage

evolution for this well-studied antigen using data from two studies and to compare the lineage trees inferred by each method. The first dataset (NP-KLH-1) was generated from C57BL/6 mice that were immunized with NP-KLH and total germinal center B cells were extracted 14 days after immunization [20]. The other two datasets came from a single study in which C57BL/6 mice were immunized with NP-KLH (NP-KLH-2a and NP-KLH-2b) and NP-specific germinal center B cells were extracted 13 days after immunization [31]. We utilized the standard $10\times$ Cell Ranger [21] single-cell bioinformatics pipeline to generate sequence $\mathbf{a}_i$ and isotype b_i for each cell i. We used Dandelion [32] to remove doublets, reassign alleles, and cluster the cells into clonotypes. We identified clonotype MSAs $\mathbf{A}_1, \ldots, \mathbf{A}_k$ based on shared V(D)J alleles for the heavy chain using the dowser package [14]. Finally, we excluded clonotypes with fewer than 5 cells. This yielded a total of $n = 2670$ sequenced B cells clustered into $k = 295$ clonotypes. We exclude methods that rely on sequence abundance as a key signal, such as GCTree [12] and ClonalTree [17] as we observed very few duplicated sequences within each clonotype. Fig. 4a shows the distribution of isotypes by dataset and Table S1 includes a more detailed summary of each dataset.

We used dnapars [7] to infer TRIBAL's input set $\mathcal{T}_j$ for each clonotype j. We found that TRIBAL's use of isotype information significantly reduced the number of optimal solutions identified by dnapars (mean: 31.5 vs. 1.3, max: 4310 vs. 8) — see Fig 4b. While IgPhyML, a maximum likelihood method using the HLP19 codon-substitution model [10, 11], infers only a single tree per clonotype, it is important to note that there might be multiple trees with maximum likelihood in the solution space. Indeed, we found high concordance of HLP19 likelihoods between the TRIBAL and IgPhyML inferred lineage trees, with a small overall mean absolute deviation of 0.97 (Fig. 4c, Fig. S8). We even observed that TRIBAL had a greater likelihood than IgPhyML in 59.3% of the clonotypes. Thus, TRIBAL resulted in a significant reduction in the size of the solution space compared to the maximum parsimony method dnapars with similar (and sometimes better) HLP19 likelikelihood as IgPhyML, illustrating how isotype information can be used to effectively reduce phylogenetic uncertainty.

Next, we assessed whether the lineage trees inferred by TRIBAL recapitulated expected biological trends for the NP-KLH model system. Previous research has indicated that the IGHV1-72*01 (VH186.2) variable gene in the heavy chain is preferentially used in the anti-NP response in C57BL/6 mice [33] via two distinct mutation paths: one in which a W33L mutation greatly increases affinity to NP, and another in which high affinity is achieved through K59R and the accumulation of several other mutations, such as S66N [34, 35]. Consequently, we expect that for clonotypes containing both W33L and K59R, these mutations would tend towards occurring in distinct lineages of the tree. We analyzed the inferred pairwise relationships between W33L & K59R in the 26 lineage trees that contained both mutations. Specifically, we categorized the relationship as K59R $\rightarrow$ W33L if K59R was ancestral to W33L, W33L $\rightarrow$ K59R if W33L was ancestral to K59R, *incomparable* if W33L and K59R occurred on distinct lineages of the tree and *same* if they were introduced on the same edge of the lineage tree. Indeed, we confirmed the tendency for mutual exclusivity of W33L and K59R by finding that the proportion of pairwise introductions categorized as *incomparable* was 0.69 and 0.67 for TRIBAL and IgPhyML, respectively (Fig. 4d). Additionally, it has been suggested that W33L mutations appear relatively early during the anti-NP response, whereas the K59R and S66N mutations typically appear later in the evolutionary history [36]. Defining level as the length of the shortest path from the MRCA of all B cells, we observed that W33L occurred at a median level of 1 for both TRIBAL and IgPhyML while both the K59R and S66N mutation occurred at a median level 2 for both methods (Fig. S9). This indicated that W33L was typically introduced earlier in the evolutionary history of a clonotype than K59R and S66N. Thus, both TRIBAL and IgPhyML trees recapitulate expected mutation patterns for this model system.

We next assessed the extent of agreement with isotype information. While TRIBAL infers isotype labels of ancestral nodes, IgPhyML does not have this capability. Therefore, we developed a new metric called *average isotype clade entropy*, which is computed with respect to the isotype labeling of the leaf set. For this metric, we compute the entropy of clade u in tree T with respect to all isotype leaf labels that are descendants of node u, taking the average entropy over all non-trivial clades, which excludes the root and the leaves (Appendix E.1). As IgPhyML returns bifurcating trees, we collapse edges with zero branch length for a more fair comparison of this metric. We observed lower average isotype clade entropy for TRIBAL (median: 0.82) versus IgPhyML (median 0.91) 4f. Fig. 4e depicts the lineage tree inferred by TRIBAL and IgPhyML for the NP-KLH-2a dataset (clonotype B_34_1_5_41_1_5). The TRIBAL inferred tree for this clonotype had lower isotype clade entropy than IgPhyML (TRIBAL: 0.51 vs IgPhyML: 0.86) while also resulting in a greater HLP19 likelihood (TRIBAL: -366.5 vs IgPhyML: -369.5). Thus, we find that the trees identified by TRIBAL are in better agreement with the leaf isotypes than IgPhyML.

In addition to the inferred B cell lineage trees, TRIBAL also inferred isotype transition probabilities $\mathbf{P}$ for each dataset (Fig. 4g, Fig S10). All three inferred isotype transition probability matrices more closely matched a CSR model of direct switching as opposed to a strictly sequential model. To compare the consistency of these estimates across

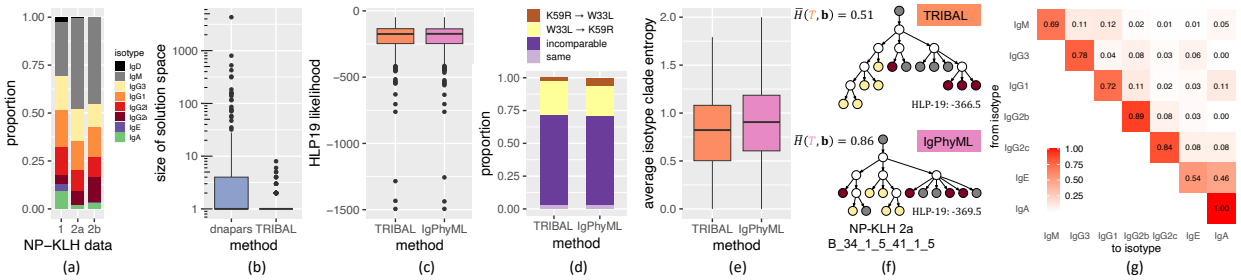


Figure 4: **Comparison between TRIBAL and IgPhyML on the NP-KLH data.** (a) The distribution of isotypes $\mathbf{b}$ in each dataset. (b) A comparison of the solution space of dnapars versus TRIBAL. (c) The distribution of the HLP19 codon-substitution likelihood [11] for lineage trees inferred by TRIBAL and IgPhyML. (d) Observed distribution of evolutionary relationships between the W33L and K59 in clonotypes where both mutations are present. (e) The distribution of the average clade entropy with respect to an isotype labeling $\mathbf{b}$ of the leafset. (f) A comparison of lineage trees inferred for clonotype NP-KLH-2a B_34_1_5_41_1_5 with the average isotype clade entropy $\bar{H}(T, \mathbf{b})$ reported for each inferred tree. (g) TRIBAL inferred isotype transition probabilities $\mathbf{P}$ for NP-KLH-1.

datasets, we computed the Jensen-Shannon divergence (JSD) between the distribution of isotype transition probabilities for each isotype starting state IgM through IgG2c for each dataset pair. We observed low JSD (median: 0.029) across a total of 15 pairwise comparisons, suggesting consistent estimates between isotype transition probabilities.

In summary, these analyses show that the inclusion of isotype information and tree refinement has the potential to yield high quality lineage tree inference, even under a simpler model of SHM, i.e., parsimony. Moreover, the TRIBAL inferred lineage trees additionally optimize an evolutionary model of CSR, yielding lower isotype entropy partitions of the leaf set than IgPhyML. Finally, the additional inference of isotype transitions probabilities $\mathbf{P}$ has the potential to distinguish between direct versus sequential switching events.

4.3 Age-associated B cell (ABC) datasets

We evaluated TRIBAL on three scRNA-seq datasets with V region sequencing that investigated the relationship between age-associated B cells (ABCs) and autoimmune disorders [37]. For each dataset, B cells were extracted from the spleen of a MRL/lpr female mouse and sequenced using $10\times$ $5'$ scRNA-seq. The data was processed by the $10\times$ Cell Ranger [21] single-cell bioinformatics pipeline to generate sequence $\mathbf{a}_i$ and isotype b_i for each cell i. Nickerson et al. [37] identified clonotype MSAs $\mathbf{A}_1, \ldots, \mathbf{A}_k$ based on shared V(D)J alleles for the heavy chain using the dowser package [14] and inferred B cell lineage trees using IgPhyML for each clonotype. After filtering out clonotypes with fewer than 5 sequences, we retained 599 B cells and 54 clonotypes across the three datasets (Table S2). Fig. 5a shows the proportion of isotypes and annotations by mouse for the retained B cells. Of these 54 clonotypes, 35 had more than one distinct isotype across the sequenced B cells, with a median of 3 distinct isotypes per clonotype.

We ran TRIBAL separately on each of the three mouse datasets, obtaining a maximum parsimony forest $\mathcal{T}_j$ for each clonotype j via dnapars. Similar to our NP-KLH analysis, we found that TRIBAL effectively utilized the additional isotype data to reduce the number of optimal solutions identified by dnapars (mean: 8.1 vs. 1.3, max: 165 vs. 4) — see Fig. 5b. The HLP19 likelihood of the TRIBAL inferred lineage trees had high concordance with the IgPhyML inferred trees (mean absolute deviation: 0.97), with TRIBAL yielding a higher likelihood for 53% of the clonotypes (Fig. 5c, Fig. S11). The average isotype clade entropy for the 35 clonotypes with more than one distinct isotype was significantly lower for TRIBAL than for IgPhyML (median: 0.49 vs. 0.77) — see Fig. 5d. An example comparison is shown in Fig. 5e,f for clonotype Mouse-1 775. The tree refinement step of TRIBAL yielded a tree with a significantly lower average isotype clade entropy when compared to IgPhyML (0.65 vs. 1.2) while both trees had identical HLP19 likelihoods (-41.4). Finally, we observed that the isotype transition probabilities reveal evidence for both direct and sequential switching of isotypes (Fig. S12).

In summary, both TRIBAL and IgPhyML yield lineage trees with very similar HLP19 likelihoods, giving support for the validity of the TRIBAL inferred lineage trees in terms of sequence evolution. However, TRIBAL jointly optimizes evolutionary models for both SHM and CSR, yielding trees with lower average isotype clade entropy.

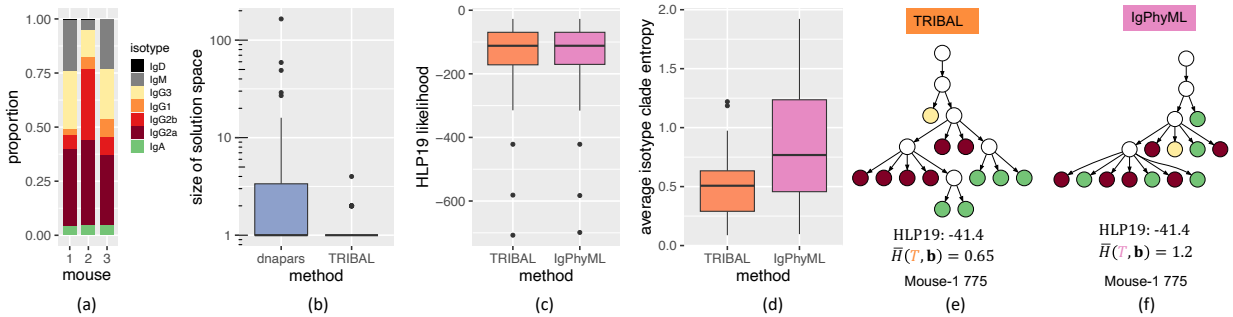


Figure 5: **Comparison between TRIBAL and IgPhyML on ABC data.** (a) Distribution of B cell isotypes. (b) A comparison of the solution space of dnapars versus TRIBAL. (c) The distribution of the HLP19 codon-substitution likelihood [11] for lineage trees inferred by TRIBAL and IgPhyML. (d) Comparison of average isotype clade entropy for TRIBAL versus IgPhyML. (e,f) Comparison of inferred B cell lineage trees by TRIBAL (e) and IgPhyML (f) for clonotype Mouse-1 775 — see (a) for color legend.

