TREECLONE: RECONSTRUCTION OF TUMOR SUBCLONE PHYLOGENY BASED ON MUTATION PAIRS USING NEXT GENERATION SEQUENCING DATA[1]

BY TIANJIAN ZHOU[*,†,‡,2], SUBHAJIT SENGUPTA[†,2], PETER MÜLLER[‡] AND YUAN JI[*,†]

The University of Chicago[*], *NorthShore University HealthSystem*[†] *and The University of Texas at Austin*[‡]

We present TreeClone, a latent feature allocation model to reconstruct tumor subclones subject to phylogenetic evolution that mimics tumor evolution. Similar to most current methods, we consider data from next-generation sequencing of tumor DNA. Unlike most methods that use information in short reads mapped to single nucleotide variants (SNVs), we consider subclone phylogeny reconstruction using pairs of two proximal SNVs that can be mapped by the same short reads. As part of the Bayesian inference model, we construct a phylogenetic tree prior. The use of the tree structure in the prior greatly strengthens inference. Only subclones that can be explained by a phylogenetic tree are assigned non-negligible probabilities. The proposed Bayesian framework implies posterior distributions on the number of subclones, their genotypes, cellular proportions and the phylogenetic tree spanned by the inferred subclones. The proposed method is validated against different sets of simulated and real-world data using single and multiple tumor samples. An open source software package is available at http://www.compgenome.org/treeclone.

1. Introduction. Initiated by somatic mutations in a single cell, cancer arises through Darwinian-like natural selection. The accumulation of subsequent genetic aberrations and the effects of selection over time result in the sequential clonal expansions of cells, finally leading to the development of a genetically aberrant tumor [Nowell (1976)]. This process, known as tumorigenesis, leads to genetically divergent subpopulations of tumor cells, also known as subclones [Bonavia et al. (2011), Marusyk, Almendro and Polyak (2012)].

Based on the set of somatic mutations they have accumulated [Nik-Zainal et al. (2012)], deep next generation sequencing (NGS) of bulk tumor DNA samples makes it possible to examine the evolutionary history of individual tumors, This is implemented by deconvoluting observed genomic data from a tumor into constituent signals corresponding to various subclones and by then reconstructing the

relationship of these subclones in a phylogeny [Deshwar et al. (2015), Marass et al. (2016)]. Apart from phylogenetic relationship, tumor purity, subclones' genotypes and cellular proportions are also coupled quantities to infer. Uncovering subclonal heterogeneity and their relationship is clinically important for better prognosis [Aparicio and Caldas (2013), Schwarz et al. (2015)]. Therefore, there is a pressing need to develop robust methods for a reliable interpretation.

Numerous methods have been proposed for the subclonal reconstruction problem, including SciClone [Miller et al. (2014)], CloneHD [Fischer et al. (2014)], PyClone [Roth et al. (2014)], PyloWGS [Deshwar et al. (2015), Jiao et al. (2014)], Clomial [Zare et al. (2014)], BayClone [Sengupta et al. (2015)], Cloe [Marass et al. (2016)] and PairClone [Zhou et al. (2019)]. The reconstruction is typically based on short reads that are mapped to single nucleotide variants (SNVs) (few methods, e.g., CloneHD, also consider somatic copy number aberrations, SCNA). Many methods based on SNV data utilize variant allele fractions (VAFs), that is, the fractions of alleles (or short reads) at each locus that carry mutations. Since humans are diploid, the expected VAFs of short reads for a homogeneous cell population should be 0, 0.5 or 1.0 for any locus in copy number neutral (copy number $= 2$) regions and after adjusting for tumor purity. Observing VAFs very different from 0, 0.5 or 1.0 is therefore evidence for heterogeneity. Most methods use only marginal SNVs. Recently, Zhou et al. (2019) have proposed to use pairs of proximal SNVs mapped by the same short reads which carry more information than marginal SNVs for more accurate subclone reconstruction. In terms of methodology, existing subclone reconstruction methods can be mainly divided into two categories, clustering-based and feature-allocation-based. The two categories are also referred to as indirect and direct reconstructions in Marass et al. (2016), depending on whether the subclonal genotypes are indirectly or directly inferred. Clustering-based methods, including SciClone, PyClone and PhyloWGS, first infer SNV clusters based on observed VAFs and then reconstruct subclonal genotypes based on the clusters. The phylogenetic relationship among the subclones can be inferred by imposing hierarchy among the SNV clusters. On the other hand feature-allocation-based methods (e.g., Clomial, BayClone, Cloe or PairClone) treat subclones as latent features and directly infer subclonal genotypes. Most of the feature-allocation-based methods assume that the features (subclones) are conditionally independent and thus are not able to infer the phylogenetic relationship among the subclones. Recently, Marass et al. (2016) have developed a model allowing for dependency among the features to infer the tumor phylogenetic tree.

In the upcoming discussion we assume that the available data are from T ($T \geq 1$) samples from a single patient and the main inference goal is intratumor heterogeneity. We present a novel approach to reconstruct tumor subclones and their corresponding phylogenetic tree based on mutation pairs. Here, a mutation pair refers to a pair of proximal SNVs on the genomes that can be simultaneously mapped by the same paired-end short reads with one SNV on each end. In other words mutation pairs can be phased by short reads. See Figure 1 for an illustration.
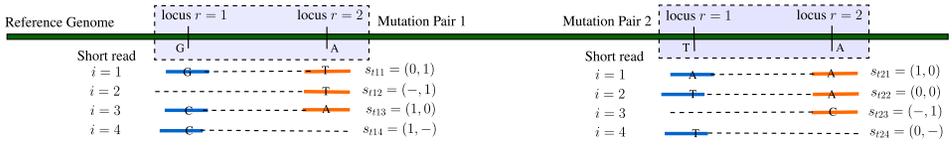
FIG. 1. *Short reads data from mutation pairs using NGS. Here, s_{tki} denotes the ith read for the kth mutation pair in sample t. Each s_{tki} is a two-dimensional vector which corresponds to the two proximal SNVs in the mutation pair, and each component of the vector takes values 0, 1 or –, representing wild type, variant or missing genotype respectively.*

Short reads mapped to only one of the SNV loci are treated as partially missing paired-end reads and are not excluded from our approach. This idea of working with phased mutation pairs was introduced in Zhou et al. (2019). We build on this work and develop a novel and entirely different inference approach by explicitly modeling the underlying phylogenetic relationship. That is, we model tumor heterogeneity based on a representation of a phylogenetic tree of tumor cell subpopulations. A prior probability model on such phylogenetic trees induces a dependent prior on the mutation profiles of latent tumor cell subpopulations. Part of this construction is that the phylogenetic tree of tumor cell subpopulations is included as a random quantity in the Bayesian model. Like most existing methods, we only consider mutation pairs in copy number neutral region, that is, copy number two. The proposed inference aims to reconstruct: (i) subclones defined by the haplotypes across all the mutation pairs, (ii) cellular proportion of each subclone and (iii) a phylogenetic tree spanned by the subclones.

Next, we introduce some notation to formally represent the described data and model structure. Consider an NGS data set with K mutation pairs shared across all T ($T \geq 1$) samples. We assume that the samples are composed of C homogeneous subclones. The number of subclones C is unknown and becomes one of the model parameters. We use a $K \times C$ matrix $\mathbf{Z}$ to represent the subclones in which each column of $\mathbf{Z}$ represents a subclone and each row represents a mutation pair. That is, the (kc) element z_{kc} of the matrix corresponds to subclone c and mutation pair k. Each z_{kc} is itself again a matrix. It is a 2×2 matrix that represents the genotypes of the two alleles of the mutation pair. See Figure 2(b). An important step in the model construction is that the columns (subclones) of $\mathbf{Z}$ form a phylogenetic tree $\mathcal{T}$. The tree encodes the parent-child relationship across the subclones. A detailed construction of the tree and a prior probability model of $\mathcal{T}$ and $\mathbf{Z}$ are introduced later. Lastly, we denote with $\boldsymbol{w}_t = (w_{t1}, \ldots, w_{tC})$ the cellular proportions of the subclones in sample t, where $0 < w_{tc} < 1$ for all c and $\sum_{c=0}^{C} w_{tc} = 1$. Using NGS data, we infer $\mathcal{T}$, C, $\mathbf{Z}$ and $\boldsymbol{w}$ based on a simple idea that variant reads can only arise from subclones with variant alleles consistent with an underlying phylogeny. We develop a tree-based *latent feature allocation model* (LFAM) to implement this reconstruction. Mutation pairs are the objects of the LFAM, and subclones are the latent features chosen by the mutation pairs.
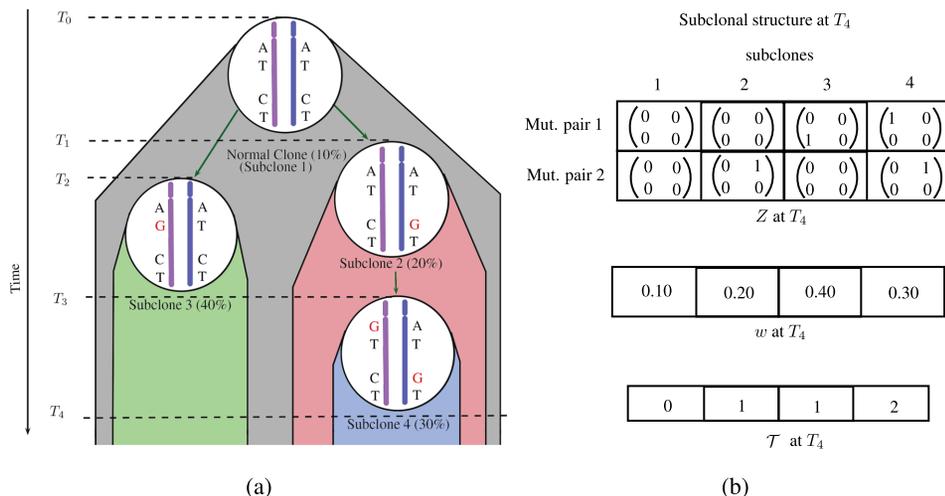
FIG. 2. *Schematic of subclonal evolution and subclone structure. Panel* (a) *shows the evolution of subclones over time. Panel* (b) *shows the subclonal structure at T_4 with genotypes $\mathbf{Z}$, cellular proportions $\mathbf{w}$ and parent vector $\mathcal{T}$. For each mutation pair k and subclone c, the entry z_{kc} of $\mathbf{Z}$ is a 2×2 matrix corresponding to the arrangement in the figure in panel* (a), *that is, with alleles in the two columns and SNVs in the rows.*

The previous brief outline of data and model structure already highlights two key features of the proposed approach: the use of phylogenetic tree structure and data on mutation pairs. The latter has important advantages. Mutation pairs contain phasing information that improves the accuracy of subclone reconstruction. If two nucleotides reside on the same short read, we know that they must appear in the same DNA strand in a subclone. For example, consider a scenario with one mutation pair and two subclones. Suppose the reference genome allele is (A, G) for that mutation pair, with the notion that A and G are phased by the same DNA strand. Suppose the two subclones have diploid genomes at the two loci and the genotypes for both DNA strands are ((C, G), (A, T)) for subclone $c = 1$, and ((C, T), (A, G)) for $c = 2$. Since in NGS short reads are generated from a single DNA strand, short reads could be any of the four haplotypes (C, G), (A, T), (C, T) or (A, G) for this mutation pair. If indeed relative large counts of short reads with each haplotype are observed, one can reliably infer that there are heterogeneous cell subpopulations in the tumor sample and the mutation pairs are subclonal. In contrast if we ignore the phasing information and only consider the (marginal) VAFs for each SNV, then the observed VAFs for both SNVs are 0.5, which could be explained by heterogeneous clonal mutations, that is, the SNVs are present in all tumor cells. In this paper we leverage the power of using mutation pairs over single SNVs to incorporate partial phasing information in our model. We assume that mutation pairs and their mapped short reads counts have been obtained using a bioinformatics pipeline such as LocHap [Sengupta et al. (2016)]. Our aim is to

use short reads mapping data on mutation pairs to reconstruct tumor subclones and their phylogeny.

