Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section??

Tumor Subclonal Reconstruction with Aldous' Beta-Splitting

Levi Boyles [1,*], Amit Deshwar [2], Quaid Morris [2], Max Welling [3], Yee Whye Teh [1]

[1]Department of Statistics, University of Oxford, Oxford, OX1 3TG, United Kingdom and
[2]Department, Institution, City, Post Code, Country and
[3]Informatics Institute, University of Amsterdam, Amsterdam, 1098 XH, The Netherlands.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: We develop beta-splitting trees, first introduced by Aldous, as Bayesian models for the problem of tumor subclonal reconstruction. Due to the complex discrete-continuous nature of the model, posterior simulation is quite involved and we develop a Markov chain Monte Carlo methodology based on Hamiltonian Monte Carlo, reversible jump, and the Wang-Landau algorithm.
Results: CPABS is shown to provide better reconstructions than other tools in a variety of scenarios.
Availability: Software is available in the CPABS package for the Julia technical computing language: https://github.com/leviboyles/cpabs
Contact: boyles@stats.ox.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

One of the leading models for tumor evolution is the *subclonal* model, in which a tumor is a heterogeneous population comprised of multiple subpopulations (or subclones) with different selective advantages [13, 15, 9]. These subpopulations are characterized by driver mutations, which are in turn associated with cancer development [7, 8], and related to each other via a phylogeny which describes the evolutionary history of the tumor. The task of subclonal reconstruction is to establish the subclonal composition of a tumor from high throughput sequencing data, with the aim of providing insights into potential treatment options [3, 4].

There are a number of existing methods for subclonal reconstruction. (TODO: non Bayesian approaches?) [15] models subclones with a "flat clustering" model – there is no explicit parent/child relationship between subclones defined by the model. [9] provides a model allowing for hierarchical interpretations of the subclonal composition, utilizing Tree Structured Stick Breaking (TSSB) [], a Bayesian prior on trees.

In this work, we explore a model utilizing Beta-Splitting trees [2] as a prior for the subclonal hierarchy. There are many choices for a prior distribution on latent tree structures for Bayesian modelling. Often it is desirable that these prior distributions possess particular properties that make them well suited to the problem at hand. For example, in hierarchical clustering it is often useful to choose a prior that prefers balanced trees over imbalanced ones in order to encourage sharing of statistical strength across clusters. Beta-splitting trees enjoy a number of useful properties, including projectivity, flexibility in modelling both balanced and imbalanced trees, and are related to a number of existing tree models, including the uniform tree distribution and Kingman's coalescent. We introduce consistent time variables for these trees that can capture the inverse relationship between subpopulation size and mutation number. Additionally, there are some computational benefits over other choices of prior such as Kingman's coalescent.

In Section 2, we describe our model for the evolutionary history of a tumor, given a tree describing the shape of the phylogeny. Section 3 reviews our prior on phylogenies, namely Aldous' beta-splitting. Section 4 we review related prior distributions. In Section 5 we describe the inference methodology used, and in Section 6 we compare our model against a state-of-the-art model for subclonal reconstruction. Finally, we conclude in Section 7.

2 Subclonal Reconstruction

Consider the following generative description of tumor evolution (see also Figure 2):

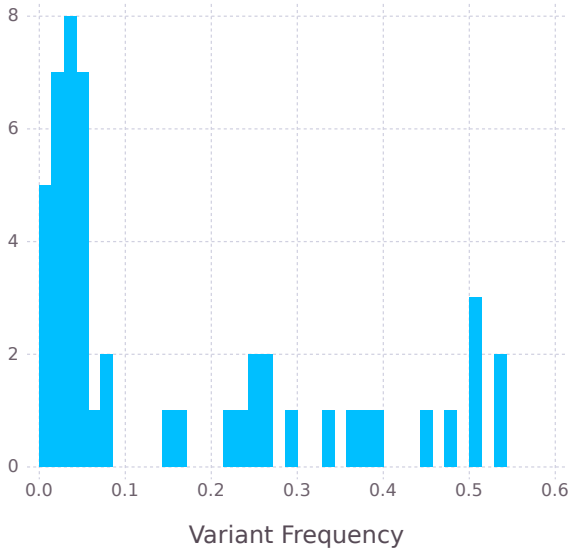

Fig. 1. Variant mutation frequency histogram for an example simulated dataset. The x-axis corresponds to the population frequency for each mutation, specifically we plot b_m/d_m for each mutation. There are five ground truth clones in this dataset, three of which share the low-frequency end of the histogram.

1. Given a large population of cells, any individual cell lineage acquires "driver mutations" – single-site mutations (SSMs) associated with cancer – with mutation rate λ_d, and "passenger mutations" with mutation rate λ_p. Passenger mutations are those introduced by the enhanced growth of the tumor but are selectively neutral.
2. When a cell acquires a driver mutation, the resulting descendent subpopulation becomes a significant proportion of the overall population. All cells in the descendant subpopulation have the driver mutation as well as all passenger mutations associated with that driver.
3. We observe only mutations that are present in entire subpopulations with a significant proportion of the overall population.

