The Annals of Applied Statistics
2015, Vol. 9, No. 2, 621–639
DOI: 10.1214/15-AOAS817

A BAYESIAN FEATURE ALLOCATION MODEL FOR TUMOR HETEROGENEITY

BY JUHEE LEE*, PETER MÜLLER[†,1], KAMALAKAR GULUKOTA[‡] AND YUAN JI[‡,§,1]

*University of California Santa Cruz**, *University of Texas, Austin*[†], *NorthShore University HealthSystem*[‡] *and University of Chicago*[§]

We develop a feature allocation model for inference on genetic tumor variation using next-generation sequencing data. Specifically, we record single nucleotide variants (SNVs) based on short reads mapped to human reference genome and characterize tumor heterogeneity by latent haplotypes defined as a scaffold of SNVs on the same homologous genome. For multiple samples from a single tumor, assuming that each sample is composed of some sample-specific proportions of these haplotypes, we then fit the observed variant allele fractions of SNVs for each sample and estimate the proportions of haplotypes. Varying proportions of haplotypes across samples is evidence of tumor heterogeneity since it implies varying composition of cell subpopulations. Taking a Bayesian perspective, we proceed with a prior probability model for all relevant unknown quantities, including, in particular, a prior probability model on the binary indicators that characterize the latent haplotypes. Such prior models are known as feature allocation models. Specifically, we define a simplified version of the Indian buffet process, one of the most traditional feature allocation models. The proposed model allows overlapping clustering of SNVs in defining latent haplotypes, which reflects the evolutionary process of subclonal expansion in tumor samples.

1. Introduction. We propose a feature allocation model [Broderick, Jordan and Pitman (2013, 2013)] to describe tumor heterogeneity using next-generation sequencing (NGS) data. We use a variation of the Indian buffet process (IBP) [Griffiths and Ghahramani (2005), Teh, Görür and Ghahramani (2007)]. The feature allocation in our model is latent. That is, the features are not directly observed. We record point mutations as single nucleotide variants (SNVs), each of which is defined as a DNA locus that possesses a variant sequence from that on the reference human genome. We use the feature allocation model to describe unobserved haplotypes, defined as a collection of single nucleotide variants (SNVs) scaffolded on a homologous genome. In a tumor sample, having more than two haplotypes is evidence of heterogeneous cell subpopulations with a distinct genome. This is the case because humans are diploid and we would therefore only observe up to two

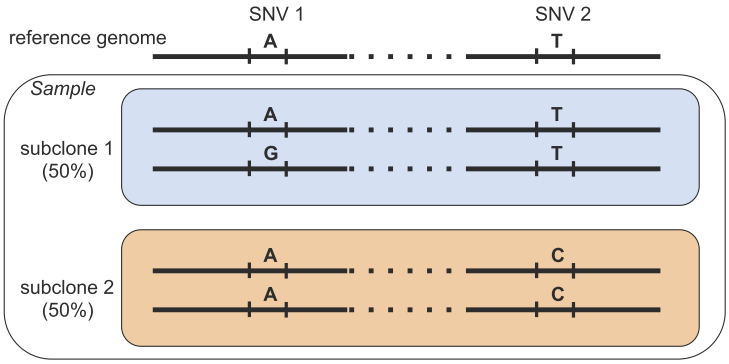
Received July 2014; revised January 2015.
[1]Supported in part by NIH R01 CA132897.
Key words and phrases. Haplotypes, feature allocation models, Indian buffet process, Markov chain Monte Carlo, next-generation sequencing, random binary matrices, variant calling.

haplotypes if all cells in a tumor sample were genetically homogeneous. In the proposed application of feature allocation models to inference for tumor heterogeneity, the haplotypes are the features and the SNVs are the experimental units that select the features. The number of features is unknown. Each tumor sample is composed of an unknown proportion of each of these haplotypes. The top level sampling model for the observed SNV counts is then defined as binomial sampling with a proportion for each SNV that is implied by this composition. In summary, we solve a deconvolution problem to explain the observed SNV frequencies for each sample by compositions of latent haplotypes.