5 Conclusion

The development and application of methods for inferring B cell lineage trees and isotype transition probabilities from scRNA-seq data is crucial for advancing our understanding of the immune system. In this work, we introduced TRIBAL, a method to infer B cell lineage trees and isotype transition probabilities from scRNA-seq data. TRIBAL makes use of existing maximum parsimony methods to optimize an evolutionary model for SHM, then incorporates isotype data to find the most parsimonious refinement, i.e., maximizing the CSR likelihood, among the input set of trees. We proved that the subproblem of finding a refinement maximizing the CSR likelihood is NP-hard. We demonstrated the effectiveness of TRIBAL via *in silico* experiments and on experimental data. On *in silico* experiments, we showed the importance of tree refinement for both accurately estimating isotype transition probabilities and lineage tree inference. Furthermore, we demonstrated on experimental data that TRIBAL returns lineage trees that have similar HLP19 likelihoods, despite utilizing a less complex model for sequence evolution but yield a reduction in the entropy of the isotype leaf labelings.

There are several directions for future research. First, integration of germline "sterile" transcripts may offer a way to initialize the TRIBAL inferred isotype transition probabilities [38]. Second, many existing B cell lineage inference methods, such as IgPhyML, yield multifurcating trees when zero length branches are collapsed. There exists an opportunity to combine likelihood or distance based inference methods with the tree refinement step of TRIBAL. Third, the MPTR problem has a more general formulation with the potential for wider applications beyond the problem of B cell lineage inference. For example, sample location is useful in refining tumor phylogeny with polytomies [15]. On a related note, we hypothesize that there are special cases of the MPTR problem and its more general formulation that are in P. Such special cases may include a weight matrix with unit costs and an upper triangular weight matrix that adheres to the triangle inequality. Fourth, the assumption that a single isotype transition probability matrix is shared by all clonotypes could be relaxed to allow the inference of multiple matrices per experiment and an assignment of clonotypes to an inferred matrix. Fifth, TRIBAL could be extended to allow for the correction of inaccurately clonotyped B cells. Finally, more robust evolutionary models for SHM are needed to capture the presence of complex mutations, such as indels, introduced during affinity maturation [39, 40].

Acknowledgments. Authors thank Harinder Singh, Ken Hoehn, Mark Shlomchik and Margie Ackerman for insightful discussions. This work was partially supported by NIH grant DP2AI177884 (A.A.K.) and by the National Science Foundation grant CCF-2046488 (M.E-K.). This work used resources, services, and support provided via the Greg Gulick Honorary Research Award Opportunity supported by a gift from Amazon Web Services.

References

[1] Murphy, K. & Weaver, C. *Janeway's immunobiology* (Garland science, 2016).

[2] Meffre, E., Casellas, R. & Nussenzweig, M. C. Antibody regulation of B cell development. *Nature immunology* **1**, 379–385 (2000).

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

[4] Victora, G. D. & Nussenzweig, M. C. Germinal centers. *Annu Rev Immunol* **30**, 429–457 (2012).

[5] Stavnezer, J., Guikema, J. E. & Schrader, C. E. Mechanism and regulation of class switch recombination. *Annual review of immunology* **26**, 261 (2008).

[6] Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology* **7**, 1–8 (2007).

[7] Felsenstein, J. PHYLIP (phylogeny inference package) version 3.6. distributed by the author. *http://www.evolution.gs.washington. edu/phylip.html* (2004).

[8] Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**, 268–274 (2015).

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

[10] Hoehn, K. B., Lunter, G. & Pybus, O. G. A phylogenetic codon substitution model for antibody lineages. *Genetics* **206**, 417–427 (2017).

[11] Hoehn, K. B. *et al.* Repertoire-wide phylogenetic models of b cell molecular evolution reveal evolutionary signatures of aging and vaccination. *Proceedings of the National Academy of Sciences* **116**, 22664–22672 (2019).

[12] DeWitt III, W. S., Mesin, L., Victora, G. D., Minin, V. N. & Matsen IV, F. A. Using genotype abundance to improve phylogenetic inference. *Molecular Biology and Evolution* **35**, 1253–1265 (2018).

[13] Davidsen, K. & Matsen IV, F. A. Benchmarking tree and ancestral sequence inference for B cell receptor sequences. *Frontiers in immunology* **9**, 2451 (2018).

[14] Hoehn, K. B., Pybus, O. G. & Kleinstein, S. H. Phylogenetic analysis of migration, differentiation, and class switching in b cells. *PLoS computational biology* **18**, e1009885 (2022).

[15] El-Kebir, M., Satas, G. & Raphael, B. J. Inferring parsimonious migration histories for metastatic cancers. *Nature Genetics* **50**, 718–726 (2018).

[16] Slatkin, M. & Maddison, W. P. A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics* **123**, 603–613 (1989).

[17] Abdollahi, N., Jeusset, L., de Septenville, A., Davi, F. & Bernardes, J. S. Reconstructing B cell lineage trees with minimum spanning tree and genotype abundances. *BMC bioinformatics* **24**, 70 (2023).

[18] Canzar, S., Neu, K. E., Tang, Q., Wilson, P. C. & Khan, A. A. BASIC: BCR assembly from single cells. *Bioinformatics* **33**, 425–427 (2017).

[19] Neu, K. E., Tang, Q., Wilson, P. C. & Khan, A. A. Single-cell genomics: approaches and utility in immunology. *Trends in immunology* **38**, 140–149 (2017).

[20] Reiman, D. *et al.* Pseudocell tracer—a method for inferring dynamic trajectories using scRNAseq and its application to B cells undergoing immunoglobulin class switch recombination. *PLoS computational biology* **17**, e1008094 (2021).

[21] Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049 (2017). URL `https://doi.org/10.1038/ncomms14049`.

[22] Hamming, R. W. Error detecting and error correcting codes. *The Bell system technical journal* **29**, 147–160 (1950).

[23] Horns, F. *et al.* Lineage tracing of human B cells reveals the in vivo landscape of human antibody class switching. *Elife* **5**, e16578 (2016).

[24] Ronen, O., Rohlicek, J. R. & Ostendorf, M. Parameter estimation of dependence tree models using the EM algorithm. *IEEE Signal Processing Letters* **2**, 157–159 (1995).

[25] Chow, C. & Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory* **14**, 462–467 (1968).

[26] Sridhar, S., Lam, F., Blelloch, G. E., Ravi, R. & Schwartz, R. Mixed integer linear programming for maximum-parsimony phylogeny inference. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **5**, 323–331 (2008).

[27] Kullback, S. & Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics* **22**, 79–86 (1951).

[28] Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Mathematical biosciences* **53**, 131–147 (1981).

[29] Sankoff, D. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics* **28**, 35–42 (1975).

[30] Cumano, A. & Rajewsky, K. Clonal recruitment and somatic mutation in the generation of immunological memory to the hapten NP. *The EMBO journal* **5**, 2459–2468 (1986).

[31] Chen, D. *et al.* Coupled analysis of transcriptome and BCR mutations reveals role of OXPHOS in affinity maturation. *Nature Immunology* **22**, 904–913 (2021). URL `https://doi.org/10.1038/s41590-021-00936-y`.

[32] Stephenson, E. *et al.* Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine* **27**, 904–916 (2021). URL `https://doi.org/10.1038/s41591-021-01329-2`.

[33] Allen, D., Simon, T., Sablitzky, F., Rajewsky, K. & Cumano, A. Antibody engineering for the analysis of affinity maturation of an anti-hapten response. *The EMBO journal* **7**, 1995–2001 (1988).

[34] Weiser, A. A. *et al.* Affinity maturation of B cells involves not only a few but a whole spectrum of relevant mutations. *International immunology* **23**, 345–356 (2011).

[35] Furukawa, K., Akasako-Furukawa, A., Shirai, H., Nakamura, H. & Azuma, T. Junctional amino acids determine the maturation pathway of an antibody. *Immunity* **11**, 329–338 (1999).

[36] Jacob, J., Przylepa, J., Miller, C. & Kelsoe, G. In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl) acetyl. iii. the kinetics of V region mutation and selection in germinal center B cells. *The Journal of experimental medicine* **178**, 1293–1307 (1993).

[37] Nickerson, K. M. *et al.* Age-associated B cells are heterogeneous and dynamic drivers of autoimmunity in mice. *Journal of Experimental Medicine* **220**, e20221346 (2023).

[38] Ng, J. C. *et al.* sciCSR infers B cell state transition and predicts class-switch recombination dynamics using single-cell transcriptomic data. *bioRxiv* 2023–02 (2023).

[39] Tan, J. *et al.* A LAIR1 insertion generates broadly reactive antibodies against malaria variant antigens. *Nature* **529**, 105–109 (2016).

[40] Lupo, C., Spisak, N., Walczak, A. M. & Mora, T. Learning the statistics and landscape of somatic mutation-induced insertions and deletions in antibodies. *PLOS Computational Biology* **18**, e1010167 (2022).

[41] Foulds, L. R. & Graham, R. L. The Steiner problem in phylogeny is NP-complete. *Advances in Applied mathematics* **3**, 43–49 (1982).

[42] Karp, R. M. *Reducibility among combinatorial problems* (Springer, 2010).

[43] Warnow, T. *Computational phylogenetics: an introduction to designing methods for phylogeny estimation* (Cambridge University Press, 2017).

A Evolutionary model for class switch recombination

A dependence tree $\bar{T}$ with n nodes, sometimes referred to as a state tree, is a tree that defines the conditional independence structure of the random variables associated with the nodes of the tree [24, 25]. Simply put, it is a type of Bayesian network, where the underlying directed acyclic graph is a tree. For each node i in dependence tree $\bar{T}$, we associate a random variable $Y_i \in S$, where S is a discrete state space. Additionally, Y_1 is the random variable associated with the root node. The joint probability $\mathbf{Y} = (Y_1, \ldots, Y_n)$ of these random variables given the underlying structure of dependence tree $\bar{T}$ is defined as follows

$$P(\mathbf{Y} \mid \bar{T}) = P(Y_1) \prod_{i=2}^{n} P(Y_i \mid Y_{\phi(i)}), \tag{5}$$

where $[n] = \{1, \ldots, n\}$ and $\phi(i)$ is a function that returns the parent of node i specified by dependence tree $\bar{T}$. Like Markov chains, this model is parameterized by a distribution over the starting state, $\pi_s = P(Y_1 = s)$, where $\sum_{i \in [r]} \pi_s = 1$, and transition probabilities $p_{s,t} = P(Y_i = t \mid Y_\phi(i) = s)$ from state s to t. Transition probabilities $\mathbf{P} = [p_{s,t}] \in [0,1]^{|S| \times |S|}$ have the property that $\sum_{t \in S} p_{s,t} = 1$ for every state s.

We model class switch recombination with a dependence tree $\bar{T}$ for each lineage tree T with nodes $V(\bar{T}) = V(T)$ and edges $E(\bar{T}) = E(T)$ and isotype labels $\beta(v) = Y_v$ for every node v. In words, the corresponding dependence tree $\bar{T}$ for B cell lineage tree T has the same topology but the dependence tree $\bar{T}$ has a random variable for isotype associated with each node.

For the leaf nodes $v_i \in L(T)$, i.e., the sequenced B cells, we observe isotype b_i directly from scRNA-seq data, i.e., $P(Y_{v_i} = b_i) = 1$. Additionally, root node v_0 in T, represents the naive B cell post V(D)J recombination and therefore $P(Y_{v_0} = 1) = 1$, meaning that root of the dependence tree has isotype state IgM and $\pi_1 = 1$. This means that any dependence tree without $Y_{v_0} = 1$ has zero probability and we omit the initial state probability term. For any other node u in T, we have $\beta(u) = Y_u$. Thus, to compute (5) for isotype labels $\beta(v)$, it suffices to know the isotype transition probabilities $\mathbf{P}$, which we formally define below.

(Main Text) Definition 2. An $r \times r$ matrix $\mathbf{P} = [p_{s,t}]$ is an *isotype transition probability matrix* provided for all isotypes $s, t \in [r]$ it holds that (i) $p_{s,t} = 0$ if $t > s$, (ii) $p_{s,t} \geq 0$, and (iii) $\sum_{t=1}^{r} p_{s,t} = 1$ for all isotypes $s \in [r]$.