The data are gathered from a deep read sequencing of a tumor biopsy, with multiple coverage per site (**better term/phrasing for this?**) . We assume that the only mutations we observe with high enough frequency to be distinguished from noise are driver mutations and passenger mutations who precede a driver mutation on a cell lineage. Additionally, we make an infinite sites assumption: no mutation event will introduce an SSM that is already present elsewhere in the tumor, nor will it remove any SSMs.

For a particular tumor, we have a finite set of M mutations. We may have biopsied the tumor multiple times giving a number of samples S. We observe the number of reads $d_{m,s}$ covering the location of mutation m in sample s. Out of these, $b_{m,s}$ are observed to have the mutation, and thus $a_{m,s} = d_{m,s} - b_{m,s}$ do not.

The aim of the reconstruction task is to infer the set of subpopulations that comprise the tumor and the assignment of mutations to the subpopulations from which they originated. In many models for subclonal reconstruction, the latent subpopulations are related to the data by modelling the *cellular prevalence* of each subpopulation, that is, the proportion of cells that are descendants of that subpopulation. Cellular prevalence is clearly related to the variant counts $b_{m,s}$ – we would expect to see larger variant allele frequencies (VAFs) $b_{m,s}/d_{m,s}$ for mutations associated with larger subpopulations. A key distinction between existing subclonal reconstruction tools is in the way in which the cellular prevalences, hereon denoted ϕ, are modelled. PyClone [] models the cellular prevalences with a Dirichlet process, assuming a flat clustering model for the mutations. PhyloSub [9] introduces a latent hierarchy that

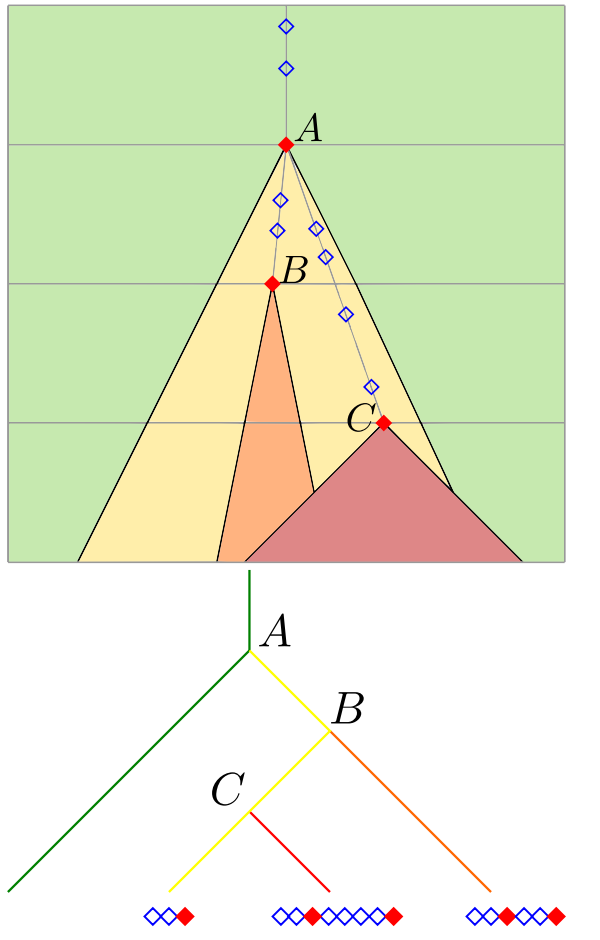

Fig. 2. Assumed process for tumor evolution (best viewed in color). The process begins with a single lineage within a reference population (green area) which acquires passenger mutations (blue diamonds) over time. Eventually, this lineage acquires a driver mutation (red-filled diamonds) that spawns a new subpopulation A that begins to compete with the reference population. This occurs 2 more times, spawning from within the A subpopulation to create two additional subpopulations B and C. This particular outcome corresponds to the binary tree shown below, if right subtrees correspond to newly spawned subpopulations and the left subtrees correspond to old subpopulations. Each colored line corresponds to the path from each population at the leaves to its point of divergence from its parent population. The length of the path from the divergence point to the earliest ancestor that can be reached from a path of a single color is the time available for the population to acquire passenger mutations. The mutations associated with a population are depicted at that population's leaf.

models the nested structure of the subpopulations, requiring that the cellular prevalence of a parent subpopulation is greater than that of the sum of cellular prevalences of its children. In these models, the reference read counts, are modelled given the cellular prevalences. If $\phi_{k,s}$ is the cellular prevalence for subpopulation k in sample s, z_m is the subpopulation to which mutation m belongs, and η^r_m and η^v_m are the probabilities of observing a reference read from the reference and variant populations, respectively[1], then the reference count is modelled by:

$$a_{m,s} \sim \text{Binomial}(d_{m,s}, (1 - \phi_{z_m,s})\eta^r_m + \phi_{z_m,s}\eta^v_m). \quad (1)$$

[1] Typically, $\eta^r \approx 1.0$ and $\eta^v \approx 0.5$, accounting for the possibility of reading the non mutated chromatid in a variant type cell.