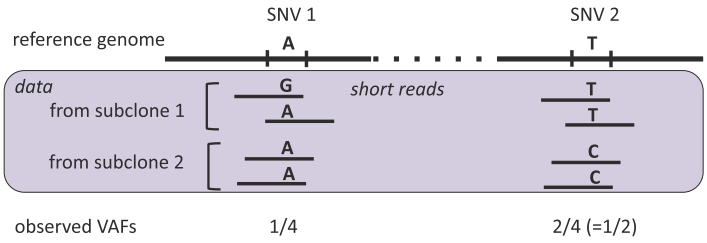
Heterogeneity in cancer tissue has been hypothesized over the past few decades [Wersto et al. (1991)] and has been demonstrated elegantly using NGS technology over the past few years [Gerlinger et al. (2012)]. Genetic variation in a tumor occurs due to evolutionary processes that drive tumor progression. Specifically, tumors include distinct clonal subpopulations of cells that arise stochastically by a sequence of randomly acquired mutations. Substantial genetic heterogeneity between tumors (inter-tumor) or within a tumor (intra-tumor) can be explained by differences in clonal subpopulations and varying proportions of those subpopulations [Landau et al. (2013), Marusyk and Polyak (2010), Russnes et al. (2011)]. For example, Navin et al. (2010) reported clonal genomic heterogeneity in breast cancers.

Data derived from NGS experiments include SNVs, small indels and copy number variations [Ng and Kirkness (2010), Wheeler et al. (2008)]. Many researchers use SNV data to investigate genes and genomic regions related to cancer phenotypes [Engle, Simpson and Landers (2006), Erichsen and Chanock (2004)]. In this paper, we utilize whole-genome sequencing data measuring variant allele fractions (VAFs) at SNVs to understand tumor heterogeneity by proposing inference on how haplotypes may be distributed within a tumor.

In an NGS experiment, millions of short DNA reads are generated and are then aligned to the reference genome. At certain positions of the genome, some or all of the mapped reads will show a sequence different from what is in the reference genome. At each genomic locus, the proportion of short reads bearing a variant sequence is called the VAF. If the VAF at a locus is nonzero, an SNV may be "called" at that locus, based on statistical inference [Li et al. (2009)]. The raw experimental data comprises the total number of reads (N) that are mapped to the locus and the number of those reads (n) indicating a variation from the reference sequence. Then $\text{VAF} = n/N$. If a tumor sample is homogeneous, that is, having a single clone, the VAF values of all the SNVs should be close to 0, 0.5 or 1, reflecting the three possible homozygous and heterozygous alleles (i.e., AA, AB, BB) at any SNV. Different VAF values imply heterogeneity of the cellular genome in the tumor sample (see Figure 1 for an example). We propose to study inference to deconvolute the VAFs from multiple SNVs and back out the latent haplotypes.



(a) Multiple haplotypes as evidence of a heterogeneous tumor



(b) Hypothetical short reads data

FIG. 1. *A hypothetical example explaining how NGS data can be used to infer heterogeneous tumor samples*. (a) *shows that there are two subclones* (*cell subpopulations*) *in the tumor sample with different haplotypes consisting of two SNVs*: *For subclone* 1, *there are two haplotypes*, *AT and GT*. *For subclone* 2, *there is only one haplotype*, *AC*. *Thus*, *there are a total of three haplotypes in the tumor sample*, *implying heterogeneous cell populations since a population of homogeneous cells would only support up to two haplotypes*. *Here sequence G for SNV* 1 *and sequence C for SNV* 2 *are mutations*. (b) *shows hypothetical short reads for this sample if it is sequenced*, *assuming that the proportions of the two subclones are equal*. *The short reads counts are summarized as observed VAFs*, *which are used for our statistical inference*.