Besides the use of mutation pairs, the other key feature of the proposed approach is that the model is built around phylogenetic tree structure. Imposing the phylogenetic tree structure in the prior of $\mathbf{Z}$ greatly strengthens inference. First, the tree structure improves biological interpretability of the inferred subclones as the evolutionary relationship among the subclones is explicitly modeled. Second, the tree structure improves the identifiability of the problem. In a subclone reconstruction problem, the input signals (observed VAFs) are usually relatively weak, especially when only $T = 1$ sample is available. Different subclone architectures can yield very similar observed data. By explicitly modeling the tree we can put higher prior probability on a subclone structure that follows a more likely phylogenetic tree. Third, the tree structure improves the mixing performance of the Markov chain Monte Carlo (MCMC) simulation used to infer the unknown quantities. As noted in Marass et al. (2016), the likelihood surface of the genotype matrix $\mathbf{Z}$ is highly multimodal with sharp peaks. Imposing the tree structure, in the MCMC simulation we only need to search from the space of $\mathbf{Z}$, where the tree structure is satisfied which greatly reduces the dimension of the parameter space of $\mathbf{Z}$ thus improves mixing of the Markov chain.

Finally, for clarification we briefly comment on the proposed model structure versus a traditional use of phylogenetic trees. Phylogenetic trees are usually used to approximate perfect phylogeny for a fixed number of haplotypes [Bafna et al. (2003)]. Most methods lack assessment of tree uncertainties and report a single tree estimate. Also, methods based on SNVs put the observed mutation profile of SNV at the leaf nodes. This is natural if the splits in the tree create subpopulations that acquire or do not acquire a new mutation (or set of mutations). In contrast we define a tree with all descendant nodes differing from the parent node by some mutations. That is, all nodes, including interior nodes, correspond to tumor cell subpopulations. Finally, we note that the prior structure in our model is different from the phylogenetic Indian Buffet Process (pIBP) [Miller, Griffiths and Jordan (2008)] which models phylogeny of the objects rather than the features.

The rest of the paper is organized as follows: Section 2 and Section 3 describe the latent feature allocation model and posterior inference respectively. Section 4 presents three simulation studies. Section 5 reports analysis results for an actual experiment. We conclude with a discussion in Section 6.

2. Statistical model.

2.1. *Representation of subclones.* Figure 2 presents a stylized example of temporal evolution of a tumor, starting from time T_0 and evolving until time T_4 with the normal clone (subclone $c = 1$) and three tumor subclones ($c = 2, 3, 4$). Each tumor subclone is marked by two mutation pairs with distinct somatic mutation profiles. In Figure 2 the true phylogenetic tree is plotted connecting the stylized

subclones. The true population frequencies of the subclones are marked in parentheses. In panel (b) subclone genomes, their population frequencies and the phylogenetic relationship, are represented by $\mathbf{Z}$, $\mathbf{w}$ and $\mathcal{T}$. Next, we explain in detail the definition of these parameters.

The entries of $\mathcal{T}$ report for each subclone the index of the parent subclone (with $\mathcal{T}_1 = 0$ for the root clone $c = 1$). Suppose K mutation pairs with C subclones are present. The subclone phylogeny can be visualized with a rooted tree with C nodes. We use a C-dimensional *parent* vector $\mathcal{T}$ to encode the parent-child relationship of a tree, where $\mathcal{T}_c = \mathcal{T}[c] = j$ means that subclone j is the parent of subclone c. The parent vector uniquely defines the topology of a rooted tree. We assume that the tumor evolutionary process always starts from the normal clone, indexed by $c = 1$. The normal clone does not have a parent, and we denote it by $\mathcal{T}_1 = 0$. For example, the parent vector representation of the subclone phylogeny in Figure 2 is $\mathcal{T} = (0, 1, 1, 2)$.

We use the $K \times C$ matrix $\mathbf{Z}$ to represent the subclone genotypes. Each column of $\mathbf{Z}$ defines a subclone, and each row of $\mathbf{Z}$ corresponds to a mutation pair. The entry z_{kc} records the genotypes for mutation pair k in subclone c. Since each subclone has two alleles $j = 1, 2$, and each mutation pair has two loci $r = 1, 2$, the entry z_{kc} is itself a 2×2 matrix,

$$z_{kc} = (z_{kc1}, z_{kc2}) = \left[\begin{pmatrix} z_{kc11} \\ z_{kc12} \end{pmatrix} \begin{pmatrix} z_{kc21} \\ z_{kc22} \end{pmatrix} \right],$$

where $\begin{pmatrix} z_{kc11} \\ z_{kc12} \end{pmatrix}$ and $\begin{pmatrix} z_{kc21} \\ z_{kc22} \end{pmatrix}$ represent mutation pairs of allele 1 and allele 2 respectively. That is, in z_{kcjr}, j and r index the two alleles and the two loci respectively. Theoretically, each z_{kcjr} can be any one of the four nucleotide sequences, A, C, G, T. However, at a single locus the probability of having more than two sequences is negligible since it would require the same locus to be mutated twice throughout the life span of the tumor which is extremely unlikely. Therefore, we assume z_{kcjr} can only take two possible values with $z_{kcjr} = 1$ (or 0) indicating that the corresponding locus has a mutation (or does not have a mutation) compared to the reference genome respectively. For example, in Figure 2 we have $K = 2$ mutation pairs and $C = 4$ subclones. For mutation pair $k = 2$ in subclone $c = 4$, the allele $j = 1$ harbors no mutation, while the allele $j = 2$ has a mutation at the first locus $r = 1$ which translates to $z_{24} = (00, 10)$ (writing 00 as a shorthand for $(0, 0)^T$, etc.). Altogether, z_{kc} can take $2^4 = 16$ possible values $z_{kc} \in \{(00, 00), (00, 01), \ldots, (11, 11)\}$. Since we do not have phasing information across mutation pairs, the z_{kc} values having mirrored columns lead to exactly the same data likelihood and thus are indistinguishable. This reduces the list of possible values of z_{kc} to the $Q = 10$ values, $z^{(1)} = (00, 00)$, $z^{(2)} = (00, 01)$, $z^{(3)} = (00, 10)$, $z^{(4)} = (00, 11)$, $z^{(5)} = (01, 01)$, $z^{(6)} = (01, 10)$, $z^{(7)} = (10, 10)$, $z^{(8)} = (01, 11)$, $z^{(9)} = (10, 11)$ and $z^{(10)} = (11, 11)$.

We assume that the normal subclone has no mutation $z_{k1} = z^{(1)}$ for all k, indicating all mutations are somatic.

Finally, we introduce notation for mixing proportions $\boldsymbol{w}$. Suppose T tissue samples are dissected from the same patient. We assume that the samples are admixtures of C subclones, each sample with a different set of mixing proportions (population frequencies). We use a $T \times C$ matrix $\boldsymbol{w}$ to record the proportions, where w_{tc} represents the population frequencies of subclone c in sample t, $0 < w_{tc} < 1$ and $\sum_{c=1}^{C} w_{tc} = 1$. The proportions w_{t1} denotes the proportion of normal cells contamination in sample t (and later we will still add a weight w_{t0} for a background clone $c = 0$).

2.2. *Sampling model.* Let $\boldsymbol{N}$ be a $T \times K$ matrix with N_{tk} representing read depth for mutation pair k in sample t. It records the number of times any locus of the mutation pair is covered by sequencing reads (see Figure 1). Let $\boldsymbol{s}_{tki} = (s_{tkir}, r = 1, 2)$ be a specific short read, where $r = 1, 2$ index the two loci in a mutation pair, $i = 1, 2, \ldots, N_{tk}$. Compared to the reference genome, we use $s_{tkir} = 1$ (or 0) to denote a variant (reference) sequence at the read. An important feature of the data is that read i may overlap in only one locus. We use $s_{tkir} = -$ to represent the missing other sequence on the read. Reads that do not overlap with either of the two loci are not included in the model as they do not contribute any information about the mutation pair. In summary $\boldsymbol{s}_{tki}$ can take $G = 8$ possible values,

$$\boldsymbol{s}_{tki} \in \{\boldsymbol{s}^{(1)}, \ldots, \boldsymbol{s}^{(8)}\} = \{00, 01, 10, 11, -0, -1, 0-, 1-\}.$$

Among all N_{tk} reads let $n_{tkg} = \sum_i I(\boldsymbol{s}_{tki} = \boldsymbol{s}^{(g)})$ be the number of short reads having genotype $\boldsymbol{s}^{(g)}$. For example, in Figure 1 out of a total of $N_{t1} = 4$ reads, we have $n_{t12} = 1, n_{t13} = 1, n_{t16} = 1$ and $n_{t18} = 1$.

We assume a multinomial sampling model for the observed read counts

$$(2.1) \qquad (n_{tk1}, \ldots, n_{tk8}) \mid N_{tk} \sim \text{Mn}(N_{tk}; p_{tk1}, \ldots, p_{tk8}),$$

where p_{tkg} is the probability of observing a short read $\boldsymbol{s}_{tki}$ with genotype $\boldsymbol{s}^{(g)}$. Next, we link p_{tkg} with the underlying subclone structures.

For a short read $\boldsymbol{s}_{tki}$, depending on whether it covers both loci or only one locus, we consider three cases: (i) a read covers both loci, taking values $\boldsymbol{s}_{tki} \in \{\boldsymbol{s}^{(1)}, \ldots, \boldsymbol{s}^{(4)}\}$ (complete read); (ii) a read covers the second locus, taking values $\boldsymbol{s}_{tki} \in \{\boldsymbol{s}^{(5)}, \boldsymbol{s}^{(6)}\}$ (left missing read); and (iii) a read covers the first locus, taking values $\boldsymbol{s}_{tki} \in \{\boldsymbol{s}^{(7)}, \boldsymbol{s}^{(8)}\}$ (right missing read). Let $v_{tk1}, v_{tk2}, v_{tk3}$ denote the probabilities of observing a short read of type (i), (ii) and (iii) respectively. Conditional on cases (i), (ii) or (iii), let $\tilde{p}_{tkg}$ be the conditional probability of observing $\boldsymbol{s}_{tki} = \boldsymbol{s}^{(g)}$. We have $p_{tkg} = v_{tk1}\tilde{p}_{tkg}, g = 1, \ldots, 4$, $p_{tkg} = v_{tk2}\tilde{p}_{tkg}, g = 5, 6$ and $p_{tkg} = v_{tk3}\tilde{p}_{tkg}, g = 7, 8$. We assume noninformative missingness and do not make inference on v's; so they remain constants in the likelihood.