2.1 Mutation Model

We extend the latent representation used to model tumor evolution to include time variables in addition to a discrete hierarchy, where there is time t_i associated with each node i in the hierarchy. These times are used to model the number of passenger mutations that occur between driver events. These time variables are not to be interpreted as physical time; the physical times and mutation rates of each subpopulation are not identifiable given the mutation read counts. Instead, time between two nodes can be interpreted as the product of the physical time between the driver events and the mutation rate of the subpopulation spawned from the parent driver event. In other words, the times variables t_i can be interpreted as time rescaled to a globally constant mutation rate.

Let π denote a binary tree, and let $\psi = (\pi, \{t_i\})$ denote the collection of the tree along with the time variables. Beta splitting defines a prior on ψ, and Section 3 defines this prior; we first consider the likelihood model for the data given a draw from this prior. For a binary tree π, we will index the set of nodes by strings in $\{l, r\}^*$, with the root being the empty string ϵ, and for each i in $\{l, r\}^*$ its two children being il and ir.

Note that the subpopulation hierarchy is not necessarily binary, but our prior is over binary trees π. Thus we make a distinction between the binary tree π and the subpopulation hierarchy H (a multifurcating tree,) and relate the two in the following way. A node i in π has the subpopulation $h_i \in H$ associated with it. The right and left subtrees of a node i are treated differently, the right subtree corresponds to the new subpopulation h_i which has acquired a cluster of SSMs and the left subtree corresponds the old subpopulation $h_{\text{parent}(i)}$ without any new SSMs. This is the natural result of applying a binary tree prior to this problem, see Figure 2. Thus, the parent node parent(i) does not necessarily correspond to the parent population of the population associated with node i, that is it is *not* necessary that $h_{\text{parent}(i)} = \text{parent}(h_i)$.

Let $\tau(i)$ be the most recent ancestor of node i from which the subtree containing i is the right subtree. Thus, $\tau(i)$ is the node associated with the parent population of subpopulation i: $h_{\tau(i)} = \text{parent}(h_i)$. Then $v_i = t_{\tau(i)} - t_i$ is the time after which the subclone associated with node i spawned from its parent subpopulation at $\tau(i)$. We assume passenger mutations occur according to a Poisson process, so that under the generative process we have that the number of passenger mutations associated with population i is $u_i \sim \text{Poisson}(\lambda_p v_i)$. However, the total number of passenger mutations is known for a given tree, so the vector of passenger mutation counts u_i is distributed according to a Multinomial:

$$u \sim \text{Multinomial}\left(M, \frac{v}{\sum_i v_i}\right). \tag{2}$$

Let $z_m = i$ if mutation m belongs to population h_i. There are two ways a mutation may have been assigned to a population, either as a driver mutation, or a passenger mutation. Let $K = N - 1$ be the number of driver mutations, then we have

$$p(z_m = i|\psi) = \frac{K}{M}\frac{1}{K} + \frac{M-K}{M}\frac{v_i}{\sum_j v_j} = \frac{1}{M} + \frac{M-K}{M}\frac{v_i}{\sum_j v_j}. \tag{3}$$

2.2 Modelling Subpopulation Sizes and Read Counts

In practice, only a subset of the tumor is biopsied, and this may be performed multiple times giving multiple "samples" of the tumor. Following [13], we model each sample as an admixture of the subclonal populations, with differing mixing proportions for each sample. This reflects the heterogeneity of the composition of each sample.

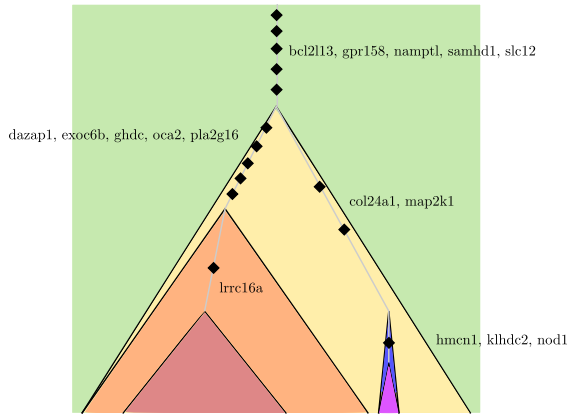

Fig. 3. Representative sample from the posterior for the CLL077 dataset. The width of each subpopulation at the base of the figure reflects the cellular prevalence of that subpopulation. Also shown are the unique mutations associated with each subclone. Time variables are not accurately depicted in this figure.

The SSM population frequency $\phi_{i,s}$ for population i and sample s are constructed as

$$\phi_{i,s} = v_{ir,s} \prod_{j \in \overline{\text{An}}(i)} v_{j,s} \qquad v_{j,s} \sim \text{Beta}(\alpha\nu_j + 1, \alpha(1-\nu_j)+1). \tag{4}$$

where $\overline{\text{An}}(i) = \text{An}(i) \cup \{i\}$. α is a hyperparameter of the model, and $v_{jl,s} = 1 - v_{jr,s}$. Larger values of α will shrink the variance of $v_{j,s}$, so that the proportions for each sample more closely match the beta-splitting proportions. Note that $v_{j,s}$ is uniformly distributed when $\alpha = 0$. Thus, the ϕ variables are more strongly tied across samples than if they were treated as independent variables when $\alpha > 0$. This will still enforce the SSM population constraints, that is, $\phi_{il,s} \leq \phi_{i,s}$ and $\phi_{ir,s} \leq \phi_{i,s}$.