The additional conditions on the transition probabilities beyond row stochasticity on the transition probabilities are to properly model the irreversibility of class switch recombination. Using the above model for class switch recombination, we compute the likelihood $\mathrm{CSR}(T, \beta, \mathbf{P})$ for observed isotypes $\mathbf{b}$ given a lineage tree with isotypes β and isotype transition probabilities as follows.

$$\begin{aligned} \mathrm{CSR}(T, \beta, \mathbf{P}) &= \Pr(\mathbf{b} \mid T, \beta, \mathbf{P}) \\ &= \prod_{(u,v) \in E(T)} p_{\beta(u), \beta(v)} \\ &= \prod_{v \in V(T) \setminus \{v_0\}} \prod_{(s,t) \in [r] \times [r]} p_{s,t}^{\mathbf{1}(\beta(v)=t, \beta(\phi(v))=s)} \\ &= \prod_{(s,t) \in [r] \times [r]} p_{s,t}^{N_{s,t}} \end{aligned} \tag{6}$$

where $N_{s,t}$ is the count of occurrences in lineage tree T such that $\beta(v) = t$ and $\beta(\phi(v)) = s$. This is easily extended for a forest of k lineage trees $T_1, \ldots, T_k$ with corresponding isotypes $\beta_1, \ldots, \beta_k$. Given isotype transition probabilities $\mathbf{P}$, the joint probabilities $\mathrm{CSR}(T_1, \beta_1, \mathbf{P}), \ldots, \mathrm{CSR}(T_k, \beta_k, \mathbf{P})$ are conditionally independent, resulting in the joint

likelihood

$$\begin{aligned}\prod_{j=1}^{k} \mathrm{CSR}(T_j, \beta_j, \mathbf{P}) &= \prod_{j=1}^{k} \Pr(\mathbf{b}_j \mid T_j, \beta_j, \mathbf{P}) \\ &= \prod_{j=1}^{k} \prod_{(u,v)\in E(T_j)} p_{\beta_j(u),\beta_j(v)} \\ &= \prod_{j=1}^{k} \prod_{v\in V(T_j)\setminus\{v_0\}} \prod_{(s,t)\in[r]\times[r]} p_{s,t}^{\mathbf{1}(\beta_j(v)=t,\beta_j(\phi(v))=s)} \\ &= \prod_{(s,t)\in[r]\times[r]} p_{s,t}^{\sum_{j=1}^{k} N_{j,s,t}}\end{aligned} \tag{7}$$

where $N_{j,s,t}$ is the count of occurrences in lineage tree T_j such that $\beta(v) = t$ and $\beta(\phi(v)) = s$.

B Combinatorial characterization and complexity results

B.1 B cell lineage forest inference

Recall the B CELL LINEAGE FOREST INFERENCE PROBLEM (BLFI) from Sec. 2, restated below for convenience.

(Main Text) Problem 1 (B CELL LINEAGE FOREST INFERENCE (BLFI))**.** Given MSAs $\mathbf{A}_1, \ldots, \mathbf{A}_k$ and isotypes $\mathbf{b}_1, \ldots, \mathbf{b}_k$ for k clonotypes, find isotype transition probabilities $\mathbf{P}^*$ for r isotypes and lineage trees $T_1^*, \ldots, T_k^*$ for $(\mathbf{A}_1, \mathbf{b}_1), \ldots, (\mathbf{A}_k, \mathbf{b}_k)$ whose nodes are labeled by sequences $\alpha_1^*, \ldots, \alpha_k^*$ and isotypes $\beta_1^*, \ldots, \beta_k^*$, respectively, such that $\sum_{j=1}^{k} \mathrm{SHM}(T_j^*, \alpha_j^*)$ is minimum and then $\prod_{j=1}^{k} \mathrm{CSR}(T_j^*, \beta_j^*, \mathbf{P}^*)$ is maximum.

Theorem 1. The BLFI problem is NP-hard even if $k = 1$ and $r = 1$.

We prove that the BLFI problem is NP-hard via a simple reduction from the LARGE PARSIMONY problem [41]. Although this problem is well known, we restate it here for completeness.

Problem 3 (LARGE PARSIMONY (LP))**.** Given a matrix $\mathbf{A} \in \{0,1\}^{n\times m}$, find a rooted tree T whose nodes are labeled by sequences $\alpha : V(T) \to \{0,1\}^m$ such that the n leaves are labeled by the rows of $\mathbf{A}$ and $\sum_{(u,v)\in E(T)} D(\alpha(u), \alpha(v))$ is minimium.

The reduction to BLFI proceeds by using the same MSA $\mathbf{A}$ directly for a single clonotype, i.e., $k = 1$. Additionally, we restrict the number r of isotypes to 1, and set isotypes $\mathbf{b} = [1]^n$.

Lemma 1. Tree T and node labeling α form an optimal solution to LP instance $\mathbf{A}$ if and only if tree T, sequences α and isotypes β, the isotype transition probabilities $\mathbf{P}$ form an optimal solution to BLFI instance $(\mathbf{A}, \mathbf{b})$.

Proof. ($\Rightarrow$) Let tree T and sequence labeling α be an optimal solution to the LP problem. We will show that T and α can be augmented to form an optimal solution to the corresponding BLFI problem. We set $\mathbf{P} = [1]$. We also set $\beta(v) = 1$ for all nodes $v \in T$. We claim that $(T, \alpha, \beta, \mathbf{P})$ form an optimal solution to BLFI. Assume for a contradiction there exists a solution $(T', \alpha', \beta', \mathbf{P}')$ such that $\mathrm{SHM}(T', \alpha') < \mathrm{SHM}(T, \alpha)$, or $\mathrm{SHM}(T', \alpha') = \mathrm{SHM}(T, \alpha)$ and $\mathrm{CSR}(T', \beta', \mathbf{P}') > \mathrm{CSR}(T, \beta, \mathbf{P})$. Clearly, any feasible solution to BLFI must use $\beta(v) = 1$ for all nodes v and $\mathbf{P} = [1]$ as $r = 1$. This means that any feasible solution to BLFI will have a CSR objective value of 1. Therefore, $\mathrm{CSR}(T', \beta', \mathbf{P}') = \mathrm{CSR}(T, \beta, \mathbf{P}) = 1$. Hence, $\mathrm{SHM}(T', \alpha') < \mathrm{SHM}(T, \alpha)$. As can be seen in (1), the SHM objective equals the objective of the LP problem. Therefore, T' and α' have a lower parsimony score than T and α, a contradiction.

($\Leftarrow$) Let $(T, \alpha, \beta, \mathbf{P})$ be an optimal solution to BLFI. Again, as the SHM objective equals the objective of the LP problem, it directly follows that (T, α) form an optimal solution to the LP problem instance. □

B.2 Most parsimonious tree refinement

B.2.1 Combinatorial characterization

Recall from the main text the definition of isotype transition probabilities $\mathbf{P}$, the CSR log-likelihood for isotypes $\mathbf{b}$ of a tree T with nodes labeled by isotypes β, and the MOST PARSIMONIOUS TREE REFINEMENT problem, provided below for convenience.

(Main Text) Definition 2. An $r \times r$ matrix $\mathbf{P} = [p_{s,t}]$ is an *isotype transition probability matrix* provided for all isotypes $s, t \in [r]$ it holds that (i) $p_{s,t} \geq 0$, (ii) $p_{s,t} = 0$ if $s > t$, and (iii) $\sum_{t=1}^{r} p_{s,t} = 1$ for all isotypes $s \in [r]$.

$$\log \mathrm{CSR}(T, \beta, \mathbf{P}) = \log \prod_{(u,v)\in E(T)} p_{\beta(u),\beta(v)} = \sum_{(u,v)\in E(T)} \log p_{\beta(u),\beta(v)}.$$

(Main Text) Problem 2 (MOST PARSIMONIOUS TREE REFINEMENT (MPTR))**.** Given a tree T on n leaves, isotypes $\mathbf{b} = [b_0, \ldots, b_n]$ and isotype transition probabilities $\mathbf{P}$, find a tree T' with root v'_0 and isotype labels $\beta' : V(T') \to [r]$ such that (i) T' is a refinement of T, (ii) $\beta'(v'_0) = b_0 = 1$, (iii) $\beta'(v'_i) = b_i$ for each leaf $v'_i \in \{v'_1, \ldots, v'_n\}$ and (iv) $\log \mathrm{CSR}(T', \beta', \mathbf{P})$ is maximum.

Let σ be a mapping from $V(T')$ to $V(T)$ that reverses all EXPAND operations of each node u' in refinement T' in order to obtain back the node $\sigma(u') = u$ from which it was derived in the original tree T. We say that an isotype labeling $\beta' : V(T') \to [r]$ of T' is transitory if along each directed edge (u', v') of T' either the isotype changes or u' and v' correspond to two distinct nodes of T. More formally, we have the following definition.

Definition 3. Let T' be a refinement of a tree T whose leaves are labeled by isotypes $\mathbf{b}$. Then, an isotype labeling β' of T' is *transitory* provided (i) $\beta'(v'_0) = 1$ where v'_0 is the root of T', (ii) $\beta'(v') = b_{\sigma(v')}$ for each leaf $v' \in L(T')$, (iii) $\beta'(u') \leq \beta(v')$ for each edge (u', v') of T', and (iv) $\beta'(u') = \beta'(v')$ only if $\sigma(u') \neq \sigma(v')$ for each edge (u', v') of T'.

Importantly, among the set of optimal solutions (T', β') to each MPTR problem instance $(T, \mathbf{b}, \mathbf{P})$ there exist solutions where β' is transitory.

Lemma 2. Let $(T, \mathbf{b}, \mathbf{P})$ be an MPTR problem instance. There exist an optimal solution (T', β') where β' is transitory.

Proof. We prove this by contradiction. Let (T', β') be an optimal solution where β' is not transitory. First, observe that it holds that $\beta'(u') \leq \beta(v')$ for each edge (u', v') of T'. To see why, if there were an edge (u', v') such that $\beta'(u') > \beta(v')$ then $\mathrm{CSR}(T', \beta', \mathbf{P}) = -\infty$ as $\log p_{s,t} = -\infty$ if $s > t$. However, setting $\beta'(u') = 1$ for nodes u' would result in log-likelihood greater than $-\infty$. Since (T', β') is a feasible solution to MPTR respecting irreversibility of isotype transitions, it means that condition (iv) of Definition 3 is violated. Let (u', v') be an edge such that $\beta'(u') = \beta'(v')$ and $\sigma(u') = \sigma(v')$. We can contract this edge, retaining the isotype labeling β' for the remaining nodes, such that the resulting tree remains a refinement of T and the objective value remains unchanged as $\log p_{s,s} = 0$. Repeating this procedure for all edges (u', v') such that $\beta'(u') = \beta'(v')$ and $\sigma(u') = \sigma(v')$ results in (T'', β''), where T'' is a refinement of T labeled by β'', with the same optimal score as (T', β'). Clearly, (T'', β'') is transitory, proving the lemma. □

B.2.2 Complexity

Note that maximizing the CSR log-likelihood is equivalent to maximizing the CSR likelihood, which is the objective function we will use in this subjection. That is,

$$\mathrm{CSR}(T, \beta, \mathbf{P}) = \prod_{(u,v)\in E(T)} p_{\beta(u),\beta(v)}.$$

We now prove the following theorem.

Theorem 2. The MPTR problem is NP-hard.

We show that MPTR is NP-hard by reduction from SET COVER.

Problem 4 (SET COVER)**.** Given a universe $\mathcal{U}$ of elements $\{u_1, \dots u_{|\mathcal{U}|}\}$ and a collection $\mathcal{S}$ of subsets $\{S_1, \dots, S_{|\mathcal{S}|}\}$ such that $\bigcup_{i=1}^{|\mathcal{S}|} S_i = \mathcal{U}$, find a cover $\mathcal{C} \subseteq \mathcal{S}$ such that $\bigcup_{S \in \mathcal{C}} S = \mathcal{U}$ and the size $|\mathcal{C}|$ of the cover is minimum.

Note that while the order of the subsets in collection $\mathcal{S}$ does not matter for SET COVER, our reduction will assume the subsets to be in an arbitrary but fixed order. Similarly, we will assume $\mathcal{U}$ to be ordered arbitrarily. SET COVER has been proven to be NP-hard in Karp's 21 NP-complete problems [42]. We describe a polynomial time reduction from SET COVER to MPTR. To that end, given the set $\mathcal{U}$ of elements and the collection $\mathcal{S}$ of subsets, we construct a tree T with $|\mathcal{U}| + 1$ leaves, $r = |\mathcal{U}| + |\mathcal{S}| + 2$ isotypes, observed isotypes $\mathbf{b} \in [r]^{|\mathcal{U}|+1}$, and $r \times r$ transition probabilities $\mathbf{P}$. The steps are as follows.

1. To construct tree T, we begin by adding the root node v_0. Following that, we attach two children, denoted as $\bar{v}_0$ and $v_{|\mathcal{U}|+1}$, to the root node v_0. Finally, for each element $u_q \in \mathcal{U}$, we add an edge $(\bar{v}_0, v_q)$ in tree T. The constructed tree T has $|\mathcal{U}| + 3$ nodes and $|\mathcal{U}| + 2$ edges.