We express $\tilde{p}_{tkg}$ in terms of $\boldsymbol{Z}$ and $\boldsymbol{w}$ based on the following generative model in three steps. Consider a sample t. To generate a short read, we first select a subclone c with probability w_{tc} (step 1). Next, we select with probability 0.5 one

of the two alleles $j = 1, 2$ (step 2). Finally, we record the read $s^{(g)}$, $g = 1, 2, 3$ or 4, corresponding to the chosen allele $z_{kcj} = (z_{kcj1}, z_{kcj2})$ (step 3). In the case of left (or right) missing locus, we observe $s^{(g)}$, $g = 5$ or 6 (or $g = 7$ or 8), corresponding to the observed locus of the chosen allele. Reflecting steps 2 and 3, we denote the probability of observing a short read $s^{(g)}$ from subclone characterized by z_{kc} by

$$(2.2) \qquad A(s^{(g)}, z_{kc}) = \sum_{j=1}^{2} 0.5 \times I(s_1^{(g)} = z_{kcj1}) I(s_2^{(g)} = z_{kcj2}),$$

with the understanding that $I(- = z_{kcjr}) \equiv 1$ for missing reads. Depending on the arguments, implicit in (2.2) is the restriction $A(s^{(g)}, z_{kc}) \in \{0, 0.5, 1\}$. Finally, proceeding as in step 1 we use the conditional probabilities $A(\cdot)$ to obtain the marginal probability of observing a short read $s^{(g)}$ from the tumor sample t with C subclones with cellular proportions $\{w_{tc}\}$ as

$$(2.3) \qquad \tilde{p}_{tkg} = \sum_{c=1}^{C} w_{tc} A(s^{(g)}, z_{kc}) + w_{t0}\rho_g.$$

The first term in equation 2.3 states that the probability of observing a short read with genotype $s^{(g)}$ is a weighted sum of the A's across all the subclones. The last term introduces the notion of a background subclone, indexed as $c = 0$ and without biological meaning, to account for noise and artifacts in the NGS data (sequencing errors, mapping errors, etc.) and also for tiny subclones that are not detectable given the sequencing depth. In (2.3) $w_{t0}\rho_g$ stands for the probability of observing $s^{(g)}$ due to this random noise. We assume the random noise does not differ across different mutation pairs, thus ρ_g does not have an index k. Note that $\rho_1 + \cdots + \rho_4 = \rho_5 + \rho_6 = \rho_7 + \rho_8 = 1$.

Finally, we note that, if desired, it is straightforward to incorporate data for marginal SNV reads in the sampling model by treating such reads as, for example, right missing reads, that is, $s_{tki2} = -$. In this case $n_{tk1} = \cdots = n_{tk6} = 0$ and the multinomial sampling model reduces to a binomial model. The addition of marginal SNV counts does not typically improve inference. See more details in Zhou et al. (2019).

2.3. *Prior model.* We construct a hierarchical prior model, starting with $p(C)$, then a prior on the tree for a given number of nodes, $p(\mathcal{T} \mid C)$ and finally a prior on the subclonal genotypes given the phylogenetic tree $\mathcal{T}$. We assume a geometric prior for the number of subclones, $p(C) = (1 - \alpha)^{C-1}\alpha$, $C \in \{1, 2, 3, \ldots\}$. Conditional on C, the prior on the tree, $p(\mathcal{T} \mid C)$ is as in Chipman, George and McCulloch (1998). For a tree with C nodes, we let

$$p(\mathcal{T} \mid C) \propto \prod_{c=1}^{C} (1 + \eta_c)^{-\beta},$$

where η_c is the depth of node c or the number of generations between node c and the normal subclone 1. The prior penalizes deeper trees and thus favors parsimonious representation of subclonal structure.

Conditional on $\mathcal{T}$ we define a prior for $\mathbf{Z}$. The subclone genotype matrix $\mathbf{Z}$ can be thought of as a feature allocation for categorical matrices. The mutation pairs are the objects, and the subclones are the latent features chosen by the objects. Each feature has 10 categories corresponding to the $Q = 10$ different genotypes. Given $\mathcal{T}$ the construction of the subclone genotype matrix needs to introduce dependence across features to respect the assumed phylogeny. We construct a prior for $\mathbf{Z}$ based on the following generative model. We start from a normal subclone denoted by $z_{\cdot 1} = \mathbf{0}$. Now, consider a subclone $c > 1$ and defined by $z_{\cdot c}$. The subclone preserves all mutations from its parent $z_{\cdot \mathcal{T}_c}$ but also gains a Poisson number of new mutations. We assume the new mutations happen randomly at the unmutated loci of the parent subclone. A formal description of prior of $\mathbf{Z}$ follows.

For a subclone c let $\ell_{kc} = \sum_{j,r} z_{kcjr}$ denote the number of mutations in mutation pair k, and let $\mathcal{L}_c = \{k : \ell_{kc} < 4\}$ denote the mutation pairs in subclone c that have less than four mutations. Let $m_{kc} = \ell_{kc} - \ell_{k\mathcal{T}_c}$ denote the number of new mutations that mutation pair k gains compared to its parent, and let $m_{\cdot c} = \sum_k m_{kc}$. We assume: (i) The child subclone should acquire at least one additional mutation compared with its parent (otherwise subclone c would be identical to its parent $\mathcal{T}_c$). (ii) If the parent has already acquired all four mutations for a given k, then the child can not gain any more new mutation. That is, if $\ell_{k\mathcal{T}_c} = 4$, then $m_{kc} = 0$. (iii) Each mutation pair can gain at most one additional mutation in each generation, $m_{kc} \in \{0, 1\}$. Based on these assumptions, given a parent subclone $z_{\cdot \mathcal{T}_c}$, we construct a child subclone $z_{\cdot c}$ as follows. Let $\mathcal{M}_c = \{k : m_{kc} = 1\}$ be the set of mutation pairs in subclone c where new mutations are gained. Let $\text{Choose}(\mathcal{L}, m)$ denote a uniformly chosen subset of $\mathcal{L}$ of size m, and let $X \sim \text{Trunc-Pois}(\lambda; [a, b])$ represent a Poisson distribution with mean λ, truncated to $a \leq X \leq b$. We assume

$$m_{\cdot c} \mid z_{\cdot \mathcal{T}_c}, \mathcal{T}, C \sim \text{Trunc-Pois}(\lambda; [1, |\mathcal{L}_{\mathcal{T}_c}|]),$$

(2.4) $$\mathcal{M}_c \mid m_{\cdot c}, z_{\cdot \mathcal{T}_c}, \mathcal{T}, C \sim \text{Choose}(\mathcal{L}_{\mathcal{T}_c}, m_{\cdot c}).$$

The lower bound and upper bound of the truncated Poisson reflect assumptions (i) and (ii) respectively. Also, equation 2.4 implicitly captures assumption (iii).

Next, for a mutation pair that gains one new mutation, we assume the new mutation randomly arises in any of the unmutated loci in the parent subclone. Let $\mathcal{Z}_{kc} = \{(j, r) : z_{kcjr} = 0\}$, and let $\text{Unif}(A)$ denote a uniform distribution over the set A. We first choose

$$(j^*, r^*) \mid z_{\cdot \mathcal{T}_c}, \mathcal{T}, C \sim \text{Unif}(\mathcal{Z}_{k\mathcal{T}_c}),$$

and then set $z_{kcj^*r^*} = 1$. So we have

$$p(\mathbf{Z} \mid \mathcal{T}, C) \propto \prod_{c=2}^{C} \text{Trunc-Pois}\big(m_{\cdot c}; [1, |\mathcal{L}_{\mathcal{T}_c}|]\big) \cdot \frac{1}{\binom{|\mathcal{L}_{\mathcal{T}_c}|}{m_{\cdot c}}} \cdot \prod_{k \in \mathcal{M}_c} \frac{1}{|\mathcal{Z}_{k\mathcal{T}_c}|}.$$

The prior construction is completed with a prior for $\boldsymbol{w}$ and $\boldsymbol{\rho}$. We design the prior $p(\boldsymbol{w})$ of $\boldsymbol{w}$ in such a manner that we could put an informative prior for w_{t1}, if a reliable estimate for tumor purity is available based on some prior bioinformatics pipeline (e.g., Carter et al. (2012), Van Loo et al. (2010)). Recall that $c = 1$ is the normal subclone, that is, w_{t1} is the normal subclone proportion (or 1 minus tumor purity) and that $\sum_{c=0, c\neq 1}^{C} w_{tc} + w_{t1} = 1$. We assume a Beta-Dirichlet prior on $\boldsymbol{w}$ such that

$$w_{t1} \sim \text{Be}(a_p, b_p) \quad \text{and} \quad \frac{w_{tc}}{(1 - w_{t1})} \sim \text{Dir}(d_0, d, \ldots, d),$$

where $c = 0, 2, 3, \ldots, C$. We set $d_0 \ll d$ as w_{t0} is only a correction term to account for background noise and model mis-specification term.

Finally, to specify a prior for $\boldsymbol{\rho} = \{\rho_g\}$, we consider complete reads, left missing reads and right missing reads separately and assume $(\rho_1, \ldots, \rho_4) \sim \text{Dir}(d_1, \ldots, d_1)$, $(\rho_5, \rho_6) \sim \text{Dir}(2d_1, 2d_1)$ and $(\rho_7, \rho_8) \sim \text{Dir}(2d_1, 2d_1)$.

3. Posterior inference. Let $\boldsymbol{x} = (\boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{\rho})$ denote the unknown parameters except for the number of subclones C and the tree $\mathcal{T}$. MCMC simulation from the posterior $p(\boldsymbol{x} \mid \boldsymbol{n}, \mathcal{T}, C)$ is used to implement posterior inference. Gibbs sampling transition probabilities are used to update $\boldsymbol{Z}$, and Metropolis–Hastings transition probabilities are used to update $\boldsymbol{w}$ and $\boldsymbol{\rho}$. For example, we update $\boldsymbol{Z}$ by row with

$$p(\boldsymbol{z}_{k\cdot} \mid \boldsymbol{z}_{-k\cdot}, \ldots) \propto \prod_{t=1}^{T} \prod_{g=1}^{G} \left[\sum_{c=1}^{C} w_{tc} A(\boldsymbol{h}_g, z_{kc}) + w_{t0}\rho_g \right]^{n_{tkg}}$$
$$\cdot p(\boldsymbol{z}_{k\cdot} \mid \boldsymbol{z}_{-k\cdot}, \mathcal{T}, C),$$

where $\boldsymbol{z}_{k\cdot}$ is a row of $\boldsymbol{Z}$ satisfying the phylogeny $\mathcal{T}$.

Since the posterior distribution $p(\boldsymbol{x} \mid \boldsymbol{n}, \mathcal{T}, C)$ is expected to be highly multimodal, we utilize parallel tempering [Geyer (1991)] to improve the mixing of the chain. Specifically, we use OpenMP parallel computing API [Dagum and Menon (1998)] in C++ to implement a scalable parallel tempering algorithm.

Updating C and $\mathcal{T}$ is more difficult. In general posterior MCMC on tree structures can be very challenging to implement [Chipman, George and McCulloch (1998), Denison, Mallick and Smith (1998)]. However, the problem here is manageable since plausible numbers for C constrain $\mathcal{T}$ to moderately small trees. We assume that the number of nodes is a priori restricted to $C_{\min} \leq C \leq C_{\max}$. Conditional on the number of subclones C, the number of possible tree topologies is finite. Let $\mathscr{T}$ denote the (discrete) sample space of $(\mathcal{T}, C)$. Updating the values of $(\mathcal{T}, C)$ involves transdimensional MCMC. At each iteration we propose new values for $(\mathcal{T}, C)$ from a uniform proposal, $q(\tilde{\mathcal{T}}, \tilde{C} \mid \mathcal{T}, C) \sim \text{Unif}(\mathscr{T})$.