Finally, given ϕ and a number of reads $d_{m,s}$, we draw the number of reference reads $a_{m,s}$ as in (1).

3 Beta-Splitting Trees

Beta-splitting defines an infinite tree-structure in a top-down fashion. Let $f(x)$ be a symmetric density on the unit interval. Begin by placing N points, or individuals, uniformly at random on the unit interval; let y_i denote the location of point i. Then repeat until all points are isolated:

1. Draw $x \sim f$.
2. Individuals with $y_i < x$ are designated to the left branch, all others designated to the right.
3. Recurse on each subinterval (after rescaling each to a unit interval), until all points are assigned their own branches as leaves.

See Figure 4 for an illustration. This process defines an infinite tree if we let $N \to \infty$. [2] specializes f to a one-parameter Beta distribution, so that $x \sim \text{Beta}(\beta + 1, \beta + 1)$. Picking β and marginalizing out x for $-2 \leq \beta < \infty$ (the discrete distribution on trees obtained by integrating out x is still sensible for $-2 \leq \beta \leq -1$) gives rise to many familiar priors, for example $\beta = 0$ corresponds to Kingman's Coalescent prior, and $\beta = -\frac{3}{2}$ corresponds to the uniform distribution on trees.

In the generative process of the beta-splitting tree, it is possible that the split point x may be trivial, that is, that the individual points all lie to the left of x, or all to the right of x. In such a situation the structure of the tree, as defined by the hierarchical separation of the individuals into leaves of the tree, does not change, but the length of the interval containing the

individuals does. Let N_i be the number of individuals at node i, D_i be the number of trivial splits prior to node i, $\tilde{\nu}_i$ be the proportion of the original interval still containing the individual points before the non-trivial split, and μ_i be the resulting length of the interval just before the (non-trivially) split, (see Figure 4 (right).) Finally, let ν_{il} and ν_{ir}, with $\nu_{il} + \nu_{ir} = 1$, be the proportions on the left and right children of i respectively.

In the case where f is a uniform distribution on $[0, 1]$, corresponding to Kingman's coalescent, it turns out to be possible to marginalize out both the N individual locations $\{y_k\}$ and the number of trivial splits $\{D_i\}$ before each node of the tree, and derive an explicit distribution for $\tilde{\nu}_i$. First, note that the marginal probability of failing to split a set of N_i points is:

$$2\int x^{N_i}(1-x)^0 f(x)dx = \frac{2}{N_i+1}. \tag{5}$$

Let $\xi_{N_i} = 1 - \frac{2}{N_i+1}$. Then $D_i \sim \text{Geometric}(\xi_{N_i})$ gives the number of failures until the N_i points are split. The distribution of $\tilde{\nu}_i$ is thus the product of D_i uniform draws. If we take $U_k \sim \text{Uniform}(0, 1)$, and $U^{(d)} = \prod_{k=1}^{d} U_k$, then $-\ln U^{(d)}$ is the sum of d Exponential draws, thus $-\ln U^{(d)} \sim \Gamma(d, 1)$, and so

$$p(-\ln\tilde{\nu}_i = y) = \sum_{d=1}^{\infty}(1-\xi_{N_i})^d\xi_{N_i}\frac{1}{(d-1)!}y^{d-1}e^{-y} + \xi_{N_i}\delta_0(y) \tag{6}$$

$$= \xi_{N_i}\delta_0(y) + (1-\xi_{N_i})\xi_{N_i}e^{-\xi_{N_i}y} \tag{7}$$

which gives

$$\tilde{\nu}_i \sim \xi_{N_i}\delta_1(\tilde{\nu}_i) + (1-\xi_{N_i})\text{Beta}(\xi_{N_i}, 1) \tag{8}$$

that is, $\tilde{\nu}_i$ is a mixture between a point mass at 1 and a $\text{Beta}(\xi_{N_i}, 1)$. This result agrees with intuition: if N_i is large, then ξ_{N_i} is close to 1, and $\tilde{\nu}_i$ is close to 1. ν_i is simply $\text{Uniform}(0, 1)$ and $\nu_{ir} = 1 - \nu_{il}$. If N_i points are split into groups of sizes N_{il} and N_{ir}, then

$$p(\nu_{il}, N_{il}, N_{ir}) = \frac{2}{(N_{il}+N_{ir})(N_{il}+N_{ir}-1)}\nu_{il}^{N_{il}-1}(1-\nu_{il})^{N_{ir}-1}. \tag{9}$$

This form arises from the fact that we are conditioning on a split occurring, that is, $N_{il} > 0$ and $N_{ir} > 0$. Finally, we can write μ_i:

$$\mu_i = \nu_i\tilde{\nu}_i\prod_{j\in\text{An}(i)}\nu_j\tilde{\nu}_j. \tag{10}$$

In summary, we have a finite tree with N leaves and each internal node having two children. At each internal node i we have μ_i, the interval length before it splits, given by (10), where $\tilde{\nu}_i$ is given by (8) and the joint distribution of the proportion of the interval and numbers of leaves of the two children of i are given by (9).