2. We consider a total of $r = |\mathcal{S}| + |\mathcal{U}| + 2$ isotypes, each corresponding to either a subset $S_i \in \mathcal{S}$, an element $u_q \in \mathcal{U}$, or one of the special symbols $\top$ or $\bot$. Specifically, the first isotype stands for the special symbol $\top$, followed by $|\mathcal{S}|$ isotypes representing each subset $S_i \in \mathcal{S}$, succeeded by $|\mathcal{U}|$ isotypes representing each element $u_q \in \mathcal{U}$, and concluding with the last isotype signifying the special symbol $\bot$. For convenience, we define a function $R : \mathcal{S} \cup \mathcal{U} \cup \{\top, \bot\} \to [r]$ to map the subsets $S_i \in \mathcal{S}$, the elements $u_q \in \mathcal{U}$, and the special symbols $\top$ and $\bot$ to their representative isotype indices as follows.

$$R(X) = \begin{cases} 1, & \text{if } X = \top, \\ i + 1, & \text{if } X = S_i, \\ |\mathcal{S}| + q + 1, & \text{if } X = u_q, \\ |\mathcal{S}| + |\mathcal{U}| + 2, & \text{if } X = \bot. \end{cases}$$

3. For the observed isotypes, we set $b_0 = b_{|\mathcal{U}|+1} = R(\top) = 1$, and $b_q = R(u_q)$ for $1 \le q \le |\mathcal{U}|$.

4. We define ϵ to be a constant such that $0 < \epsilon \le 1/(|\mathcal{S}| + |\mathcal{U}| + 1)$. Next, we construct the isotype transition probabilities $\mathbf{P}$ parameterized by ϵ as follows.

   (a) We set the transition probability from $R(\top)$ to $R(\top)$ or $R(S_i)$ for any set $S_i \in \mathcal{S}$ to be ϵ and to $R(u_q)$ for any $u_q \in \mathcal{U}$ to be 0.

   $$\begin{aligned} p_{R(\top),R(\top)} &= \epsilon, & \\ p_{R(\top),R(S_i)} &= \epsilon & \forall 1 \le i \le |\mathcal{S}|, \\ p_{R(\top),R(u_q)} &= 0 & \forall 1 \le q \le |\mathcal{U}|. \end{aligned}$$

   (b) We set the transition probability from $R(\top)$ to $R(\bot)$ to be $1 - (1 + |\mathcal{S}|)\epsilon$.

   $$p_{R(\top),R(\bot)} = 1 - (1 + |\mathcal{S}|)\epsilon.$$

   (c) We set the transition probability $p_{R(S_i),R(S_j)}$ for any $S_i, S_j \in \mathcal{S}$ to be ϵ if $i < j$, and 0 otherwise.

   $$p_{R(S_i),R(S_j)} = \begin{cases} \epsilon, & \text{if } i < j, \\ 0, & \text{if } i \ge j, \end{cases} \qquad \forall 1 \le i, j \le |\mathcal{S}|.$$

   (d) We set the transition probability from $R(S_i)$ to $R(u_q)$ for any set $S_i \in \mathcal{S}$ and any element $u_q \in \mathcal{U}$ to be ϵ if $u_q \in S_i$, and 0 otherwise.

   $$p_{R(S_i),R(u_q)} = \begin{cases} \epsilon, & \text{if } u_q \in S_i, \\ 0, & \text{if } u_q \notin S_i, \end{cases} \qquad \forall 1 \le i \le |\mathcal{S}|, 1 \le q \le |\mathcal{U}|.$$

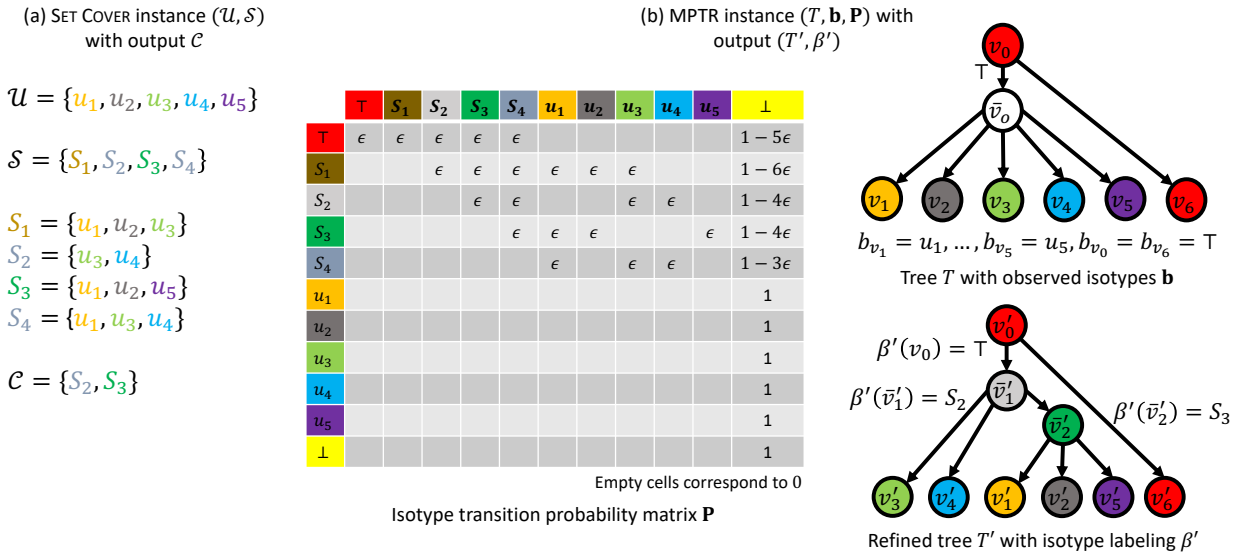


Figure S1: **Polynomial time reduction from SET COVER to MPTR.** (a) shows a SET COVER instance $(\mathcal{U}, \mathcal{S})$, with the corresponding minimum set cover $\mathcal{C}$. The constructed MPTR instance $(T, \mathbf{b}, \mathbf{P})$, along with the output (T', β') is shown in (b). Isotypes are indicated through colors. The mapping function R is omitted, with the isotypes directly represented by elements, subsets, $\top$, or $\bot$. The empty boxes in the transition probability matrix $\mathbf{P}$ corresponds to 0.

(e) For each $S_i \in \mathcal{S}$, we set the transition probability from $R(S_i)$ to $R(\top)$ to be 0 and to $R(\bot)$ to be $1 - (|\mathcal{S}| - i + |S_i|)\epsilon$.

$$p_{R(S_i),R(\top)} = 0 \qquad 1 \le i \le |\mathcal{S}|,$$
$$p_{R(S_i),R(\bot)} = 1 - (|\mathcal{S}| - i + |S_i|)\epsilon \qquad 1 \le i \le |\mathcal{S}|.$$

(f) For any $u_q \in \mathcal{U}$, we set the transition probability from $R(u_q)$ to any other isotype except $\bot$ to be 0. We set $p_{R(u_q),R(\bot)}$ for any $u_q \in \mathcal{U}$ to be 1.

$$p_{R(u_q),R(X)} = 0 \qquad \forall 1 \le q \le |\mathcal{U}|, X \in \mathcal{S} \cup \mathcal{U} \cup \{\top\},$$
$$p_{R(u_q),R(\bot)} = 1 \qquad \forall 1 \le q \le |\mathcal{U}|.$$

(g) Last, we set the transition probability $p_{R(\bot),R(\bot)}$ to be 1.

$$p_{R(\bot),R(\bot)} = 1$$

Clearly, by construction matrix $\mathbf{P}$ obtained from a SET COVER instance $(\mathcal{U}, \mathcal{S})$ is an isotype transition probability matrix as $\mathbf{P}$ is upper triangular, each entry is non-negative and each row sums to 1. In addition, this reduction takes polynomial time.

To prove hardness, let (T', β') be an optimal solution to the MPTR instance composed of the input tree T, observed isotypes $\mathbf{b}$, and isotype transition probabilities $\mathbf{P}$ corresponding to SET COVER instance $(\mathcal{U}, \mathcal{S})$.

Lemma 3. $\mathrm{CSR}(T', \beta', \mathbf{P}) > 0$ for the refined tree T' and the isotype labeling β' inferred by MPTR.

Proof. We prove this by showing that for any constructed input tree T, observed isotypes $\mathbf{b}$ and isotype transition probabilities $\mathbf{P}$, there exists a refined tree T' and isotype labeling β' such that $\mathrm{CSR}(T', \beta', \mathbf{P}) > 0$. We provide a proof by constructing a refined tree T' with isotype labeling β'. The tree T' will expand the unique polytomous node $\bar{v}_0$ into a chain $\bar{v}'_1 \to \ldots \to \bar{v}'_{|\mathcal{S}|}$. We leave the remaining nodes $v_0, v_1, \ldots, v_{|\mathcal{U}|+1}$ of T unaltered, letting $v'_0, v'_1, \ldots, v'_{|\mathcal{U}|+1}$ denote their corresponding nodes in T'. Next, for each $1 \le q \le |\mathcal{U}|$, we pick a subset S_i such that $u_q \in S_i$, and add edge $(\bar{v}'_i, v'_q)$ in T' and set $\beta'(v'_q) = R(u_q)$. We add the edges $(v'_0, v'_{|\mathcal{U}|+1})$ and $(v'_0, \bar{v}'_1)$. Finally, we set $\beta'(v'_0) = \beta'(v'_{|\mathcal{U}|+1}) = R(\top)$. Clearly all the edges in T' have nonzero isotype transition probabilities, so $\mathrm{CSR}(T', \beta', \mathbf{P}) > 0$. □

Corollary 2. The root v'_0 of T' is labeled by isotype $\top$.

Proof. Due to the presence of leaf $v_{|\mathcal{U}|+1}$ with isotype $b_{|\mathcal{U}|+1} = R(\top)$, the root v'_0 of T' must be labeled by isotype $\beta'(v'_0) = R(\top)$, otherwise there would be a zero-probability edge. □

Corollary 3. No node v' of T' is labeled by isotype $\bot$.

Corollary 4. Each edge (v', v'') of T' has an isotype transition probability of $p_{\beta'(v'),\beta'(v'')} = \epsilon$.

Observe that $\bar{v}_0$ is the only polytomous node in T. We will now prove that $\bar{v}_0$ is the only node of T that is expanded in the refined tree T'.

Lemma 4. Node $\bar{v}_0$ is the only node of T that is expanded in T'.

Proof. By Lemma 2, we may assume that β' is transitory. Let v'_0 be the root of T'. We prove this lemma by contradiction. Let $v \neq \bar{v}_0$ be a distinct node of T that is expanded in T'. We distinguish the following three cases.

- $v = v_{|\mathcal{U}|+1}$: In this case, v equals the leaf node $v_{|\mathcal{U}|+1}$ whose parent is the root v_0. Consider the corresponding node $v'_{|\mathcal{U}|+1}$ of T' such that $\sigma(v'_{|\mathcal{U}|+1}) = v_{|\mathcal{U}|+1}$ and $v'_{|\mathcal{U}|+1}$ is a leaf of T'. Since β' is transitory, we have that $\beta'(v'_0) = \beta'(v'_{|\mathcal{U}|+1}) = R(\top)$. Since node $v_{|\mathcal{U}|+1}$ was expanded, node $v'_{|\mathcal{U}|+1}$ has a unique parent $v''_{|\mathcal{U}|+1} \neq v'_0$. As β' is transitory and $\beta'(v'_{|\mathcal{U}|+1}) = R(\top)$ and $R(\top) \leq s$ for all $s \in [r]$, we must have that $\beta'(v''_{|\mathcal{U}|+1}) = R(\top)$. This, however, implies that β' is not transitory as $\sigma(v''_{|\mathcal{U}|+1}) = \sigma(v'_{|\mathcal{U}|+1}) = v_{|\mathcal{U}|+1}$ and $\beta'(v''_{|\mathcal{U}|+1}) = \beta'(v'_{|\mathcal{U}|+1}) = R(\top)$, which yields a contradiction.
- $v \in \{v_1, \ldots, v_{|\mathcal{U}|}\}$: Note that v is a leaf of T. Consider the corresponding node v' of T' such that $\sigma(v') = v$ and v' is a leaf of T'. The parent of v in T is node $\bar{v}_0$. Since node v was expanded, node v' has a unique parent v'' such that $\sigma(v'') = v$. Let v''' be the unique parent of v''. By Corollary 4, we have that the two edges (v'', v') and (v''', v'') both have probabilities ϵ, contributing a factor of 2ϵ to the overall probability $\mathrm{CSR}(T', \beta', \mathbf{P})$. However, by contracting the edge (v'', v') and removing the node v'', we obtain another solution with higher probability, leading to a contradiction.
- $v = v_0$: Consider the corresponding node v'_0 such that $\sigma(v'_0) = v_0$ and v'_0 is the root of T'. There are two cases two consider. Let v''_0 be a child of v'_0 such that $\sigma(v''_0) = v_0$. We distinguish two cases.
  - First, $\beta'(v'_0) = \beta'(v''_0)$. By Corollary 2, we have that $\beta'(v'_0) = \beta'(v''_0) = R(\top)$. By Corollary 4, we have that the edge (v'_0, v''_0) contributes a factor of ϵ to the overall probability $\mathrm{CSR}(T', \beta', \mathbf{P})$. We can remove this factor by simply contracting the edge (v'_0, v''_0), resulting in a more optimal solution, which is a contradiction.
  - Second, $\beta'(v'_0) \neq \beta'(v''_0)$. By Corollary 2, we have that $\beta'(v'_0) = R(\top)$. By Lemma 3, we have $\beta'(v''_0) \in \{R(S_1), \ldots, R(S_{|\mathcal{S}|})\}$. Again, by the same lemma, all children of v''_0 will be labeled by isotypes different than v''_0. In particular, each child of v''_0 will either correspond to node v_0 or $\bar{v}_0$ of T, labeled from the set $\{R(S_1), \ldots, R(S_{|\mathcal{S}|})\} \setminus \{\beta'(v''_0)\}$. Thus, we may contract the edge (v'_0, v''_0), with probability ϵ, and remove the node v''_0, reassigning all children of v''_0 to v'_0. The resulting tree and isotype labeling will have a larger probability, a contradiction.