In order to search the space $\mathscr{T}$ for the number of subclones and trees that best explain the observed data, we follow a similar approach as in Lee et al. (2015), Zhou et al. (2019) (motivated by fractional Bayes' factor in O'Hagan (1995)) that

splits the data into a training set and a test set. Recall that $\boldsymbol{n}$ represents the read counts data. We split $\boldsymbol{n}$ into a training set $\boldsymbol{n}'$ with $n'_{tkg} = bn_{tkg}$ and a test set $\boldsymbol{n}''$ with $n''_{tkg} = (1-b)n_{tkg}$. Let $p_b(\boldsymbol{x} \mid \mathcal{T}, C) = p(\boldsymbol{x} \mid \boldsymbol{n}', \mathcal{T}, C)$ be the posterior evaluated on the training set only. We use p_b in two instances. First, p_b is used as an informative prior instead of the original prior $p(\boldsymbol{x} \mid \mathcal{T}, C)$, and second, p_b is used as a proposal distribution for $\tilde{\boldsymbol{x}}$, $q(\tilde{\boldsymbol{x}} \mid \tilde{\mathcal{T}}, \tilde{C}) = p_b(\tilde{\boldsymbol{x}} \mid \tilde{\mathcal{T}}, \tilde{C})$. Finally, the acceptance probability of proposal $(\tilde{\mathcal{T}}, \tilde{C}, \tilde{\boldsymbol{x}})$ is evaluated on the test set. Importantly, in the acceptance probability the (intractable) normalization constant of p_b cancels out and makes this approach computationally feasible:

$$p_{\mathrm{acc}}(\mathcal{T}, C, \boldsymbol{x}, \tilde{\mathcal{T}}, \tilde{C}, \tilde{\boldsymbol{x}}) = 1 \wedge \frac{p(\boldsymbol{n}'' \mid \tilde{\boldsymbol{x}}, \tilde{\mathcal{T}}, \tilde{C})}{p(\boldsymbol{n}'' \mid \boldsymbol{x}, \mathcal{T}, C)} \cdot \frac{p(\tilde{\mathcal{T}}, \tilde{C}) \, p_b(\tilde{\boldsymbol{x}} \mid \tilde{\mathcal{T}}, \tilde{C})}{p(\mathcal{T}, C) \, p_b(\boldsymbol{x} \mid \mathcal{T}, C)}$$
$$\cdot \frac{q(\mathcal{T}, C \mid \tilde{\mathcal{T}}, \tilde{C}) \, q(\boldsymbol{x} \mid \mathcal{T}, C)}{q(\tilde{\mathcal{T}}, \tilde{C} \mid \mathcal{T}, C) \, q(\tilde{\boldsymbol{x}} \mid \tilde{\mathcal{T}}, \tilde{C})}.$$

Here we use p_b as an informative proposal distribution for $\tilde{\boldsymbol{x}}$ to achieve a better mixing MCMC simulation with reasonable acceptance probabilities. Without the use of an informative proposal, the proposed new tree is almost always rejected because the multinomial likelihood with the large sample size is very peaked. Under the modified prior $p_b(\cdot)$, the resulting conditional posterior on $\boldsymbol{x}$ remains entirely unchanged, $p_b(\boldsymbol{x} \mid \mathcal{T}, C, \boldsymbol{n}) = p(\boldsymbol{x} \mid \mathcal{T}, C, \boldsymbol{n})$ [Zhou et al. (2019)].

The described uniform tree proposal is in contrast to usual search algorithms for trees that generate proposals from neighboring trees. Such algorithms have the important drawback that they quickly gravitate toward a local mode and then get stuck. A possible approach to addressing this problem is to repeatedly restart the algorithm from different starting trees; see Chipman, George and McCulloch (1998) for more details. Our uniform tree proposal combined with the data splitting scheme is another way to mitigate this challenge, efficiently searching the tree space while keeping a reasonable acceptance probability.

All posterior inference is contained in the posterior distribution for $\boldsymbol{x}, C$ and $\mathcal{T}$. For example, the marginal posterior distribution of C and $\mathcal{T}$ gives updates posterior probabilities for all possible values of C and $\mathcal{T}$ but it is still useful to report point estimates. We use the posterior modes $(\hat{C}, \hat{\mathcal{T}})$ as point estimates for $(C, \mathcal{T})$, and, conditional on $\hat{C}$ and $\hat{\mathcal{T}}$, we use the maximum a posteriori (MAP) estimator as an estimation for the other parameters. The MAP is approximated as the MCMC sample with highest posterior probability. Let $\{\boldsymbol{x}^{(l)}, l = 1, \ldots, L\}$ be a set of MCMC samples of $\boldsymbol{x}$, and

$$\hat{l} = \underset{l \in \{1, \ldots, L\}}{\arg\max} \, p(\boldsymbol{n} \mid \boldsymbol{x}^{(l)}, \hat{\mathcal{T}}, \hat{C}) p(\boldsymbol{x}^{(l)} \mid \hat{\mathcal{T}}, \hat{C}).$$

We report point estimates as $\hat{\boldsymbol{Z}} = \boldsymbol{Z}^{(\hat{l})}$, $\hat{\boldsymbol{w}} = \boldsymbol{w}^{(\hat{l})}$ and $\hat{\boldsymbol{\rho}} = \boldsymbol{\rho}^{(\hat{l})}$.

4. Simulation studies. We carry out three simulation studies to assess the performance of TreeClone. We simulate multiple datasets with different number of subclones (C), number of samples (T), average read depths ($\bar{N}_{tk}$) and left and right missing rates (v_{tk2}, v_{tk3}) to test the performance of the proposed model in different scenarios. In all simulation studies we generate hypothetical read count data for dozens of mutation pairs ($K = 100$ for simulations 1 and 2, and $K = 69$ for simulation 3) which is a typical number of mutation pairs in a real tumor sample.

4.1. *Simulation* 1, *convergence diagnostic and sensitivity analysis.*

Simulation 1(a). We first validate TreeClone on nine simulated datasets, one for each combination of $C \in \{3, 4, 5\}$ and average read depth $\bar{N}_{tk} \in \{50, 200, 1000\}$. For all nine datasets we consider $T = 5$ samples and $K = 100$ mutation pairs. We randomly generate the phylogenetic tree $\mathcal{T}$ and the genotype matrix $\mathbf{Z}$ from the prior model. The subclone proportions are simulated from $\mathbf{w}_t \sim$ Dir($0.01, \sigma(15, 10, 5)$), Dir($0.01, \sigma(15, 10, 8, 5)$) and Dir($0.01, \sigma(15, 10, 8, 5, 3)$) for $C = 3, 4$ and 5 respectively. Here $\sigma(x_1, \ldots, x_n)$ stands for a random permutation of $x_1, \ldots, x_n$. The noise fractions $\boldsymbol{\rho}$ are generated from the prior with $d_1 = 1$. We mimic a typical rate of left (or right) missing reads by setting $v_{tk2} = v_{tk3} = 0.3$ for all t and k. The read depth N_{tk} is generated from a negative-binomial distribution, $N_{tk} \sim \text{NB}(r_N, p_N)$, to reflect the over dispersion of read depth in sequencing data. We fix r_N and p_N such that $\text{E}(N_{tk}) = \bar{N}_{tk}$ and $\text{sd}(N_{tk}) = \bar{N}_{tk}/5$ for $\bar{N}_{tk} = 50$, 200 and 1000. Finally, the read counts $\{n_{tkg}\}$ are simulated from the multinomial sampling model (2.1).

We fit the model with the following hyperparameters: $\alpha = 0.5, \beta = 0.5, d = 0.5$, $d_0 = 0.03$ and $d_1 = 1$, where the values of α and β imply mild penalty for deep and bushy trees [Chipman, George and McCulloch (1998)]. We set $a_p = d, b_p = d_0 + (C - 1)d$ for a given C and set $\lambda = 2K/C$ to express our prior belief that about half of the mutations occur uniformly at each generation. We use $C_{\min} = 2$ and $C_{\max} = 5$ as the range of C for computational efficiency. In principle $C_{\max}$ can be any finite number. However, the size of the tree space grows exponentially with $C_{\max}$, so that the time needed for achieving a good mixing Markov chain grows exponentially. We set the training set fraction (in the transdimensional MCMC implementation) to $b = 0.95$, which we found to perform well in all simulation studies; see Supplementary Section 2 for a discussion [Zhou et al. (2019)]. We run a total of 8000 MCMC iterations. Discarding the first 3000 draws as initial burn-in, we have a Monte Carlo sample with 5000 posterior draws.

We evaluate the performance of TreeClone in estimating the number of subclones C, phylogeny $\mathcal{T}$, genotype $\mathbf{Z}$ and subcone proportions $\mathbf{w}$. To this end, we define the reconstruction error rates $C_{\text{err}} = I(\hat{C} \neq C), \mathcal{T}_{\text{err}} = I(\hat{\mathcal{T}} \neq \mathcal{T})$,

$$Z_{\text{err}} = \frac{1}{K(C-1)} \min_{\sigma} \left(\sum_{k,c} I(\hat{z}_{k\sigma(c)} \neq z_{kc}) \right),$$

TABLE 1

Simulation 1(a). *Summary of simulation results for nine combinations of C and $\bar{N}_{tk}$. Each cell of the table reports $(C_{\mathrm{err}}, \mathcal{T}_{\mathrm{err}}, Z_{\mathrm{err}}, w_{\mathrm{err}})$ as an average over three runs of TreeClone with different random number seeds*

$\bar{N}_{tk}$	C		
	3	4	5
50	0, 0, 0.00, 0.09	0, 0, 0.03, 0.07	0, 0, 0.07, 0.06
200	0, 0, 0.00, 0.15	0, 0, 0.00, 0.09	0, 0, 0.00, 0.05
1000	0, 0, 0.00, 0.14	0, 0, 0.00, 0.10	0, 0, 0.00, 0.08

and $w_{\mathrm{err}} = \sum_{t,c} |\hat{w}_{t\sigma(c)} - w_{tc}|/(TC)$, similar to Marass et al. (2016). Here, σ is a permutation of columns of $\boldsymbol{Z}$ to account for label switching of subclones, and the same permutation is imposed on columns of $\boldsymbol{w}$. The simulation results are summarized in Table 1. For all nine simulated datasets TreeClone nicely recovers the truth and attains reasonably low reconstruction errors.

To assess the convergence of the MCMC algorithm, for each simulated dataset we run the algorithm three times with different random seeds. We use the log-posterior values to calculate a potential scale reduction factor (PSRF) [Gelman and Rubin (1992)] for the three Markov chains. A PSRF close to 1 indicates convergence of the Markov chain to the target distribution. The results are reported in Table 2. Next, to assess the identifiability of the parameters in the TreeClone model, we calculate frequentist coverage probabilities of 95% posterior credible intervals for p_{tkg}. The results are shown as the second entry in each cell of Table 2.