Variables that denote the time from leaf to internal node are often employed in hierarchical clustering models. Typically, the parameters that generate the observations reside at the internal nodes of the tree, with shorter branch lengths corresponding to stronger correlations between parameters. If the time variables have the property that internal nodes that have few descendants are strongly encouraged to have smaller times[2] and thus are "closer" to the leaves, then the clustering model is less likely to be overly flexible and in danger of overfitting. Kingman's Coalescent and the Dirichlet Diffusion Tree [11] both employ time variables with this property.

[2] We take the convention that time runs from leaves to root

The choice of time variables is a delicate one, as any choice should leave the overall distribution consistent (in the Kolmogorov sense). One way to ensure that the times retain a consistent prior distribution is to show that the time to the most recent common ancestor (MRCA) for a pair of nodes is the same in the infinite tree as it is in a finite projection. For internal node i, we may define the time t_i as

$$t_i = g(\mu_i) \tag{11}$$

where g is a monotonically increasing function with $g(0) = 0$. Thus using μ_i to define a time variable gives a consistent prior over times. As $0 \leq \mu_i \leq 1$, choosing a strictly convex g with $g(1) = 1$ will correspond to nodes that are closer to the leaves, and thus to a less flexible prior.

This choice of time variable has two main advantages over the times used for the Coalescent model. First, the time for a node i is dependent only on the number of individuals delegated to each of its children. Coalescent times, on the other hand, are determined by starting with N individuals and recursively joining pairs of with exponential waiting times – the time for node i is directly dependent on the times of many other nodes throughout the tree. Thus, MCMC inference under Beta-Splitting can be performed in a more computationally efficient manner. Second, this construction allows for more choice in the specification of how the times are distributed for nodes that are deeper in the hierarchy, for example the choice of g can be modified to push internal nodes leaf-wards or root-wards. Thus we obtain a prior with the same distribution on tree structures as that in Kingman's Coalescent, but with a more flexible choice as to the distribution on times.

3.1 Number of Clones

The above model is defined for a fixed number of clones (equivalently, for a tree with a fixed number of leaves.) Unlike the TSSB, any two points will eventually be split by beta-splitting. In our model, a point in the beta-splitting process corresponds to an entire population. Thus, in order allow for a variable number of clones, we specify a prior on the K, namely:

$$K \sim \text{NegativeBinomial}(c, \theta). \tag{12}$$

We choose $c = 2$ and $\theta = 0.4$ so that K has a large variance.

4 Related Work

There are many choices for prior distributions over trees which have differing properties. Perhaps the simplest prior is the uniform tree distribution, where each tree with N leaves is given equal probability. This distribution has the property in that the typical draw from the uniform prior is imbalanced [2]. Another is the distribution on trees induced by Kingman's Coalescent [10], in which, given N individuals, pairs are *coalesced* uniformly at random until a full tree is formed. A typical draw from this distribution is balanced. Another interesting prior is the Dirichlet Diffusion Tree (DDT) [11], which defines a joint distribution on tree structures and time variables which cannot be factored into analytically tractable factors. The distribution on the DDT's time variables is specified by a "divergence function" which also influences the balance of the tree.

The Tree-Structured Stick Breaking (TSSB) [1] prior allows data to sit at internal nodes of the tree, thus the size of the tree (its number of nodes) is inferred in addition to the structure. The TSSB has been used in our problem of interest: tumor subclonal reconstruction [9]. However, the TSSB has the tendency to prefer shallow hierarchies with high branching factors [16], a property which is not necessarily ideal for subclonal reconstruction.