□

Assume that a series of EXPAND operations on $\bar{v}_0$ in T has generated k nodes in T', where k ranges from 1 (no EXPAND operation) to $|\mathcal{U}|$. We denote $\bar{v}'_1, \ldots, \bar{v}'_k$ to be the new nodes in T' originating from $\bar{v}_0$ in T, i.e., $\sigma(\bar{v}'_1) = \ldots = \sigma(\bar{v}'_k) = \bar{v}_0$. Let $\bar{T}'$ be the subtree of T' induced by nodes $\bar{v}'_1, \ldots, \bar{v}'_k$.

Lemma 5. The refined tree T' has $|\mathcal{U}| + k + 2$ nodes, $|\mathcal{U}| + k + 1$ edges, and $\mathrm{CSR}(T', \beta', \mathbf{P}) = \epsilon^{|\mathcal{U}|+k+1}$.

Proof. Since T has $|\mathcal{U}| + 3$ nodes, and, by Lemma 4, the only node $\bar{v}_0$ of T that is expanded, expands to k nodes $\bar{v}'_1, \ldots, \bar{v}'_k \in V(T')$, the total number of nodes in T' is $|\mathcal{U}| + 2 - 1 + k = |\mathcal{U}| + k + 2$. Similarly, the number of edges in T is $|\mathcal{U}| + 2$, and since $\bar{T}'$ is a tree containing k nodes, it has $k - 1$ edges. So the total number of edges in T' is $|\mathcal{U}| + 2 + k - 1 = |\mathcal{U}| + k + 1$. It follows from Corollary 4 that $\mathrm{CSR}(T', \beta', \mathbf{P}) = \epsilon^{|\mathcal{U}|+k+1}$. □

Lemma 6. Nodes $\bar{v}'_1, \ldots, \bar{v}'_k$ of T' are labeled by k distinct isotypes from the set $\{R(S_1), \ldots, R(S_{|\mathcal{S}|})\}$.

Proof. By construction of $\mathbf{P}$, $R(u_q)$ can only be transitioned into from $R(S_i)$ with nonzero probability where $u_q \in S_i$. So if there is an edge $(\bar{v}'_j, v'_q)$ in T' connecting expanded node $\bar{v}'_j$ with leaf v'_q labeled with $R(u'_q)$ then $\beta'(\bar{v}'_j) = S_i$ for some $S_i \in \mathcal{S}$. Using the observation, we begin by showing that each expanded node $\bar{v}'_i$ has at least one child $v'_q \in L(T')$. We do so by contradiction. Suppose the refined tree T' has an expanded node $\bar{v}'_i$ that does not have any leaf $v'_q \in L(T')$ as a child. Without loss of generality, assume that $\bar{v}'_i$ has a child $\bar{v}''_i$, which, in turn, is the parent of a leaf $v'_q \in L(T')$. This means that $\bar{v}''_i$ is labeled with $\beta'(\bar{v}''_i) = R(S_i)$ for some $S_i \in \mathcal{S}$. Since $R(S_i)$ can only be transitioned into from $R(S_j)$, where $j < i$, or $R(\top)$ with nonzero probability, it holds that $\beta(\bar{v}'_i)$ is either $R(S_j)$ where $j < i$ or $R(\top)$. Similarly, the parent of $\bar{v}'_i$ should also be labeled either with $R(S_{j'})$ where $j' < j$ or $R(\top)$. Now we create a new tree T'' by (i) adding the children of $\bar{v}'_i$ as the children of the parent of $\bar{v}'_i$, and (ii) deleting the edge between $\bar{v}'_i$ and its parent. Clearly T'' has nonzero transition probabilities on all the edges, but has one fewer edge than T'. So $\mathrm{CSR}(T'', \beta', \mathbf{P}) < \mathrm{CSR}(T', \beta', \mathbf{P})$, which contradicts with the premise that T' minimizes $\mathrm{CSR}(T', \beta', \mathbf{P})$. So each expanded node $\bar{v}'_j$ is labeled with $R(S_i)$ for some $S_i \in \mathcal{S}$.

It remains to show that the k nodes $\bar{v}'_1, \ldots, \bar{v}'_k$ are labeled by k distinct isotypes from the set $\{R(S_1), \ldots, R(S_{|\mathcal{S}|})\}$. To see why, observe that, by construction of $\mathbf{P}$, the incident nodes of each edge among nodes $\bar{v}'_1, \ldots, \bar{v}'_k$ must be labeled by distinct isotypes from the set $\{R(S_1), \ldots, R(S_{|\mathcal{S}|})\}$, as $p_{R(S_i),R(S_i)} = 0$ for all $S_i \in \mathcal{S}$. □

Lemma 7. There exists an minimum set cover of size k if and only if there is an optimal solution (T', β') such that $\mathrm{CSR}(T', \beta', \mathbf{P}) = \epsilon^{|\mathcal{U}|+k+1}$.

Proof. ($\Rightarrow$) Let $\mathcal{C} = \{S^*_1, \ldots, S^*_k\}$ be a set cover of minimum size k. Without loss of generality, we further assume that $R(S^*_i) < R(S^*_{i+1})$ for any $1 \leq i \leq k-1$. Next, we build a refined tree T' with isotype labeling β' by expanding the node $\bar{v}_0 \in V(T)$ to k nodes $\bar{v}'_1, \ldots, \bar{v}'_k \in V(T')$. More specifically, we replace $\bar{v}_0$ with $\bar{v}'_1, \ldots, \bar{v}'_k \in V(T')$ such that (i) v_0 is connected to $\bar{v}'_1$ by an edge, (ii) there is an edge $(\bar{v}'_i, \bar{v}'_{i+1})$ in T' for each $1 \leq i \leq k-1$, (iii) $\bar{v}'_i$ is labeled with $R(S^*_i)$, i.e. $\beta'(\bar{v}'_i) = R(S^*_i)$, and (iv) for each child v_q of $\bar{v}_0$ in T, there exists exactly one edge $(\bar{v}'_i, v_q)$ in T' where $u_q \in S^*_i$. Clearly T' is a refinement of tree T, and all the newly added edges have nonzero transition probabilities ϵ. Hence, $\mathrm{CSR}(T', \beta', \mathbf{P}) = \epsilon^{|\mathcal{U}|+k+1}$.

All that remains to show is that (T', β') is optimal. We show this by contradiction. Let (T'', β'') be an optimal solution such that $\mathrm{CSR}(T'', \beta'', \mathbf{P}) < \mathrm{CSR}(T', \beta', \mathbf{P}) = \epsilon^{|\mathcal{U}|+k+1}$. By Lemma 4, we have that only the node $\bar{v}_0$ of T is expanded in T'' corresponding $\bar{v}''_1, \ldots, \bar{v}''_{k'}$ nodes in T''. Since $\mathrm{CSR}(T'', \beta'', \mathbf{P}) < \mathrm{CSR}(T', \beta', \mathbf{P})$, it must hold that $k'' < k$. By Lemma 6 we have that the k' labels of nodes $\bar{v}''_1, \ldots, \bar{v}''_{k'}$ correspond to k' distinct subsets of $\mathcal{S}$. By Lemma 3, we have that these k' subsets of $\mathcal{S}$ form a cover of the universe $\mathcal{U}$, leading to a contradiction. Hence, (T', β') is optimal.

($\Leftarrow$) Now assume that there exists an optimal solution (T', β') such that $\mathrm{CSR}(T', \beta', \mathbf{P}) = \epsilon^{|\mathcal{U}|+k+1}$. Note that the restriction that $\mathrm{CSR}(T', \beta', \mathbf{P}) = \epsilon^{|\mathcal{U}|+k+1}$ is without loss of generality due to Lemma 5. Now according to Lemma 6, there are k expanded nodes in T' labeled with $R(S^*_1), \ldots, R(S^*_k)$. We define $\mathcal{C} = \{S^*_1, \ldots, S^*_k\}$. Now each leaf $v'_q \in L(T')$ labeled with $R(u_q)$ is the child of an expanded node $\bar{v}'_i \in V(T')$ labeled with $R(S^*_i)$. Since $\mathrm{CSR}(T', \beta', \mathbf{P}) > 0$ by Lemma 3, the transition probability from $R(S^*_i)$ to $R(u_q)$ is strictly greater than 0, which means $u_q \in S^*_i$. So every element in $\mathcal{U}$ is covered by one of the subsets from $\mathcal{C}$. So $\mathcal{C}$ is a set cover of size k.

It remains to show that $\mathcal{C}$ is a minimum-size set cover. Assume for a contradiction that there exists a cover $\mathcal{C}' =\subseteq \mathcal{S}$ such that $|\mathcal{C}'| = k' < k = |\mathcal{C}|$. Let $\mathcal{C}' = \{C'_1, \ldots, C'_{k'}\}$ where the subsets follow the same order as in the original reduction to MPTR. We construct a refined tree T'' with isotype labeling β'' corresponding to $\mathcal{C}'$ by expanding the unique polytomous node $\bar{v}_0$ of T into a chain $\bar{v}''_1 \to \ldots \to \bar{v}''_{k'}$, with one node $\bar{v}''_i$ for each subset $C_i \in \mathcal{C}'$ labeled by $\beta''(\bar{v}''_i) = R(C_i)$, and connecting each leaf $v_q \in \{v_1, \ldots, v_{|\mathcal{U}|}\}$ to a single expanded node $\bar{v}''_i$ such that $u_q \in C'_i$. Since $\mathcal{C}'$ is a cover of $\mathcal{U}$, each leaf $v_q \in \{v_1, \ldots, v_{|\mathcal{U}|}\}$ will be connected. Moreover, tree T'' with isotype labeling β'' form a solution to MPTR. Clearly, T'' has $|\mathcal{U}| + k' + 2$ nodes and $|\mathcal{U}| + k' + 1$ edges. Moreover, each edge of T'' has a nonzero isotype transition probability equal to ϵ, so $\mathrm{CSR}(T'', \beta'', \mathbf{P}) = \epsilon^{|\mathcal{U}|+k'+1} < \epsilon^{|\mathcal{U}|+k+1} = \mathrm{CSR}(T', \beta', \mathbf{P})$, a contradiction. □

C Supplementary Methods

C.1 Tree refinement

We solve an instance $(T, \mathbf{b}, \mathbf{P})$ of the MPTR problem (Fig. S2a) by reducing it to a graph problem. Given an instance $(T, \mathbf{b}, \mathbf{P})$ of the MPTR, we construct a directed graph $G_{T,\mathbf{b}}$, called the expansion graph, with nodes $V(G_{T,\mathbf{b}}) \subseteq$

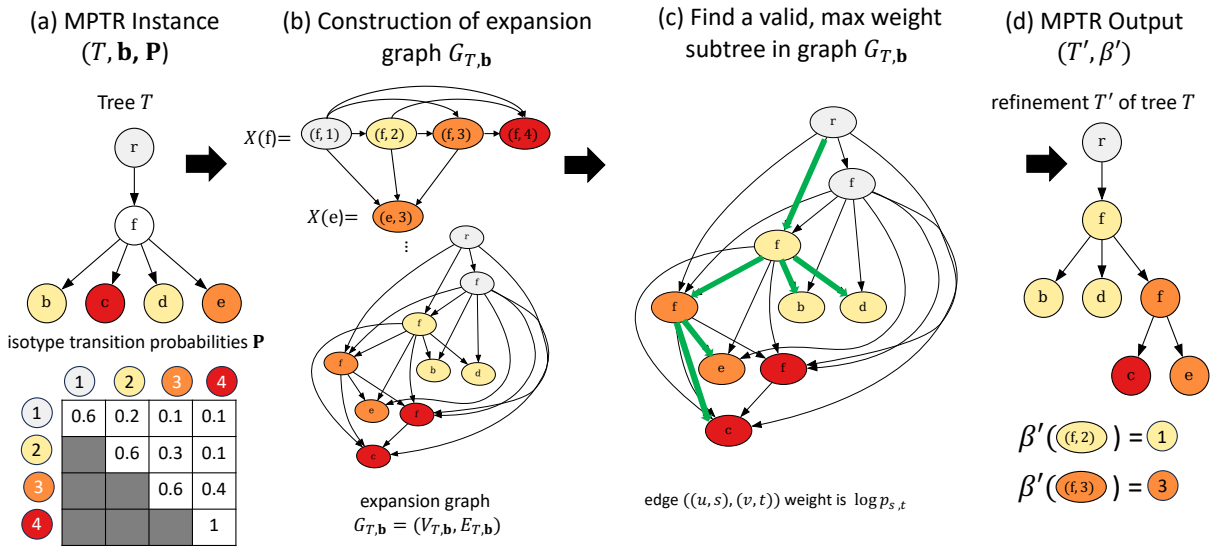


Figure S2: **Algorithm for solving the MPTR problem** (a) An instance $(T, \mathbf{b}, \mathbf{P})$ of the MPTR problem. (b) To construct the expansion graph $G_{T,\mathbf{b}}$ for tree T whose leaves have isotypes $\mathbf{b}$, each original node u in $V(T)$ corresponds to a set $X(u)$ of nodes in $G_{T,\mathbf{b}}$. Edges are added to capture all transitory refinements of tree T. (c) We use the expansion graph $G_{T,\mathbf{b}}$ with weighted edges to find a valid, maximum weight subtree in $G_{T,\mathbf{b}}$, depicted in green. (d) This selected subtree is an optimal solution (T', β') to the MPTR problem instance $(T, \mathbf{b}, \mathbf{P})$.