Simulation 1(b). Next, we vary average read depth $\bar{N}_{tk}$ and T. Again, we simulate nine datasets, one for each combination of T and $\bar{N}_{tk}$ with $T \in \{1, 3, 5\}$ and

TABLE 2

Simulation 1(a). *Convergence diagnostic and frequentist coverage probabilities for nine simulated datasets. For each simulated dataset we run TreeClone three times with different random seeds. Each cell of the table reports the PSRF (for log-posterior values) and average coverage rate (of 95% credible intervals for p_{tkg}), averaged over the three chains*

$\bar{N}_{tk}$	C		
	3	4	5
50	1.06, 100%	1.08, 85%	1.33, 96%
200	1.13, 92%	1.04, 86%	1.38, 88%
1000	1.05, 100%	1.11, 83%	3.72, 94%

TABLE 3

Simulation 1(b). *Summary of posterior inference for nine combinations of T and $\bar{N}_{tk}$. Each cell of the table reports* $(C_{\mathrm{err}}, \mathcal{T}_{\mathrm{err}}, Z_{\mathrm{err}}, w_{\mathrm{err}})$

$\bar{N}_{tk}$	T		
	1	**3**	**5**
50	0, 0, 0.29, 0.05	0, 0, 0.11, 0.09	0, 0, 0.03, 0.08
200	0, 0, 0.16, 0.04	0, 0, 0.01, 0.07	0, 0, 0.00, 0.08
1000	0, 0, 0.13, 0.02	0, 0, 0.00, 0.12	0, 0, 0.00, 0.10

$\bar{N}_{tk} \in \{50, 200, 1000\}$. We assume the same genotype matrix $\mathbf{Z}$ for all the nine datasets with $C = 4$ subclones and $K = 100$ mutation pairs. The parameters are generated from the prior model, and N_{tk} is generated from $\mathrm{NB}(r_N, p_N)$ as in simulation 1(a).

The simulation results are summarized in Table 3. Even with only one sample and low read depth, TreeClone can reliably estimate C and $\mathcal{T}$ (although it does not perfectly recover $\mathbf{Z}$ due to the limited amount of data).

Simulation 1(c).   Finally, we assess sensitivity with respect to the rates of left and right missing mutation pairs v_{tk2} and v_{tk3}. Missingness is simply due to the length of a short read is not long enough to cover both loci in a mutation pair. Thus, missingness is noninformative, and we assume $v_{tk2} = v_{tk3}$ in the simulation for simplicity. We simulate five datasets with missing rates $v_{tk2} = v_{tk3} = 0, 0.1, 0.25,$ 0.4 and 0.5. The extreme case $v_{tk2} = v_{tk3} = 0.5$ implies that no complete mutation pair reads are recorded. For all of the five datasets, we consider $T = 5$ samples and average read depth $\bar{N}_{tk} = 200$, and we assume the same genotype $\mathbf{Z}$ with $C = 4$ subclones and $K = 100$ mutation pairs. The parameters are generated from the prior model as in simulation 1(a) and (b). The simulation results are summarized in Table 4. TreeClone maintains high reconstruction accuracy across all scenarios.

Using only marginal SNVs.   We consider another simulation assuming that we observe only marginal SNVs that are not phased. We treat the marginal

TABLE 4

Simulation 1(c). *Summary of posterior inference in simulated datasets under a range of values for* $v_{tk2} = v_{tk3}$. *Each entry reports* $(C_{\mathrm{err}}, \mathcal{T}_{\mathrm{err}}, Z_{\mathrm{err}}, w_{\mathrm{err}})$

$v_{tk2} = v_{tk3}$				
0	**0.1**	**0.25**	**0.4**	**0.5**
0, 0, 0.00, 0.08	0, 0, 0.00, 0.09	0, 0, 0.00, 0.08	0, 0, 0.00, 0.07	0, 0, 0.12, 0.07

SNVs as right missing reads, that is, $v_{tk2} = 0$ and $v_{tk3} = 1$. We simulate one dataset under this scenario. The reconstruction errors are $(C_{\text{err}}, \mathcal{T}_{\text{err}}, Z_{\text{err}}, w_{\text{err}}) = (0, 0, 0.46, 0.1)$. The number of subclones and phylogenetic tree are correctly identified. The genotype reconstruction error Z_{err} is high because the data do not provide any information for inferring the unobserved locus z_{kcj2} for all k, c and j. For a fair comparison we redefine $Z_{\text{err}}^{\text{SNV}} = \min_\sigma (\sum_{k,c} I(\hat{z}_{k\sigma(c)j1} \neq z_{kcj1}$ for $j = 1$ or 2$))/(K(C-1))$. The redefined reconstruction error is $Z_{\text{err}}^{\text{SNV}} = 0.00$.

4.2. *Simulation* 2 *and comparison with alternatives.*　In the second simulation study we compare the performance of TreeClone with existing methods. We consider $T = 1$ sample which is the case for most real-world tumor cases (due to the challenge in obtaining multiple samples from a patient). In practice we notice that methods using only marginal SNV data find it hard to identify branching in a phylogenetic tree. We use this simulation to illustrate the information gain by using mutation pairs data.

We consider $K = 100$ mutation pairs and assume a simulation truth with $C = 4$ latent subclones with a true phylogenetic tree as

$$1 \diagup\!\!\!\!\diagdown \genfrac{}{}{0pt}{}{2 \to 4}{3.}$$
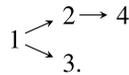
Figure 3(a) shows the true underlying subclonal genotypes. We use a heatmap to show the subclone matrix $\mathbf{Z}$, where colors from light gray to black are used to represent genotypes $z^{(1)}$ to $z^{(10)}$. The subclone weights are simulated from Dir$(0.01, \sigma(15, 10, 8, 5))$. For the single sample in this simulation, we get $\mathbf{w} = (0.000, 0.135, 0.169, 0.470, 0.226)$. We generate $\boldsymbol{\rho}$ from the prior model with $d_1 = 1$, and we set $v_{tk2} = v_{tk3} = 0.3$ for all t and k. We generate the read depth $N_{tk} \sim \text{NB}(r_N, p_N)$ with $\text{E}(N_{tk}) = 500$ and $\text{sd}(N_{tk}) = 100$.

The hyperparameters are set as in simulation 1. To explore a larger tree space, we set $C_{\max} = 6$, run a total of 13,000 MCMC iterations and discard the first 3000 draws as initial burn-in.

Posterior inference is summarized in Figure 3(b), (c). Figure 3(c) shows the top three tree topologies and corresponding posterior probabilities. The posterior mode recovers the true phylogeny. Figure 3(b) shows the estimated genotypes conditional on the posterior modes $(\hat{C}, \hat{\mathcal{T}})$. Some mismatches are due to the single sample and limited read depth. The estimated subclone proportions are $\hat{\mathbf{w}} = (0.000, 0.103, 0.162, 0.498, 0.237)$.

Comparison with alternatives.　There is no other subclone calling method based on paired-end read data that also infers phylogeny. We therefore compare with other similar model-based approaches. In particular, we use Cloe [Marass et al. (2016)], PhyloWGS [Deshwar et al. (2015), Jiao et al. (2014)] and PairClone
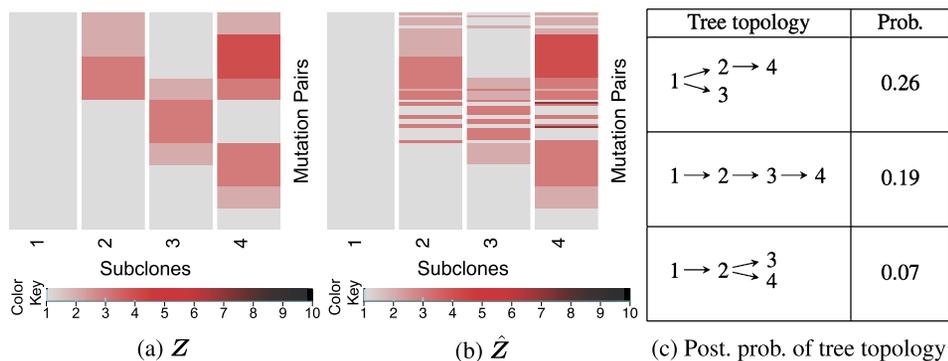
FIG. 3. *Simulation 2. True genotype matrix* $\mathbf{Z}$ *in the form of a heatmap* (a), *estimated genotype matrix* $\hat{\mathbf{Z}}$ *under TreeClone* (b), *and three tree topologies with the highest posterior probabilities estimated by TreeClone* (c). *In panels* (a), (b) *colors from light gray to black represent genotypes* $z^{(1)}$ *to* $z^{(10)}$.

[Zhou et al. (2019)] for inference with the same simulated data. Cloe and Phy-loWGS infer phylogeny but take marginal SNV data as input. For these methods we therefore discard the phasing information in mutation pairs and only record marginal mutation counts for SNVs. The simulation truth for Cloe and PhyloWGS is shown in Figure 4(a). The color means a heterozygous mutation at the corresponding SNV locus. PairClone takes the same mutant read counts and read depths for mutation pairs as input but uses a different probability model that does not allow to infer the phylogenetic tree.

Cloe infers clonal genotypes and phylogeny based on a similar feature allocation model. We run Cloe with the default hyperparameters for the same number of 13,000 iterations with the first 3000 draws as initial burn-in. Based on the MAP estimate (over $2 \leq C \leq 6$), Cloe reports $C = 3$ subclones with phylogeny $1 \rightarrow$
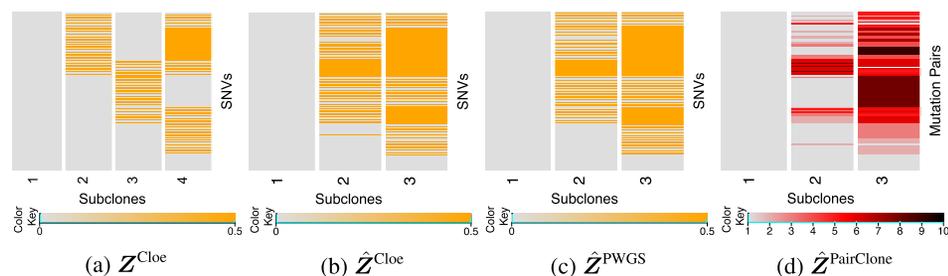

FIG. 4. *Simulation 2. True genotype matrix* $\mathbf{Z}^{\text{Cloe}}$ *in the form of a heatmap* (a), *estimated genotype matrices* $\hat{\mathbf{Z}}^{\text{Cloe}}$ *under Cloe* (b) *and* $\hat{\mathbf{Z}}^{\text{PWGS}}$ *under PhyloWGS* (c) *and estimated genotype matrix* $\hat{\mathbf{Z}}^{\text{PairClone}}$ *under PairClone* (d). *In panels* (a), (b), (c) *the colors represent a heterozygous mutation at the corresponding SNV locus. In panel* (d) *the colors represent genotypes* $z^{(1)}$ *to* $z^{(10)}$ *for the mutation pairs.*

$2 \rightarrow 3$, and the subclone genotypes are shown in Figure 4(b) with subclone proportions $\hat{\boldsymbol{w}}^{\text{Cloe}} = (0.555, 0.223, 0.222)$.