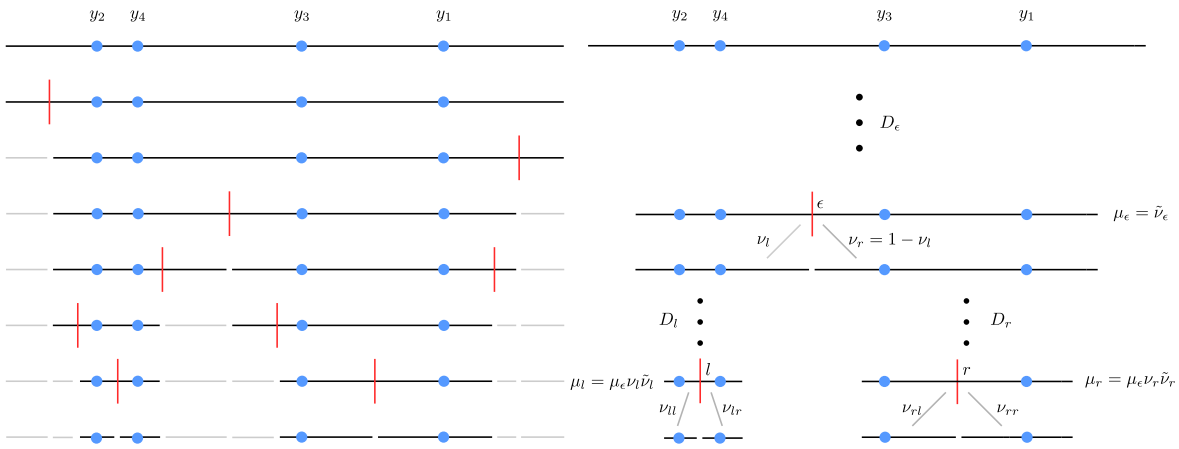

Fig. 4. Aldous' Beta-splitting. (left) Five points are dropped onto the unit interval, which is then recursively split according to beta-splitting. Note that splits may occur that do not separate any data. (right) By marginalizing out the splits that do not separate data, we get a simplified representation for the Beta-Splitting tree using fewer explicit splitting locations. In the figure, beta-splitting failed to separate the four points D_ϵ times, resulting in an interval of length $\tilde{\nu}_\epsilon$. This interval is then split into two subintervals with proportions ν_l and ν_r, and the process recurses on each subinterval. Note that we need not explicitly represent the locations y to sample from this process, as we need only determine to which subinterval each point belongs at each split event.

5 Inference

5.1 Sampling $\nu, \tilde{\nu}, \upsilon$, and z

We give brief descriptions of the inference methods used, for more detail on the updates for $\nu, \tilde{\nu}, \upsilon$, and z, see the supplementary appendix.

The ν variables are continuous variables with nondegenerate densities, and so we update each one in turn with slice sampling [12]. The $\tilde{\nu}$ variables are continuous with an atom at 1 due to the form of the (8). We update each in turn, augmenting the atom with an auxiliary variable giving a nondegenerate density from which we slice sample, projecting to the atom as appropriate (this is a similar approach to that taken in some spike and slab models.) We sample υ according to Hamiltonian Monte Carlo [5], as these variables are strongly tied and it is important to update them jointly. Finally, we update z according to Gibbs sampling.

5.2 Sampling the tree structure ψ

We employ a prune/graft sampling algorithm where we randomly choose a node to prune from the tree, and then sample a new location for the pruned subtree. We hold the ν variables fixed, so grafting will change the times associated with the nodes underneath the graft point.

Sampling ψ can be done in a more computationally efficient manner than in the Coalescent. After pruning a subtree with a number of leaves $n < N$, to evaluate the posterior probability of a potential grafting location we need to evaluate the likelihood and prior for the subtree rooted at the grafting location. For a likelihood that depends directly on the time variables (such as a Brownian motion likelihood), as the ν variables are fixed and the times are defined as products of ν variables over a node's ancestors, the posterior probabilities for all graft locations can be computed in $O(Nn)$ time using dynamic programming. The likelihood defined by (4) doesn't follow this form, however, as the ϕ variables are also defined as products over υ variables over a node's ancestors, we can still evaluate the posterior probability of all graft points in $O(Nn)$ time.

In contrast, the prior distribution for the Coalescent times all change when grafting in a new subtree, giving a $O(N^2)$ cost to evaluate the prior using a similar prune/graft scheme. In our experiments, we found it beneficial to keep $n = 1$, as the improved mixing was not commensurate with the increased computational cost.

5.3 Sampling K

In our application, the expected number of clones K will be much less than the number of observed mutations. This significantly reduces the cost of inference for a tree with a fixed number of leaves. However, the fact that we must also infer the number of leaves complicates inference in a different manner; we must sample over the joint space of the tree ψ, the assignments z, and several continuous variables, while simultaneously inferring the size of the tree. In most hierarchical clustering applications, the number of leaves is fixed to the number of observations.

We employ a grow-prune style reversible jump [6] sampler combined with the Wang-Landau (WL) algorithm [19] to sample the number of clusters. We define the WL partition to partition K into blocks: $\{1 \dots 3, 4, 5, 6 \dots \infty\}$. We found partitioning based on energy level did not improve performance due to the slower convergence of the WL partition functions. WL forces the sampler to visit all blocks; not only does this reduce the prevalence for the sampler to get stuck in a single model, it encourages sample diversity by occasionally causing the MCMC state to change dramatically.

The use of the WL algorithm comes at a cost: WL computes an online estimate of the posterior probability of each block which may take a long time to converge. Also, if one block has much higher estimated posterior probability than the others', then, effectively, only samples from that block will be used for prediction, reducing the number of effective samples of the overall procedure.

6 Experiments

To validate that our model produces reasonable and interpretable results, we first ran our method on the Chronic Lymphocytic Leukemia (CLL) dataset from [14]. Specifically, we considered the data from patient CLL077 and compared our method's results to a manually generated phylogeny by a human expert (provided in [14]). We found that most samples from the posterior look similar to that seen in Figure 3. This sample is nearly identical to the baseline tree structure, excepting (low population) clusters 4 and 5 are merged in the baseline tree.