$V(T) \times [r]$ and edges $E(G_{T,\mathbf{b}})$. At a high level, nodes of $V(G_{T,\mathbf{b}})$ are of the form (u, s) where $u \in V(T)$ is a node of the input tree T and $s \in [r]$ is an isotype state. Formally, we have the following definition.

Definition 4. A directed graph $G_{T,\mathbf{b}}$ is an *expansion graph* of a rooted tree T whose leaves are labeled by isotypes $\mathbf{b}$ provided $V(G_{T,\mathbf{b}}) = \bigcup_{u \in V(T)} X(u)$ where

$$X(u) = \begin{cases} \{(u, b_u)\}, & \text{if } u \in L(T), \\ \{(u, s) \mid s \in \{1, \ldots, \max\{b_v \mid v \in L(T_u)\}\}\}, & \text{if } u \in V(T) \setminus L(T), \end{cases} \tag{8}$$

and $E(G_{T,\mathbf{b}}) = \{((u, s), (v, t)) \mid (u, v) \in E(T), s \leq t\} \cup \{((u, s), (u, t)) \mid u \in V(T), s < t\}$.

In the above definition $X(u)$ is the set of nodes of $G_{T,\mathbf{b}}$ corresponding to node u of T, accounting for the fact that leaves u of T retain their isotype state in any refinement T' of T. On the other hand, internal nodes u of T may be subject to EXPAND operations such that the corresponding nodes of T' are assigned isotypes s ranging from state 1 to the maximum isotype state among all descendant leaves of u in T. The edges of $G_{T,\mathbf{b}}$ respect the irreversibility property of isotypes as well as the parental relationships of nodes of T. See Fig. S2c for an example expansion graph $G_{T,\mathbf{b}}$.

We now define constrained subtrees, termed valid, of the expansion graph $G_{T,\mathbf{b}}$.

Definition 5. A subtree T' of $G_{T,\mathbf{b}}$ is *valid* provided (i) T' is rooted at $(v_0, 1)$ where v_0 is the root of T and (ii) there is a unique edge $((u, s), (v, t))$ in $E(T')$ for each edge (u, v) of T.

We now show that the set of valid subtrees of $G_{T,\mathbf{b}}$ corresponding to trees T' with isotype labelings β' is equivalent to the set composed of pairs (T', β') where T' is a refinement of T and β' is a transitory isotype labeling of T'.

Lemma 8. Let T' be a refinement of T whose leaves are labeled by isotypes $\mathbf{b}$ and let β' be an isotype labeling of T'. Then, β' is transitory if and only if (T', β') induces a valid subtree of $G_{T,\mathbf{b}}$.

Proof. ($\Rightarrow$) Let β' be a transitory isotype labeling of T'. We start by showing that (T', β') induce a connected subtree of $G_{T,\mathbf{b}}$. First, let u' be a node of T' labeled by isotype $\beta(u')$. We claim that $(u', \beta(u')) \in X(u)$. We distinguish the two cases. First, $u' \in L(T')$. Let $u = \sigma(u')$ be the original leaf node u of T. Since β' is transitory, we have $\beta(u') = b_{\sigma(u')} = b_u$. Hence, $(u', \beta(u')) \in X(u)$ for each leaf node $u' \in L(T')$. Second, $u' \in V(T') \setminus L(T')$. Let

$u = \sigma(u')$ be the original internal node u of T. Suppose for a contradiction $(u', \beta'(u')) \notin X(u)$). This means that $\beta'(u') > \max\{b_v \in L(T_u)\}$. As such, there would be an edge (u'', v'') such that $\beta'(u'') > \beta'(v'')$ where u'' is a node in the subtree $T'_{u'}$ rooted at node u'. However, this would mean that β' would violate condition (iii) of Definition 3, a contradiction. Thus, $(u', \beta(u')) \in X(u)$ for each internal node $u' \in V(T' \setminus L(T')$. Hence, $(u', \beta(u')) \in V(G_{T,\mathbf{b}})$.

We now prove that each edge (u', v') of T' whose incident nodes are labeled by $(\beta'(u'), \beta'(v'))$ corresponds to an edge $((u', \beta'(u')), (v', \beta'(v')))$ of $G_{T,\mathbf{b}}$. This follows directly from conditions (iii) and (iv) of Definition 3 and the definition of $E(G_{T,\mathbf{b}})$ in Definition 4. This implies that the subgraph of $G_{T,\mathbf{b}}$ induced by (T', β') is a (connected) subtree of $G_{T,\mathbf{b}}$.

We now must show that this induced subtree of $G_{T,\mathbf{b}}$ is valid. By condition (i) of Definition 3, we have that $\beta'(v'_0) = 1$ for the root v'_0 of T'. As such, the induced subtree of $G_{T,\mathbf{b}}$ is rooted at $(v'_0, 1)$. Finally, we must show there is a unique edge $((u, s), (v, t))$ in the induced subtree of $G_{T,\mathbf{b}}$ for each original edge (u, v) of T. This follows from the fact that T' is a refinement of T. Thus the subgraph of $G_{T,\mathbf{b}}$ induced by (T', β') is a valid subtree of $G_{T,\mathbf{b}}$.

($\Leftarrow$) Consider a valid subtree of $G_{T,\mathbf{b}}$, resulting in a tree T' and isotype labeling β'. To see why T' is a refinement of T, observe that edges $((u, s), (u, t))$ correspond to an EXPAND operation on node u of T. It remains to show that β' is transitory. By condition (i) of Definition 5, we have that the root of T' is labeled by state 1, satisfying condition (i) of Definition 3. Conditions (ii) and (iii) of Definition 3 are met by construction of $G_{T,\mathbf{b}}$. Finally, condition (iv) of Definition 3 follows from condition (ii) of Definition 5. Hence, the isotype labeling β' of T' is transitory. □

The following key proposition follows from the previous two lemmas.

Proposition 2. Let $G_{T,\mathbf{b}}$ be an expansion graph of a rooted tree T whose leaves are labeled by isotypes $\mathbf{b}$. Then, given isotype transition probablities $\mathbf{P}$, a valid subtree (T', β') of $G_{T,\mathbf{b}}$ maximizing $\sum_{(u',v') \in E(T')} \log p_{\beta'(u'),\beta'(v')}$ is an optimal solution to MPTR instance $(T, \mathbf{b}, \mathbf{P})$.

To find such a valid subtree with maximum log-likelihood, we formulate the following MILP based on a multi-commodity flow formulation for modeling connectivity. We make use of two sets of decision variables. The first is $f^t_{(u,s),(v,t)} \in \mathbb{R}_{\geq 0}$, which represents the amount of flow on edge (u, v) designated for sink $q \in L(T)$. The second is $x_{(u,s),(v,t)} \in \{0, 1\}$, which indicates if edge (u, v) has non-zero flow.

$$\min \sum_{((u,s),(v,t)) \in E(G_{T,\mathbf{b}})} x_{(u,s),(v,t)} \log p_{s,t} \tag{9}$$

s.t.

$$\sum_{(v,t) \in \eta^+((u,s))} f^q_{(u,s),(v,t)} = \sum_{(v,t) \in \eta^-((u,s))} f^q_{(v,t),(u,s)}, \qquad \forall (v,s) \in V(G_{T,\mathbf{b}}) \setminus \{(v_0, 1)\}, v \notin L(T), q \in L(T), \tag{10}$$

$$\sum_{(u,s) \in \eta^-((q,b_q))} f^q_{(u,s),(q,b_q)} = 1, \qquad \forall q \in L(T), \tag{11}$$

$$\sum_{(v,t) \in \eta^+((v_0,1))} f^q_{(v_0,1),(v,t)} = 1, \qquad \forall q \in L(T), \tag{12}$$

$$f^q_{(u,s),(v,t)} \leq x_{(u,s),(v,t)}, \qquad \forall q \in L(T), ((u,s),(v,t)) \in E(G_{T,\mathbf{b}}), \tag{13}$$

$$\sum_{(u,s) \in X(u)} \sum_{(v,t) \in X(v)} x_{(u,s),(v,t)} = 1, \qquad \forall (u,v) \in E(T), \tag{14}$$

$$0 \leq f^q_{(u,s),(v,t)} \leq 1, \qquad \forall q \in L(T), ((u,s),(v,t)) \in E(G_{T,\mathbf{b}}), \tag{15}$$

$$x_{(u,s),(v,t)} \in \{0, 1\}, \qquad \forall ((u,s),(v,t)) \in E(G_{T,\mathbf{b}}), \tag{16}$$

where $\eta^+((u, s))$ is the set of direct successors of node (u, s) in graph $E(G_{T,\mathbf{b}})$ and $\eta^-((u, s))$ is the set of direct predecessors of node (u, s).

Constraints (10), (11), (12) enforce flow conversation and ensure that each terminal receives one unit of flow. Below is a description of each of the above constraints. Constraint (13) links the flow variables to the choice of edges in the resulting refinement. Finally, constraint (14) ensures that refined tree T' can be obtained from tree T via a series of EXPAND operations.

C.2 Maximum likelihood estimate of isotype transition probabilities

Given a forest $T_1, \ldots, T_k$ of lineage trees correspondingly labeled by isotypes $\beta_1, \ldots, \beta_k$, we seek the maximum likelihood estimate of isotype transition probabilities $\mathbf{P}^*$.

$$\mathbf{P}^* = \arg\max_{\mathbf{P}} \prod_{(s,t)\in[r]\times[r]} p_{s,t}^{\sum_{j=1}^{k} N_{j,s,t}} \tag{17}$$

subject to

$$\sum_{t\in[r]} p_{s,t} = 1, \qquad \forall s \in [r]. \tag{18}$$

We solve this constrained optimization problem using Lagrange multipliers λ_s for each state s. We first take the log of likelihood $\prod_{j=1}^{k} \mathrm{CSR}(T_j, \beta_j, \mathbf{P})$ with respect to isotype transition probabilities $\mathbf{P}$.

$$\begin{aligned} \log \prod_{j=1}^{k} \mathrm{CSR}(T_j, \beta_j, \mathbf{P}) &= \log \prod_{(s,t)\in[r]\times[r]} p_{s,t}^{\sum_{j=1}^{k} N_{j,s,t}} \\ &= \sum_{(s,t)\in[r]\times[r]} \left(\sum_{j=1}^{k} N_{j,s,t} \right) \log p_{s,t}. \end{aligned} \tag{19}$$

To our log-likelihood, we add the term $\lambda_s \left(\sum_{s\in[r]} p_{s,t} - 1 \right)$ for each isotype s, resulting in new objective

$$\mathcal{L}(\mathbf{P}, \lambda_1, \ldots, \lambda_r) = \left[\sum_{(s,t)\in[r]\times[r]} \left(\sum_{j=1}^{k} N_{j,s,t} \right) \log p_{s,t} + \sum_{s\in[r]} \lambda_s \left(\sum_{t\in[r]} p_{s,t} - 1 \right) \right] \tag{20}$$

Then, we set the partial derivative of $\mathcal{L}(\mathbf{P}, \lambda_1, \ldots, \lambda_r)$ with respect to each parameter $p_{s,t}$ and λ_s and solve the resulting system of equations. For each λ_s, we obtain our constraint,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \lambda_s} = 0 &= \left(\sum_{t\in[r]} p_{s,t} - 1 \right) \\ \sum_{t\in[r]} p_{s,t} &= 1 \end{aligned} \tag{21}$$

For each parameter $p_{s,t}$, we set the partial derivative to 0 and solve for $p_{s,t}$ as a function of λ_s.

$$\begin{aligned} \frac{\partial \mathcal{L}}{p_{s,t}} = 0 &= \frac{\sum_{j=1}^{k} N_{j,s,t}}{p_{s,t}} - \lambda_s \\ \lambda_s &= \frac{\sum_{j=1}^{k} N_{j,s,t}}{p_{s,t}} \\ p_{s,t} &= \frac{\sum_{j=1}^{k} N_{j,s,t}}{\lambda_s} \end{aligned}$$

Given the constraint (21), we have that

$$\sum_{t\in[r]} p_{s,t} = \frac{\sum_{t\in[r]} \sum_{j=1}^{k} N_{j,s,t}}{\lambda_s} = 1, \tag{22}$$

and

$$\lambda_s = \sum_{t \in [r]} \sum_{j=1}^{k} N_{j,s,t}.$$

This yields the following maximum likelihood estimate $p^*_{s,t}$,

$$p^*_{s,t} = \frac{\sum_{j=1}^{k} N_{j,s,t}}{\sum_{t \in [r]} \sum_{j=1}^{k} N_{j,s,t}}$$

Lastly, to account for unobserved isotype transitions where isotype $s \leq t$, we add a pseudocount of 1, resulting in updated isotype transition probabilities

$$p^*_{s,t} = \frac{\sum_{j=1}^{k} N_{j,s,t} + 1}{\sum_{t \in [r]} \left(\sum_{j=1}^{k} N_{j,s,t} + 1 \right)}. \tag{23}$$

In the main text, we additionally use the shorthand $N_{s,t} = \sum_{j=1}^{k} N_{j,s,t}$.