PhyloWGS, on the other hand, infers clusters of mutations and phylogeny. One can then make phylogenetic analysis to conjecture subclones and genotypes. Let $\tilde{\phi}_i$ denote the fraction of cells with a variant allele at locus i. The $\tilde{\phi}_i$'s are latent quantities related to the observed VAF for each SNV. PhyloWGS infers the phylogeny by clustering SNVs with matching $\tilde{\phi}_i$'s under a tree-structured prior for the unique values ϕ_j. In particular they use the tree-structured stick breaking process (TSSB) [Adams, Ghahramani and Jordan (2010)]. The TSSB implicitly defines a prior on the formation of subclones, including a prior on C and the number of novel loci that arise in each subclone (in contrast, TreeClone explicitly defines these model features, allowing easier prior control on C and $\mathcal{M}_c$). We run PhyloWGS with the default hyperparameters and 2500 iterations with a burn-in of 1000 samples. We only consider loci with VAF > 0; as for PhyloWGS the other loci do not provide information on clustering. We then report cluster sizes and phylogeny based on MAP estimate. PhyloWGS reports three subclones with phylogeny $0 \rightarrow 1(79, 0.438) \rightarrow 2(53, 0.220)$, where 0 refers to the normal subclone, and the numbers in the brackets refer to the cluster sizes and cellular prevalences. The conjectured subclone genotypes are shown in Figure 4(c) with subclone proportions $\hat{\boldsymbol{w}}^{\text{PWGS}} = (0.562, 0.218, 0.220)$.

Inferences under Cloe and PhyloWGS do not entirely recover the truth. Let $\mathcal{M}_c$ denote the new mutations that are gained by subclone c. The reason for the failure to recover the simulation truth is probably that the common mutations of subclones 2 and 4 ($\mathcal{M}_2$ with a cellular prevalence of $0.169 + 0.226$) have a similar cellular prevalence as the mutations of subclone 3 ($\mathcal{M}_3$ with a cellular prevalence of 0.470). Therefore, Cloe infers that $\mathcal{M}_2$ and $\mathcal{M}_3$ belong to the same subclone ($\mathcal{M}_2^{\text{Cloe}} \approx \mathcal{M}_2 \cup \mathcal{M}_3$ and $\mathcal{M}_3^{\text{Cloe}} \approx \mathcal{M}_4$). Similarly, PhyloWGS clusters $\mathcal{M}_2$ and $\mathcal{M}_3$ together. Using more informative mutation pairs data, TreeClone can infer that $\mathcal{M}_2$ and $\mathcal{M}_3$ belong to different subclones. The inclusion of phasing information from the paired-end read data increases statistical power in recovering the underlying structure.

PairClone uses the same mutation pairs data and same sampling model to infer clonal genotypes. However, PairClone uses a finite categorical Indian buffet process prior for $\boldsymbol{Z}$ which assumes independence among the subclones and does not infer phylogeny. PairClone infers three subclones with genotypes shown in Figure 4(d). The estimated subclone proportions are $\hat{\boldsymbol{w}}^{\text{PairClone}} = (0.594, 0.229, 0.177)$, similar to Cloe and PhyloWGS's results. PairClone does not entirely recover the truth, probably because not imposing the tree structure reduces the identifiability of the problem.

Comparison using additional marginal SNVs. An NGS dataset contains many more marginal SNVs than phased mutation pairs. These additional marginal SNVs

can be utilized by methods such as Cloe and PhyloWGS. For an illustration of the information gain by using more SNVs, we run Cloe and PhyloWGS with a larger set of SNVs that contains 400 SNVs in addition to the 200 SNVs that are part of the 100 mutation pairs, that is, a total of 600 SNVs. We assume a genotype matrix $\mathbf{Z}_+ = (\mathbf{Z}, \mathbf{Z}, \mathbf{Z})'$, that is, repeating the rows of $\mathbf{Z}$ three times and keep the other parameters unchanged from before.

Using 600 SNVs, Cloe infers two subclones with phylogeny $1 \rightarrow 2$ and proportions $\hat{\mathbf{w}}_+^{\text{Cloe}} = (0.641, 0.359)$. PhyloWGS infers three (conjectured) subclones with phylogeny $1 \rightarrow 2 \rightarrow 3$ and proportions $\hat{\mathbf{w}}_+^{\text{PWGS}} = (0.562, 0.212, 0.226)$. The results suggest that the additional 400 SNVs do not contribute much information about the branching in the phylogenetic tree.

4.3. *Simulation* 3.   In the third simulation, we evaluate the performance of the proposed approach on multiple samples. We generate the simulated data by mimicking a real-world lung cancer data (see Section 5). Following that dataset, we consider $T = 4$ tissue samples and $K = 69$ mutation pairs. The simulation truth $\mathbf{Z}$ and $\mathbf{w}$ are generated by fitting TreeClone on the lung cancer dataset. We estimate $C = 6$ subclones with phylogeny

$$1 \begin{array}{c} \nearrow \\ \searrow \end{array} \begin{array}{c} 2 \begin{array}{c} \nearrow 3 \\ \searrow 4 \end{array} \\ 5 \rightarrow 6. \end{array}$$
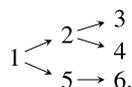
Figure 5(a), (b) shows the simulation truth $\mathbf{Z}$ and $\mathbf{w}$ in the form of heatmaps respectively. For $\mathbf{w}$ a darker color indicates higher abundance of a subclone in a sample (the proportions w_{t0} of the background subclone are tiny and not shown). Read depths $\{N_{tk}\}$ and left and right missing rates $\{v_{tk2}, v_{tk3}\}$ are taken to be exactly the same as in the real data. The average read depth is $\bar{N}_{tk} = 156$.

The hyperparameters are set as in Simulation 1. To explore a larger tree space, we set $C_{\max} = 7$, run a total of 30,000 MCMC iterations and discard the first 3000 draws as initial burn-in.

Posterior inference is summarized in Figure 5(c), (d), (e). Figure 5(c) shows the top three tree topologies and corresponding posterior probabilities. The posterior mode recovers the true phylogeny. Figure 5(d), (e) shows the estimated genotypes $\hat{\mathbf{Z}}$ and subclone proportions $\hat{\mathbf{w}}$ conditional on $(\hat{C}, \hat{\mathcal{T}})$. There are some mismatches for subclones 2, 4 and 5, as they only take small proportions in all the samples ($w_{t2}, w_{t5} < 0.02$ and $w_{t4} < 0.11$ for all t). The reconstruction errors are $Z_{\text{err}} = 0.21$ and $w_{\text{err}} = 0.01$ (considering only subclones 1, 3 and 6 the reconstruction error becomes $Z_{\text{err}}^{(1,3,6)} = 0.04$).

For comparison we again run Cloe, PhyloWGS and PairClone on the same data. Cloe infers three subclones with phylogeny $1 \rightarrow 2 \rightarrow 3$, where Cloe's subclones 1, 2 and 3 roughly correspond to the true subclones 1, 6 and 3 respectively. Cloe's result is reasonable since subclones 2, 4 and 5 have small proportions, and there is

(a) Z

(b) w

(c) Post. prob. of tree topology

(d) $\hat{Z}$

(e) $\hat{w}$
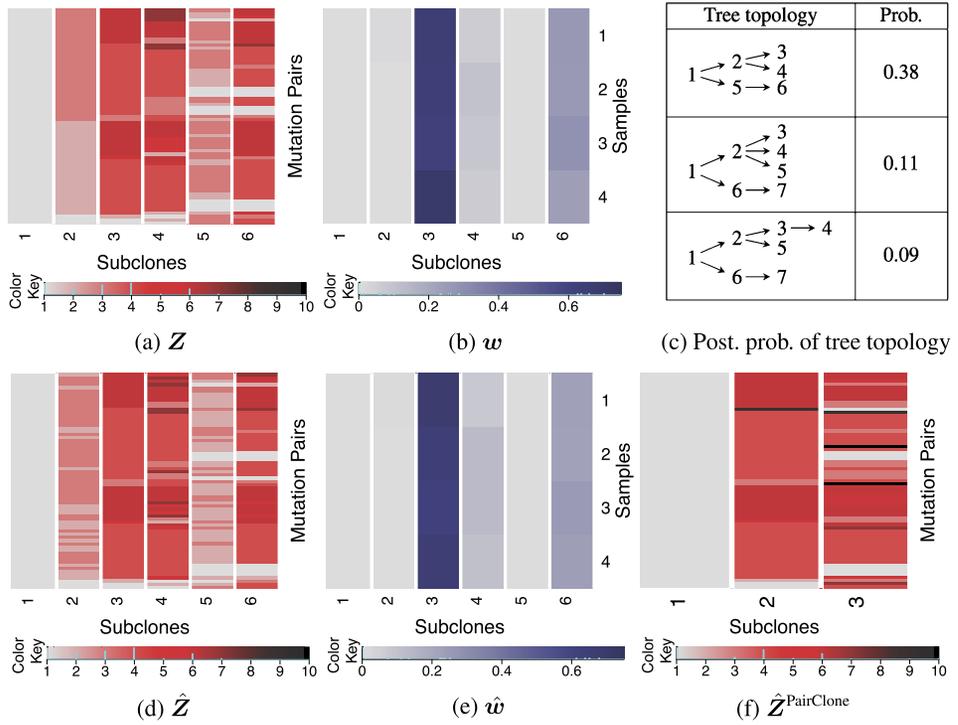
(f) $\hat{Z}^{\text{PairClone}}$

FIG. 5. *Simulation* 3. *Simulation truth Z and w in the form of heatmaps* (a), (b), *three tree topologies with the highest posterior probabilities estimated by TreeClone* (c), *estimates $\hat{Z}$ and $\hat{w}$ under TreeClone* (d), (e) *and estimate $\hat{Z}^{\text{PairClone}}$ under PairClone* (f). *In panels* (a), (d), (f) *colors from light gray to black represent genotypes $z^{(1)}$ to $z^{(10)}$. In panels* (b), (e) *the colors indicate the abundance of a subclone in a sample.*

not much statistical power to estimate them. Similar to the definition of Z_{err}, we define $Z_{\text{err, Cloe}}$ for SNVs. Comparing $\hat{Z}^{\text{Cloe}}$ with subclones 1, 6 and 3, we calculate the reconstruction error $Z_{\text{err,Cloe}}^{(1,3,6)} = 0.04$, indicating good model fit of Cloe. By allowing mutation loss Cloe infers a linear phylogenetic tree which is still reasonable. On the other hand PhyloWGS infers the phylogeny as $0 \to 1 \to 2 \begin{smallmatrix} \to 3 \\ \to 4 \end{smallmatrix}$ (details not shown) which approximates but misses some detail in the simulation truth. Finally, PairClone infers three subclones corresponding to the true subclones 1, 3 and 6, shown in Figure 5(f). PairClone also reasonably recovers the truth but does not infer phylogeny. Comparing $\hat{Z}^{\text{PairClone}}$ with subclones 1, 3 and 6 we calculate the reconstruction error $Z_{\text{err, PairClone}}^{(1,3,6)} = 0.14$ which is higher than $Z_{\text{err}}^{(1,3,6)}$.