To provide a quantitative assessment of our model, we compare our model to PhyloSub on simulated data. The data were generated given a tree with a specified number of clones $K \in \{3, 4, 5\}$, a read depth $R \in \{50, 70, 100\}$, and number of unique mutations per clone $M \in \{10, 25, 100\}$, generating 8 independent datasets for each combination

of parameters. The generating process constructs a chain-structured phylogeny of K clones, sampling relative population sizes according to a symmetric low-variance beta distribution. Given the population sizes, the reference counts are generated according to (1). This data-generating process was chosen to encourage the generation of distinguishable clusters, however, the process still generates datasets where the clones are not easily identified, see an example frequency histogram in Figure 1.

We evaluated the two models by comparing the (flat) clusterings on the mutations to the ground truth clusterings, as in [9]. However, [9] evaluates performance using area under the precision recall curve (AUPR) for the resulting co-clustering matrix. We found that the AUPR measure of clustering quality overly penalizes clusterings with smaller numbers of clusters. This preference for either small or large number of clusters is a common problem for cluster evaluation scores that don't account for the score that might have occurred due to chance [18]. Thus, we depart from [9] and compare the two models using Adjusted Mutual Information (AMI) [17] on the flat clustering induced by the sampled hierarchies. AMI is adjusted so that it does not generally prefer simple or complex models.

For a pair of partitions $C^{(1)}$ and $C^{(2)}$, define the distributions $p^{(k)} = C^{(k)}/|C^{(k)}|$. Then the AMI is defined as

$$AMI(C^{(1)}, C^{(2)}) = \frac{I(p^{(1)}, p^{(2)}) - \mathbb{E}_q\left[I(p^{(1)}, p^{(2)})\right]}{\max(H(p^{(1)}), H(p^{(2)})) - \mathbb{E}_q\left[I(p^{(1)}, p^{(2)})\right]} \tag{13}$$

where H is entropy, I is mutual information, and q is a generalized hypergeometric distribution [17].

To compare the two methods, we estimate the expected AMI:

$$\mathbb{E}\left[AMI(C, C^*)\right] \approx \frac{1}{W}\sum_{k=1}^{K} w_k AMI(C^{(k)}, C^*) \tag{14}$$

where $C^{(k)}$ is the clustering associated with the kth MCMC sample, C^* is the ground truth clustering, and $W = \sum_k w_k$. $w_k = Z_{n_k}/L_{n_k}$, where n_k is the WL block to which k belongs, Z_n is the WL partition function estimate associated with block n, and L_n is the number of samples in block n. In the case of PhyloSub, $w_k = 1$.

As seen in Figure 5, our model generally outperforms PhyloSub on AMI. The difference in performance of the two algorithms did not have a strong dependence on read depth or mutation number. However, PhyloSub's performance declines in the $K = 5$ case. Interestingly, this effect does not quickly diminish with the introduction of more data; see Figure 5. This suggests that either the PhyloSub model begins to fit poorly as data complexity increases, or that its sampler fails to mix in more complex settings.

7 Conclusion

We have presented a model for subclonal reconstruction utilizing Aldous' beta-splitting as a prior on tree structures. Marginalizing out the splitting locations that do not separate a finite set of points provides a compact representation for beta-splitting draws; furthermore, consistent time variables can be introduced to produce a projective prior over trees with times. Additionally, we have leveraged the recursive structure of beta-splitting to develop an efficient MCMC sampler. Applied to subclonal reconstruction, beta-splitting improves over the state of the art model.

Acknowledgements

Funding

This work has been supported by the... Text Text Text Text.