D Simulation details

We designed *in silico* experiments to evaluate TRIBAL with known ground-truth isotype transition probabilities $\mathbf{P}$ and lineage trees T labeled by sequences α and isotypes β. Specifically, we used an existing BCR phylogenetic simulator [13] that models SHM)(Appendix D.1 but not CSR. We generated isotype transition probabilities $\mathbf{P}$ with $r = 7$ isotypes (as in mice) under two different models of CSR (Appendix D.2). Briefly, both CSR models assume the probability of not transitioning is higher than the probability of transitioning, but in the *sequential model* there is clear preference for transitions to the next contiguous isotype, while in the *direct model* the probabilities of contiguous and non-contiguous class are similar (Fig. S3) Given $\mathbf{P}$, we evolved isotype characters down each ground truth lineage tree T.

We generated 5 replications of each CSR model for $k = 75$ clonotypes and $n \in \{35, 65\}$ cells per clonotype, resulting in 20 *in silico* experiments, yielding a total of 1500 ground truth lineage trees. We generated our *in silico* experiments to evaluate all aspects of TRIBAL while benchmarking against existing methods including dnapars [7], dnaml [7] and IgPhyML [10].

D.1 SHM simulation and benchmarking

The Davidsen and Matsen SHM simulator models the generation of B cell lineage trees via a Poisson branching process with selection towards BCRs with increased affinity [13]. We used the provided Docker Hub image container [1] to generate our ground truth B cell lineage trees T and sequence labels α. In addition, we used the provided benchmarking pipeline to run dnapars [7], dnaml [7] and IgPhyML [10]. Below is the command to generate our *in silico* experiments for $n \in \{35, 65\}$ cells and $k = 75$ clonotypes and run comparison methods.

```
simulate
    --igphyml
    --dnapars
    --dnaml
    --selection
    --target_dist=5
    --target_count=100
    --carry_cap=1000
    --T=35
    --lambda=2.0
    --lambda0=0.365
    --n={n}
    --nsim={k}
    --random_naive=sequence_data/AbPair_naive_seqs.fa
```

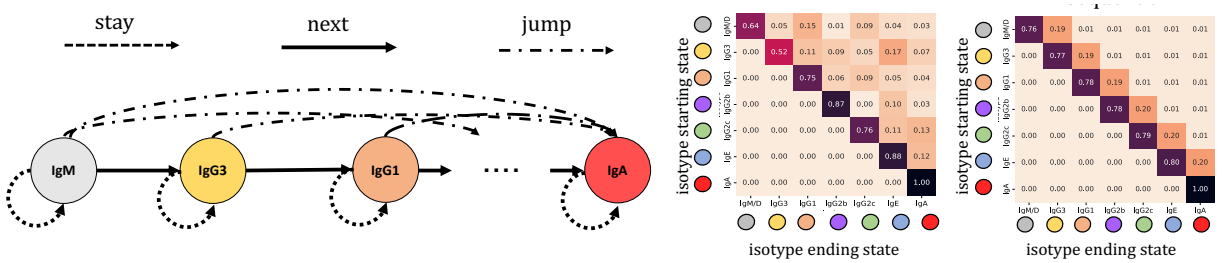
D.2 CSR simulation

After generating each ground truth B cell lineage tree T as described above, we then evolved isotype characters down each tree T using two different models for class switch recombination to obtain ground truth isotypes β. First, we describe the two different CSR models that we used to generate ground truth isotype transition probabilities $\mathbf{P}$. Then, we describe the generation of these isotype transition probability matrices under these two models.

We grouped each isotype transition probability $p_{s,t}$ where $s <= t$ into one of three categories: (i) *stay*, (ii) *next*, and (iii) *jump*(Fig S3a). In *stay*, the B cell does not undergo any class switching and the isotype does not change. In *next*, a B cell class switches to the next contiguous heavy chain locus. In *jump*, the B cell class switches by jumping to an isotype heavy chain constant locus that is not contiguous.

Next, we describe how we generated ground truth isotype transition probabilities $\mathbf{P}$ under both direct and sequential CSR models. To simulate isotype transition probabilities with direct switching, we randomly sampled a probability of transitioning $1 - \theta \in \{0.1, 0.15, \ldots, 0.35\}$. We then set the initial isotype transition probabilities as

[1] krdav/bcr-phylo-benchmark



(a) class switch category (b) direct model (c) sequential model

Figure S3: **Class switch recombination models for *in silico* experiments**. a) Examples of different isotype transition probability parameter groups. (b) Examples of simulated isotype transition probabilities $\mathbf{P}$ for the direct model of CSR. In the direct model, when a B cell class switches is no systematic preference for transition to the *next* sequential state or *jumping* to a non-contiguous isotype. (c) In the sequential model, a B cell undergoing CSR has a strong affinity for the *next* contiguous heavy chain locus.

$$p'_{s,t} = \begin{cases} 0, & \text{if } s > t, \\ \min(\theta + \epsilon, \tau), & \text{if } s = t, \\ \min(\frac{1-\theta}{r-s} + \epsilon, \tau), & \text{if } s < t, \end{cases}$$

where we add Gaussian noise $\epsilon \sim \mathcal{N}(\mu, \sigma)$ with mean $\mu = 0.05$ and standard deviation $\sigma = 0.025$ to each parameter. To avoid negative transition probabilities we set $\tau = 0.01$. Fig. S3b shows an example of a simulated isotype transition probability matrix under the direct CSR model.

$$p'_{s,t} = \begin{cases} 0, & \text{if } s > t, \\ \min(\theta + \epsilon, \tau), & \text{if } s = t, \\ \min(1 - \theta + \epsilon, \tau), & \text{if } t = s + 1, \\ \tau, & \text{otherwise.} \end{cases}$$

We then set parameter $p'_{s,t} := p'_{s,t} / \sum_{s \in [r]} p'_{s,t}$ to ensure each row in the isotype transition probability matrix $\mathbf{P}$ sums to 1. Fig. S3b shows an example of a simulated isotype transition probability matrix under the direct model. Fig. S3c shows an example of a simulated isotype transition probability matrix under the sequential CSR model.

D.3 Inference using TRIBAL

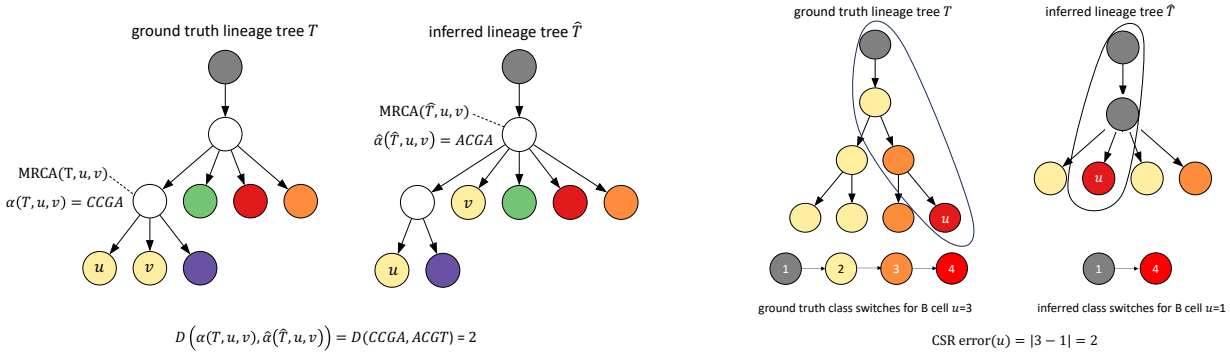
We ran TRIBAL in two ways, referred to as TRIBAL and TRIBAL-NO REFINEMENT, in order to assess the importance of the tree refinement stage of our algorithm. As the naming convention implies, the main difference between TRIBAL and TRIBAL-NO REFINEMENT, is that in TRIBAL-NO REFINEMENT the input trees are not refined and the isotypes $\hat{\beta}$ are inferred using the Sankoff [29] algorithm with weights $w_{s,t}^{(\ell)} = -\log p_{s,t}^{(\ell)}$. All other steps of TRIBAL algorithm remain the same.

Due to large input sets $\mathcal{T}_j$ for some simulated clonotypes j, we sample 50 trees from $\mathcal{T}_j$ for consideration of candidate lineage tree $T_j^{(\ell)}$ within each iteration ℓ. We additionally include the previous optimal lineage tree $T^{(\ell-1)}$ of iteration $\ell - 1$ in the sampled trees for each clonotype j to ensure convergence.

We used a convergence threshold of 0.5 and a maximum of 10 iterations per restart. A total of 5 restarts were performed by iterating through $\theta \in \{0.55, 0.65, 0.75, 0.85, 0.95\}$ for each restart.

D.4 Performance metrics

To evaluate performance of isotype transition probability inference, $\hat{\mathbf{P}}$, we utilized *Kullback-Leibler (KL) divergence*. To assess accuracy of lineage tree inference, we used normalized Robinson-Foulds (RF) distance to assess accuracy of the topology of the inferred tree $\hat{T}$, most recent common ancestor (MRCA) distance to assess accuracy of the inferred sequences $\hat{\alpha}$, and Class Switch Recombination (CSR) error to assess accuracy of the inferred isotypes $\hat{\beta}$.



(a) MRCA distance

(b) CSR error

Figure S4: **Performance metrics for B cell lineage tree inference.** (a) An example calculation for MRCA distance leaves u and v. (b) An example calculation of CSR error for lineage u.

Kullback-Leibler (KL) divergence. To evaluate accuracy of isotype transition probability inference, we used *Kullback–Leibler (KL) divergence* [27] to compare the inferred transition probability distribution $\hat{\mathbf{p}}_s$ of each isotype s to the simulated ground truth distribution $\mathbf{p}_s$. KL divergence D_{KL} is defined as

$$D_{\mathrm{KL}}(\hat{\mathbf{p}}_s || \mathbf{p}_s) = \sum_{q \in [r]} \hat{p}_{s,t} \log(\hat{p}_{s,t}/p_{s,t}) \tag{24}$$

The lower the KL divergence, the more similar the two distributions.

Normalized Robinson-Foulds (RF) distance. To assess the accuracy of topology of the inferred B cell lineage tree $\hat{T}$ with respect to simulated ground truth tree T, we used *normalized Robinson-Foulds (RF) distance*. For this metric, we treat both trees as unrooted. For an unrooted tree, if you remove an edge (but not its endpoints), it defines a bipartition of the leaf set [43]. Doing this for every edge in tree T yields a set $B(T)$ of bipartitions. RF distance is defined as the size of the symmetric difference between bipartitions $B(T)$ and $B(\hat{T})$ [28]. We then normalize this by the total number of bipartitions in each tree. Thus, normalized RF is computed as follows

$$\text{normalized RF}(T, \hat{T}) = \frac{|B(T) \triangle B(\hat{T})|}{|B(T)| + |B(\hat{T})|}. \tag{25}$$

Most Recent Common Ancestor (MRCA) distance. To assess the accuracy of the inferred ancestral sequence reconstruction $\hat{\alpha}$ with respect to simulated ground truth α, we used a metric called *Most Recent Common Ancestor (MRCA) distanc*e introduced by Davidsen and Matsen [13]. For any two simulated B cells (leaves), the MRCA distance is the Hamming distance between the MRCA sequences of these two B cells in both the ground truth and inferred lineage trees. This distance is then averaged over all pairs of simulated B cells. A graphical depiction of this metric is show in Fig. S4a.