5. Lung cancer data. We use whole-exome sequencing (WES) data generated from four ($T = 4$) surgically dissected tumor samples taken from a single

patient diagnosed with lung adenocarcinoma. DNA is extracted from all four samples, and the exome library is sequenced on an Illumina HiSeq 2000 platform in paired-end fashion. Each read is 100 base pairs long. We use BWA [Li and Durbin (2009)] and GATK's UniformGenotyper [McKenna et al. (2010)] for mapping and variant calling respectively. In order to find mutation pair locations along with their genotypes and the number of reads supporting them, we use a bioinformatics tool called LocHap [Sengupta et al. (2016)]. LocHap searches for two or three SNVs that are scaffolded by the same reads. When the scaffolded SNVs, known as local haplotypes, exhibit more than two haplotypes, it is known as local haplotype variant (LHV). Using the individual BAM and VCF files, LocHap finds a few hundreds LHVs on average in a WES sample. We select LHVs with two SNVs as we are interested in mutation pairs only. On those LHVs we run the bioinformatics filters suggested by LocHap to keep the mutation pairs with high calling quality. We focus our analysis in copy number neutral regions. In the end we get 69 mutation pairs for the sample and record the read count data from the output of LocHap. Figure 6 shows the histograms of read depths, left missing rates and right missing rates. The average read depth, left missing rate and right missing rate are 156, 0.21 and 0.23 respectively. Simulation 1 showed that with $T = 4$ samples TreeClone should provide useful inference with this combination of moderate read depth and left/right missing rates.

We use the same hyperparameters as in Simulation 1. We first fit the dataset with $C_{\max} = 5$ and found that the posterior distribution of C is concentrated at the value 5. This suggests that setting $C_{\max} = 5$ is too low; see more details in Supplementary Section 3 (Zhou et al. (2019)). To explore a larger tree space, we fit the dataset again with $C_{\max} = 7$. We run a total of 30,000 MCMC iterations and discard the first 3000 draws as initial burn-in. Figure 7(c) shows the top three tree

(a) Histogram of $\{N_{tk}\}$ (b) Histogram of $\{v_{tk2}\}$ (c) Histogram of $\{v_{tk3}\}$
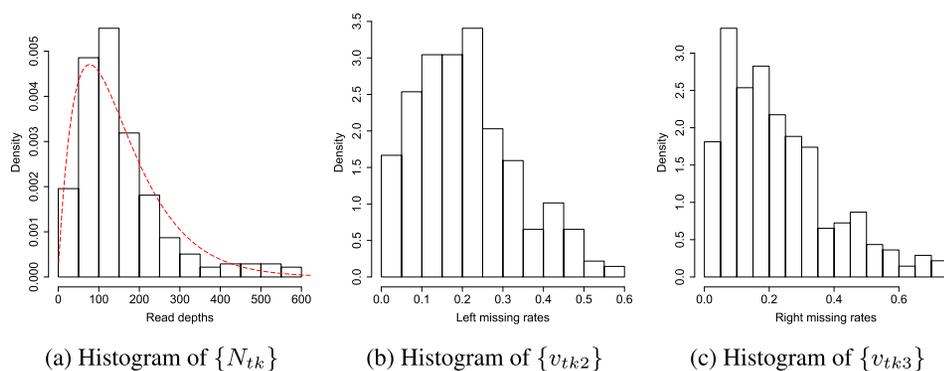
FIG. 6.   *Some summary plots of the lung cancer dataset. Histograms of read depths* (a), *left missing rates* (b) *and right missing rates* (c). *The dashed line in* (a) *is a negative binomial density, showing that it is not unreasonable to assume the read depths follow a negative binomial distribution.*

(a) $\hat{\boldsymbol{Z}}$ (b) $\hat{\boldsymbol{w}}$ (c) Post. prob. of tree topology

(d) Histogram of $\left(\hat{p}_{tkg} - \bar{p}_{tkg}\right)$ (e) Q-Q plot of R^B versus χ_7^2
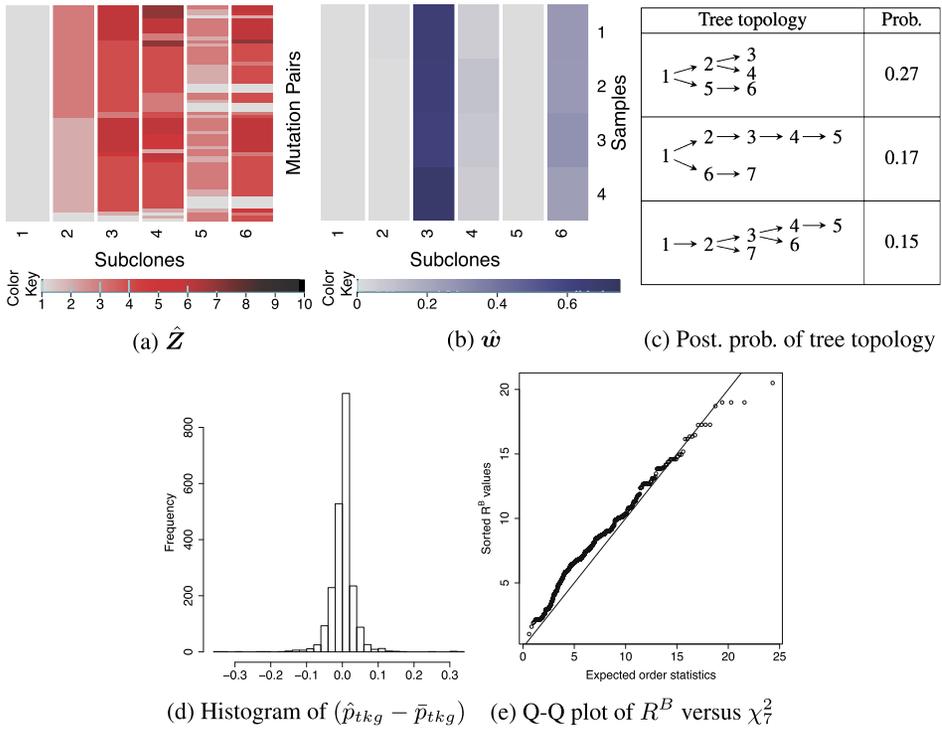
FIG. 7. *Posterior inference with TreeClone for the lung cancer dataset. Estimated genotype matrix $\hat{\boldsymbol{Z}}$ (a), cellular proportions $\hat{\boldsymbol{w}}$ (b) and three tree topologies with the highest posterior probabilities (c). Panel (d) shows a histogram of residuals which are the differences between the estimated and empirical values of p_{tkg}. Panel (e) is a quantile-quantile (Q-Q) plot of posterior samples of the test statistic R^B against expected order statistics from a χ_7^2 distribution for the Bayesian χ^2 test.*

topologies and corresponding posterior probabilities. The posterior mode is
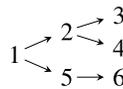
with $\hat{C} = 6$ subclones. Based on our experiences, we recommend to increase $C_{\max}$ if the posterior distribution is peaked at the current $C_{\max}$ value to avoid unnecessary truncation. Figure 7(a), (b) show the estimated subclone genotypes $\hat{\boldsymbol{Z}}$ and cellular proportions $\hat{\boldsymbol{w}}$ respectively ($\hat{w}_{t0} < 3 \times 10^{-3}$ are not shown). The rows for $\hat{\boldsymbol{Z}}$ are reordered for better display. The cellular proportions of the subclones show strong similarity across the four samples, indicating homogeneity of the samples. This is plausible as the samples are dissected from proximal sites. Subclone 1, which is the normal subclone, takes a small proportion in all four samples, indicating high purity of the tumor samples. Subclones 2 and 5 are also included in only small proportions. They have almost vanished in the samples. However, as

parents of subclones 3, 4 and 6 respectively, they are important for the reconstruction of the subclone phylogeny. Subclones 3, 4 and 6 are the three main subclones. They share a large proportion of common mutations, but each one has some private mutations, consistent with the estimated tree.

Test of fit. Finally, Figure 7(d) shows a histogram of residuals, where we calculate empirical values $\bar{p}_{tkg} = n_{tkg}/N_{tk}$ and plot the difference $(\hat{p}_{tkg} - \bar{p}_{tkg})$. The residuals are centered at zero with little variation, indicating a good model fit. For a more formal goodness-of-fit test, we carry out the Bayesian χ^2 test proposed in Johnson (2004). Recall that the observations in our case are short reads s_{tki} taking $G = 8$ discrete values $\{00, 01, 10, 11, -0, -1, 0-, 1-\}$. We count the number of short reads that fall into each of these categories. Let M_g denote these counts, and let $x^{(l)}$ be a posterior sample of $x = (Z, w, \rho)$. The statistic R^B is defined as

$$R^B(x^{(l)}) = \sum_{g=1}^{G} \left[\frac{M_g - Nq_g(x^{(l)})}{\sqrt{Nq_g(x^{(l)})}} \right]^2,$$

where $N = \sum_{t,k} N_{tk}$, and $q_g(x^{(l)}) = \sum_{t,k} N_{tk} p_{tkg}(x^{(l)})/N$ is the expected proportion of short reads in category g calculated by $x^{(l)}$. Under the null hypothesis of a good model fit, the statistic should follow a χ^2-distribution with $G - 1 = 7$ degrees of freedom. Figure 7(e) shows a quantile-quantile plot of posterior samples of R^B against expected order statistics from a χ_7^2 distribution. In addition we find the proportion of posterior samples of R^B exceeding the 95% quantile of a χ_7^2 distribution to be 0.054. There is no evidence of a lack of fit.

Cloe and PhyloWGS. For comparison we run Cloe and PhyloWGS on the same dataset with default hyperparameters. Cloe infers four subclones with phylogeny $1 \rightarrow 2 \rightarrow 3 \rightarrow 4$. Figure 8(a), (b) show the estimated genotypes $\hat{Z}^{\text{Cloe}}$ and cellular proportions $\hat{w}^{\text{Cloe}}$ respectively. The estimated subclones 2, 3 and 4 under Cloe match with subclones 6, 4 and 3 respectively, under TreeClone. Cloe infers a linear phylogenetic tree since it allows mutation loss. PhyloWGS estimates six clusters (and a cluster 0 for normal subclone) of the SNVs with phylogeny:

$$0 \rightarrow 1 \rightarrow 2 \begin{smallmatrix} \nearrow 3 \rightarrow 4 \\ \searrow 5 \rightarrow 6. \end{smallmatrix}$$
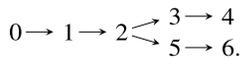
Figure 8(c) summarizes the cluster sizes and cellular prevalences. In light of the earlier simulation studies, we believe that the inference on $\mathcal{T}$ under TreeClone is more reliable. However, results from Cloe and PhyloWGS confirm that the four samples have similar proportions of all the subclones, indicating little intersample heterogeneity. Also, Cloe and PhyloWGS infer very small normal cell proportion, confirming the high tumor purity found by TreeClone. Finally, the same lung cancer dataset was analyzed under PairClone in Zhou et al. (2019). PairClone infers

(a) $\hat{\boldsymbol{Z}}^{\text{Cloe}}$ (b) $\hat{\boldsymbol{w}}^{\text{Cloe}}$ (c) Cellular prevalence (PhyloWGS)

Samples / Clusters (sizes)	1	2	3	4
0 (0)	1.0	1.0	1.0	1.0
1 (93)	0.977	0.990	0.989	0.980
2 (32)	0.854	0.816	0.832	0.852
3 (6)	0.532	0.431	0.473	0.596
4 (3)	0.433	0.408	0.445	0.386
5 (3)	0.252	0.180	0.229	0.152
6 (1)	0.112	0.108	0.056	0.127
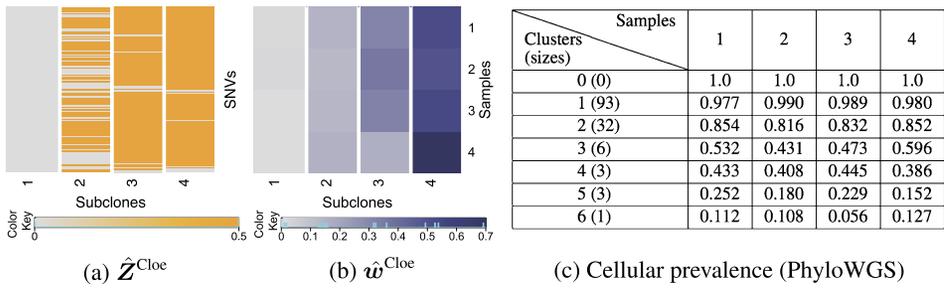
FIG. 8. *Posterior inference with Cloe and PhyloWGS for the lung cancer dataset. Panels* (a), (b) *show the estimated genotype matrix* $\hat{\boldsymbol{Z}}^{\text{Cloe}}$ *and cellular proportions* $\hat{\boldsymbol{w}}^{\text{Cloe}}$ *under Cloe. Panel* (c) *reports the SNV clusters*, *cluster sizes and cellular prevalences inferred by PhyloWGS.*

three subclones corresponding to TreeClone's subclones 1, 3 and 6 and also confirms the four tissue samples are highly homogenuous. PairClone gives reasonable result but can not infer phylogeny.