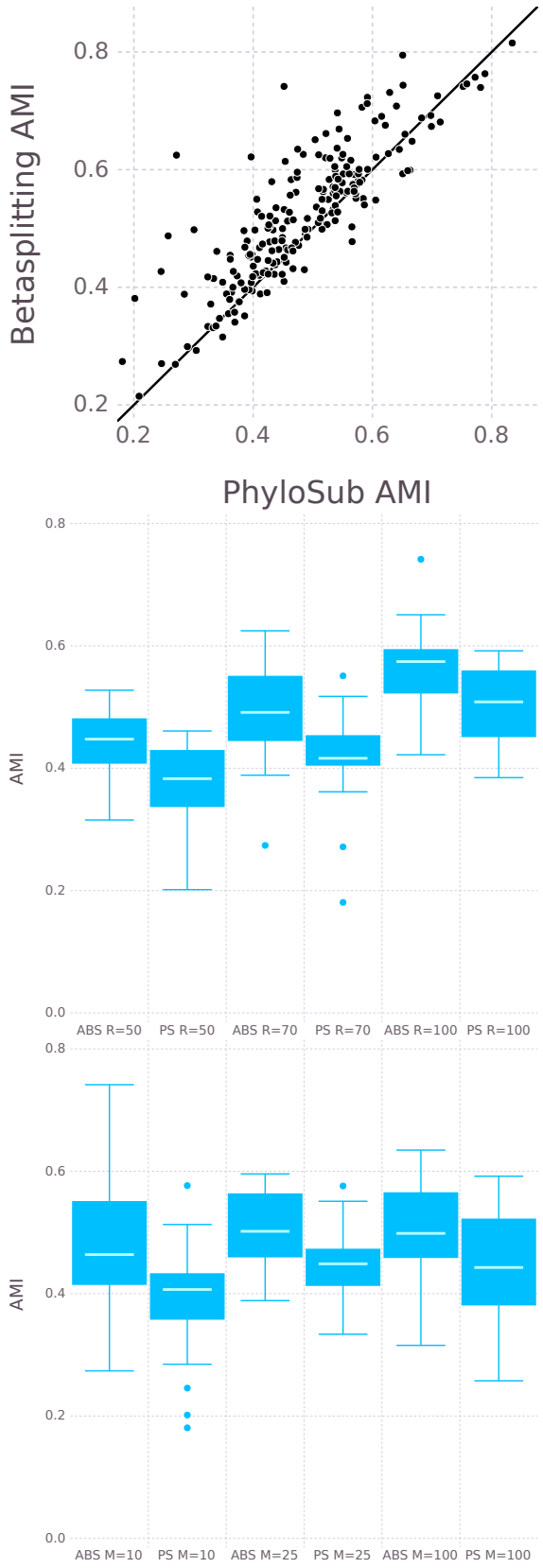

Fig. 5. (left) Scatter plot showing the Adjusted Mutual Information for PhyloSub versus Aldous' Beta-Splitting. (center, right) Effect of varying read depth and mutation number for the $K = 5$ case. "ABS" refers to Aldous' beta-splitting, "PS" refers to PhyloSub.

References

[1]RP Adams, Z Ghahramani, and MI Jordan. Tree-structured stick breaking for hierarchical data. *Advances in Neural Information Processing Systems*, 23:19–27, 2010.

[2]D Aldous. Probability distributions on cladograms. *IMA Volumes in Mathematics and its Applications*, 76, 1996.

[3]S Aparicio and C Caldas. The implications of clonal genome evolution for cancer medicine. *New England Journal of Medicine*, 2013.

[4]PL Bedard, AR Hansen, MJ Ratain, and LL Siu. Tumour heterogeneity in the clinic. *Nature*, 2013.

[5]S Duane, AD Kennedy, BJ Pendleton, and D Roweth. Hybrid monte carlo. *Physics letters B*, 1987.

[6]PJ Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 1995.

[7]D Hanahan and RA Weinberg. The hallmarks of cancer. *cell*, 2000.

[8]D Hanahan and RA Weinberg. Hallmarks of cancer: the next generation. *cell*, 2011.

[9]W Jiao, S Vembu, AG Deshwar, L Stein, and Q Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC bioinformatics*, 15(1):35, January 2014.

[10]JFC Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43, 1982.

[11]RM Neal. Density modeling and clustering using Dirichlet diffusion trees. *Bayesian Statistics*, 7:619–629, 2003.

[12]RM Neal. Slice sampling. *Annals of statistics*, 2003.

[13]JK Pritchard, M Stephens, and P Donnelly. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959, June 2000.

[14]A Schuh, J Becq, S Humphray, and A Alexa. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood*, 2012.

[15]Sohrab P Shah, Andrew Roth, Rodrigo Goya, Arusha Oloumi, Gavin Ha, Yongjun Zhao, Gulisa Turashvili, Jiarui Ding, Kane Tse, Gholamreza Haffari, Ali Bashashati, Leah M Prentice, Jaswinder Khattra, Angela Burleigh, Damian Yap, Virginie Bernard, Andrew McPherson, Karey Shumansky, Anamaria Crisan, Ryan Giuliany, Alireza Heravi-Moussavi, Jamie Rosner, Daniel Lai, Inanc Birol, Richard Varhol, Angela Tam, Noreen Dhalla, Thomas Zeng, Kevin Ma, Simon K Chan, Malachi Griffith, Annie Moradian, S-W Grace Cheng, Gregg B Morin, Peter Watson, Karen Gelmon, Stephen Chia, Suet-Feung Chin, Christina Curtis, Oscar M Rueda, Paul D Pharoah, Sambasivarao Damaraju, John Mackey, Kelly Hoon, Timothy Harkins, Vasisht Tadigotla, Mahvash Sigaroudinia, Philippe Gascard, Thea Tlsty, Joseph F Costello, Irmtraud M Meyer, Connie J Eaves, Wyeth W Wasserman, Steven Jones, David Huntsman, Martin Hirst, Carlos Caldas, Marco A Marra, and Samuel Aparicio. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486(7403):395–9, June 2012.

[16]J Steinhardt and Z Ghahramani. Flexible Martingale Priors for Deep Hierarchies. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 43, pages 61–62, 2012.

[17]NX Vinh, J Epps, and J Bailey. Information theoretic measures for clusterings comparison. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8, New York, New York, USA, June 2009. ACM Press.

[18]NX Vinh, J Epps, and J Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 2010.

[19]C Zhou, T Schulthess, S Torbrügge, and D Landau. Wang-Landau Algorithm for Continuous Models and Joint Density of States. *Physical Review Letters*, 96(12):120201, March 2006.