More formally,

$$\text{MRCA distance}(\alpha, \hat{\alpha}) = \frac{2}{n(n-1)m} \sum_{u,v \in L(T)} D(\hat{\alpha}(\hat{T}, u, v), \alpha(T, u, v)), \tag{26}$$

where in a slight abuse of notation $\alpha(T, u, v)$ is the sequence of the most recent common ancestor (MRCA) of nodes u and v in lineage tree T and m is the length of MSA.

Class switch recombination (CSR) error. We assessed the accuracy of isotype inference by a new metric called *CSR error*, which is computed for each B cell i and clonotype j and is the absolute difference between the number of ground-truth class switches and inferred number of class switches that occurred along its evolutionary path from the root (Fig. S4b). Since dnaml, dnapars and IgPhyML do not infer isotypes for internal nodes, we pair these methods with the Sankoff algorithm [29] using $w_{s,t}$ equals 1 if $s = t$, 0 if $s < t$ and ∞ otherwise.

E Supplementary Results

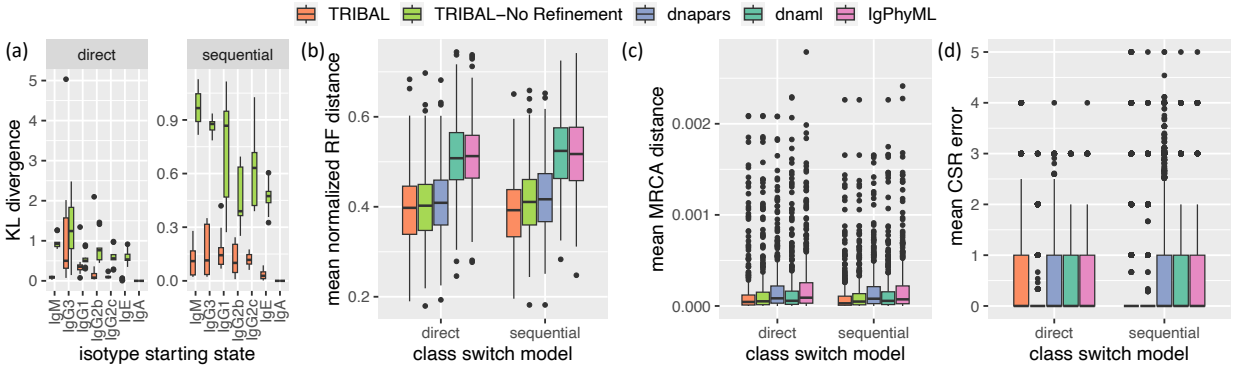


Figure S5: **Simulations results for $k = 75$ clonotypes and $n = 65$ cells per clonotype.**

(a) direct CSR model

(b) sequential CSR model

Figure S6: **KL divergence from ground truth isotype transition probabilities aggregated over all isotype starting states, except IgA, by isotype starting state with varying the number k clonotypes, the number n of cells and CSR model.**

(a) $k = 25$ clonotypes

(b) $k = 50$ clonotypes

Figure S7: **KL divergence from ground truth isotype transition probabilities by isotype starting state for $k \in \{25, 50\}$ clonotypes**

dataset	clonotypes k	total cells n	median cells per clonotype	max cells per clonotype	median distinct isotypes per clonotype
NP-KLH-1	167	1776	7	89	3
NP-KLH-2a	70	537	6	32	2
NP-KLH-2b	58	357	5	21	2

Table S1: **Summary of NP-KLH mouse scRNA-seq datasets**

E.1 Average clade entropy for a leaf labeling

We describe a metric used to assess the average entropy contained within a leaf-labeling of the clades of a tree. First, we introduce some notation. Let Σ be an alphabet. Let clade u of tree T be the subtree T_u rooted at node u. Let $\delta(u) \subseteq L(T)$ be the subset of leaves that are descendants of node u. Let $\ell : L(T) \to \Sigma$ be a leaf labeling. Given a clade u and leaf-labeling ℓ, the entropy of a clade with respect to its leaf labels is defined as

$$H(u, \ell) = -\sum_{s \in [r]} p(s) \log p(s), \tag{27}$$

where $p(s) = \sum_{v \in \delta(u)} \mathbf{1}(\ell(v) = s)/|\delta(u)|$. The average clade entropy $\bar{H}$ is computed over all clades except the leaves $L(T)$ and the root r as follows

$$\bar{H}(T, \ell) = \frac{\sum_{u \in \bar{V}} H(u, \ell)}{|\bar{V}|}, \tag{28}$$

where $\bar{V} = V(T) \setminus (\{r\} \cup L(T))$ is the set of non-trivial clades.

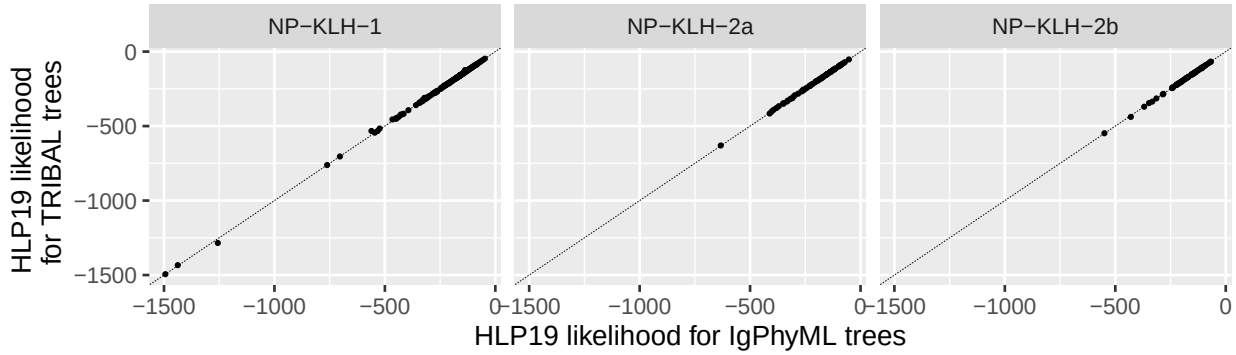


Figure S8: **Comparison of HLP19 likelihood computed for IgPhyML and TRIBAL inferred B cell lineage trees for NP-KLH datasets.**

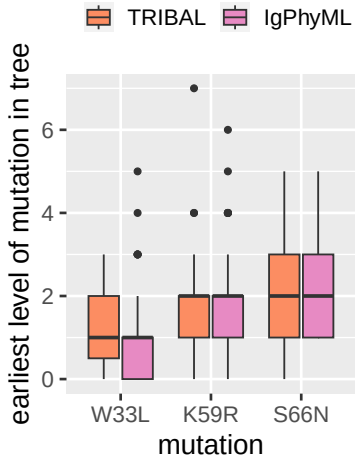


Figure S9: **Earliest observed level of mutation in a B cell lineage tree.** Level 0 represents the MRCA.

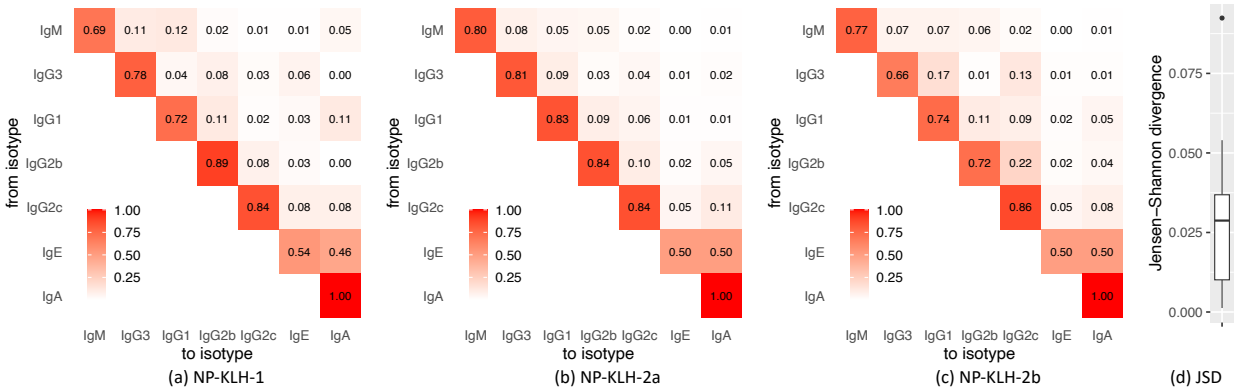


Figure S10: **TRIBAL inferred isotype transition probabilities for NP-KLH.** (a) Isotype transition probabilities for NP-KLH-1. (b) Isotype transition probabilities for NP-KLH-2a. (c) Isotype transition probabilities for NP-KLH-2b. (d) The distribution of Jensen-Shannon divergence (JSD) for pairwise comparisons of rows of the inferred isotype transition probabilty matrices for IgM through Ig2c. IgE was excluded from comparison due to a lack of observed B cells within each dataset to yield informative estimates. IgA is excluded as the inference of this row is trivial.

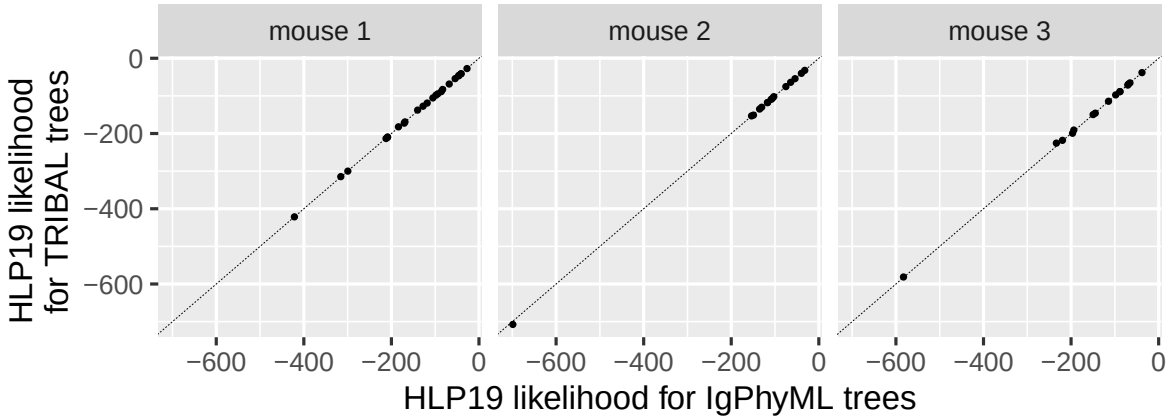


Figure S11: **Scatterplot comparing HLP19 likelihood for IgPhyML trees to the HLP19 likelihood computed for TRIBAL trees for ABC datasets.**

dataset	clonotypes k	total cells n	median cells per clonotype	max cells per clonotype	median distinct isotypes per clonotype
mouse 1	24	224	7.5	31	2
mouse 2	15	218	7	81	3
mouse 3	15	157	7	39	3

Table S2: **Summary of ABC mouse scRNA-seq datasets**

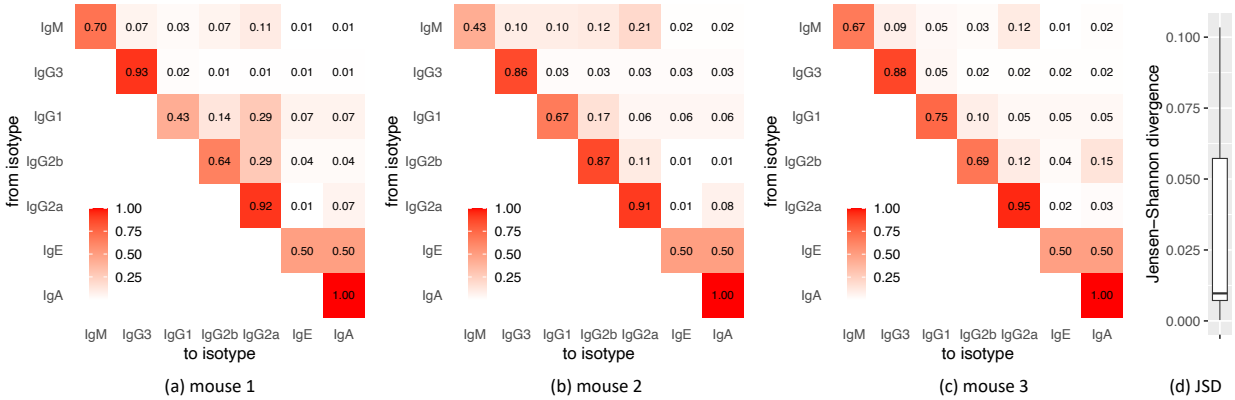


Figure S12: **TRIBAL inferred isotype transition probabilities for ABC datasets.** (a) Isotype transition probabilities for Mouse 1. (b) Isotype transition probabilities for Mouse 2. (c) Isotype transition probabilities for NP-Mouse 3. (d) The distribution of Jensen-Shannon divergence (JSD) for pairwise comparisons of rows of the inferred isotype transition probability matrices for IgM through Ig2c. IgE was excluded from comparison due to a lack of observed B cells within each dataset to yield informative estimates. IgA is excluded as the inference of this row is trivial.