6. Discussion. In this work, using a tree-based LFAM, we infer subclonal genotypes structure for mutation pairs, their cellular proportions and the phylogenetic relationship among subclones. This is the first attempt to generate a subclonal phylogenetic structure using mutation pair data. We show that more accurate inference can be obtained using mutation pairs data compared to using only marginal counts for single SNVs. The model can be easily extended to incorporate more than two SNVs. Another way of extending the model is to encode mutation times inside the length of the edges of phylogenetic tree.

Currently, the heterogeneity is measured mostly with SNV and CNA data. However, structural variants (SVs) like deletion, duplication, inversion, translocation and other large genome rearrangement arguably provide more accurate data for VAF estimation [Fan et al. (2014)] which is the key input for characterizing tumor heterogeneity. Therefore, incorporation of SVs into the current model could significantly improve the outcome of tumor heterogeneity analysis. Recently, Brocks et al. (2014) attempted to explain intratumor heterogeneity in DNA methylation and copy-number pattern by a unified evolutionary process. The current genome centric definition of tumor heterogeneity could be extended by incorporation of methylation, DNA mutation and RNA expression data in an integromics model.

Finally, in the era of big data it is important to factor computation into the research effort and build efficient computational models that could handle millions of SNVs. Linear response variational Bayes [Giordano, Broderick and Jordan (2015)] or MAD-Bayes [Broderick, Kulis and Jordan (2013), Xu et al. (2015)] methods could be considered as alternative computational strategies to tackle the problem.

SUPPLEMENTARY MATERIAL

Supplement to "TreeClone: Reconstruction of Tumor Subclone Phylogeny Based on Mutation Pairs using Next Generation Sequencing Data" (DOI: 10.1214/18-AOAS1224SUPP; .zip). We provide the R package TreeClone, a glossary of biological terms and the supplementary details referenced in the main text.

REFERENCES

ADAMS, R. P., GHAHRAMANI, Z. and JORDAN, M. I. (2010). Tree-structured stick breaking for hierarchical data. In *Advances in Neural Information Processing Systems* 19–27.

APARICIO, S. and CALDAS, C. (2013). The implications of clonal genome evolution for cancer medicine. *N. Engl. J. Med.* **368** 842–851.

BAFNA, V., GUSFIELD, D., LANCIA, G. and YOOSEPH, S. (2003). Haplotyping as perfect phylogeny: A direct approach. *J. Comput. Biol.* **10** 323–340.

BONAVIA, R., CAVENEE, W. K., FURNARI, F. B. et al. (2011). Heterogeneity maintenance in glioblastoma: A social network. *Cancer Res.* **71** 4055–4060.

BROCKS, D., ASSENOV, Y., MINNER, S., BOGATYROVA, O., SIMON, R., KOOP, C., OAKES, C., ZUCKNICK, M., LIPKA, D. B., WEISCHENFELDT, J. et al. (2014). Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Rep.* **8** 798–806.

BRODERICK, T., KULIS, B. and JORDAN, M. (2013). MAD-Bayes: MAP-based asymptotic derivations from Bayes. In *Proceedings of the* 30*th International Conference on Machine Learning* 226–234.

CARTER, S. L., CIBULSKIS, K., HELMAN, E., MCKENNA, A., SHEN, H., ZACK, T., LAIRD, P. W., ONOFRIO, R. C., WINCKLER, W. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30** 413–421.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (1998). Bayesian CART model search. *J. Amer. Statist. Assoc.* **93** 935–948.

DAGUM, L. and MENON, R. (1998). OpenMP: An industry standard API for shared-memory programming. *IEEE Comput. Sci. Eng.* **5** 46–55.

DENISON, D. G. T., MALLICK, B. K. and SMITH, A. F. M. (1998). A Bayesian CART algorithm. *Biometrika* **85** 363–377. MR1649118

DESHWAR, A. G., VEMBU, S., YUNG, C. K., JANG, G. H., STEIN, L. and MORRIS, Q. (2015). PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16** 35.

FAN, X., ZHOU, W., CHONG, Z., NAKHLEH, L. and CHEN, K. (2014). Towards accurate characterization of clonal heterogeneity based on structural variation. *BMC Bioinform.* **15** 1.

FISCHER, A., VÁZQUEZ-GARCÍA, I., ILLINGWORTH, C. J. R. and MUSTONEN, V. (2014). High-definition reconstruction of clonal composition in cancer. *Cell Rep.* **7** 1740–1752.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* 457–472.

GEYER, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics*, *Proceedings of the* 23*rd Symposium on the Interface* 156–163. Interface Foundation of North America, Fairfax Station, VA.

GIORDANO, R. J., BRODERICK, T. and JORDAN, M. I. (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Advances in Neural Information Processing Systems* 1441–1449.

JIAO, W., VEMBU, S., DESHWAR, A. G., STEIN, L. and MORRIS, Q. (2014). Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinform.* **15** 35.

JOHNSON, V. E. (2004). A Bayesian χ^2 test for goodness-of-fit. *Ann. Statist.* **32** 2361–2384. MR2153988

LEE, J., MÜLLER, P., GULUKOTA, K. and JI, Y. (2015). A Bayesian feature allocation model for tumor heterogeneity. *Ann. Appl. Stat.* **9** 621–639. MR3371328

LI, H. and DURBIN, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25** 1754–1760.

MARASS, F., MOULIERE, F., YUAN, K., ROSENFELD, N. and MARKOWETZ, F. (2016). A phylogenetic latent feature model for clonal deconvolution. *Ann. Appl. Stat.* **10** 2377–2404. MR3592061

MARUSYK, A., ALMENDRO, V. and POLYAK, K. (2012). Intra-tumour heterogeneity: A looking glass for cancer? *Nat. Rev. Cancer* **12** 323–334.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20** 1297–1303.

MILLER, K. T., GRIFFITHS, T. L. and JORDAN, M. I. (2008). The phylogenetic Indian buffet process: A non-exchangeable nonparametric prior for latent features. In *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence* 403–410.

MILLER, C. A., WHITE, B. S., DEES, N. D., GRIFFITH, M., WELCH, J. S., GRIFFITH, O. L., VIJ, R., TOMASSON, M. H., GRAUBERT, T. A., WALTER, M. J. et al. (2014). SciClone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol.* **10** e1003665.

NIK-ZAINAL, S., VAN LOO, P., WEDGE, D. C., ALEXANDROV, L. B., GREENMAN, C. D., LAU, K. W., RAINE, K., JONES, D., MARSHALL, J., RAMAKRISHNA, M. et al. (2012). The life history of 21 breast cancers. *Cell* **149** 994–1007.

NOWELL, P. C. (1976). The clonal evolution of tumor cell populations. *Science* **194** 23–28.

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison. *J. Roy. Statist. Soc. Ser. B* **57** 99–138. MR1325379

ROTH, A., KHATTRA, J., YAP, D., WAN, A., LAKS, E., BIELE, J., HA, G., APARICIO, S., BOUCHARD-CÔTÉ, A. and SHAH, S. P. (2014). PyClone: Statistical inference of clonal population structure in cancer. *Nat. Methods* **11** 396–398.

SCHWARZ, R. F., NG, C. K., COOKE, S. L., NEWMAN, S., TEMPLE, J., PISKORZ, A. M., GALE, D., SAYAL, K., MURTAZA, M., BALDWIN, P. J. et al. (2015). Spatial and temporal heterogeneity in high-grade serous ovarian cancer: A phylogenetic analysis. *PLoS Med.* **12** e1001789.

SENGUPTA, S., WANG, J., LEE, J., MÜLLER, P., GULUKOTA, K., BANERJEE, A. and JI, Y. (2015). BayClone: Bayesian nonparametric inference of tumor subclones using NGS data. In *Proceedings of the Pacific Symposium on Biocomputing* (*PSB*) **20** 467–478.

SENGUPTA, S., GULUKOTA, K., ZHU, Y., OBER, C., NAUGHTON, K., WENTWORTH-SHEILDS, W. and JI, Y. (2016). Ultra-fast local-haplotype variant calling using paired-end DNA-sequencing data reveals somatic mosaicism in tumor and normal blood samples. *Nucleic Acids Res.* **44** e25.

VAN LOO, P., NORDGARD, S. H., LINGJÆRDE, O. C., RUSSNES, H. G., RYE, I. H., SUN, W., WEIGMAN, V. J., MARYNEN, P., ZETTERBERG, A., NAUME, B. et al. (2010). Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. USA* **107** 16910–16915.

XU, Y., MÜLLER, P., YUAN, Y., GULUKOTA, K. and JI, Y. (2015). MAD Bayes for tumor heterogeneity—feature allocation with exponential family sampling. *J. Amer. Statist. Assoc.* **110** 503–514. MR3367243

ZARE, H., WANG, J., HU, A., WEBER, K., SMITH, J., NICKERSON, D., SONG, C., WITTEN, D., BLAU, C. A. and NOBLE, W. S. (2014). Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput. Biol.* **10** e1003703.

ZHOU, T., MÜLLER, P., SENGUPTA, S. and JI, Y. (2019). PairClone: A Bayesian subclone caller based on mutation pairs. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 705–725.

T. ZHOU
Y. JI
DEPARTMENT OF PUBLIC HEALTH SCIENCES
THE UNIVERSITY OF CHICAGO
5841 SOUTH MARYLAND AVE MC2000
CHICAGO, ILLINOIS 60637
USA
E-MAIL: tjzhou@uchicago.edu
          yji@health.bsd.uchicago.edu

S. SENGUPTA
RESEARCH INSTITUTE
NORTHSHORE UNIVERSITY HEALTHSYSTEM
1001 UNIVERSITY PL
EVANSTON, ILLINOIS 60201
USA
E-MAIL: subhajit06@gmail.com

P. MÜLLER
DEPARTMENT OF MATHEMATICS
THE UNIVERSITY OF TEXAS AT AUSTIN
1 UNIVERSITY STATION C1200
AUSTIN, TEXAS 78712
USA
E-MAIL: pmueller@math.utexas.edu