We propose a Bayesian feature allocation model to characterize such cellular heterogeneity in a way that explains the observed NGS data. We construct a matrix of binary features (equivalently, haplotypes) as shown in Figure 2. In the figure, columns correspond to haplotypes and rows correspond to SNVs. We define haplotype c by a binary vector $(z_{1c}, \ldots, z_{Sc})$ of indicators of whether $(z_{sc} = 1)$ or not $(z_{sc} = 0)$ a variant sequence is observed at the SNV s. Here we view SNV as a genetic locus on which either a variant or reference DNA sequence could be observed. Figure 2 illustrates the definition of five haplotypes ($C = 4$, columns) with $S = 10$ SNVs (rows). In the figure, black (white) indicates $z_{sc} = 1$ ($z_{sc} = 0$). For example, SNV 1 in Figure 2 possesses a variant sequence in the two haplotypes $c = 0$ and $c = 1$. On the other hand, SNV 9 possesses variant sequences in four haplotypes: $c = 0, 1, 2$ and 4. A prior probability model on such a binary ma-

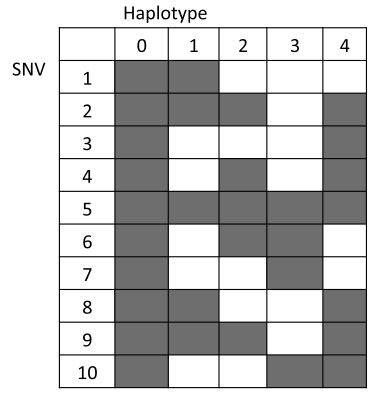


FIG. 2. *An illustration of cell types* (*binary latent features*) *in columns. A black/white block indicates a variant/reference DNA sequence at the corresponding SNV* (*row*) *for the haplotype* (*column*).

trix $\mathbf{Z} = [z_{sc}]$ is known as a feature allocation model. Here, we assume that C is unknown and place a prior on C.

Assuming that samples are composed of different proportions of C haplotypes, we aim to fit the observed VAFs of the SNVs to infer these proportions. For example, we may observe that one sample is dominated by haplotypes 1 and 4, while another is dominated by haplotypes 2 and 3. If the samples are from the same tumor, the differences in haplotypic compositions are evidence of *intra-tumor* heterogeneity; on the other hand, differences in samples from different tumors imply *inter-tumor* heterogeneity. Therefore, the proposed inference provides a unified framework to address inference for both biological concepts. Importantly, the characterization of haplotypes is based on selected SNVs only. Otherwise inference for heterogeneity between tumors in different patients would not be biologically meaningful, as cellular genomes and haplotypes are not expected to be shared across patients. However, for tumors in the same class of disease, SNVs in local disease-related genomic regions may be common to all or some of the tumors, thereby allowing for the proposed inference.

There are currently few approaches that address the problem of tumor heterogeneity. Su et al. (2012) and Larson and Fridley (2013) recognized that a tumor sample is a mixture of normal cells and tumor cells, and developed a method to estimate tumor purity levels for paired tumor-normal tissue samples using DNA sequencing data. None of the two methods considers more than two samples or unpaired samples. PurBayes [Larson and Fridley (2013)] accounts for intra-tumor heterogeneity, but it does not provide inference on the subpopulation configurations as inference on the latent matrix $\mathbf{Z}$ under the proposed model. PyClone [Roth et al. (2014)], a recently published method, proposes inference to cluster SNVs with different VAFs. An underlying assumption of PyClone is that SNVs can be arranged in clusters that inform about subclones. A key component of PyClone is the use of clustering models such as the Dirichlet process for inference on these

clusters. While such clusters are informative about heterogeneity, inference that is provided by PyClone is not meant to identify subclones or haplotypes. The primary aim of PyClone is inference on mutation clusters, defined as a group of SNVs with similar variant allele fractions.

In contrast, our proposed feature allocation model explicitly models the haplotypic genomes of subclones, allowing overlapping SNVs shared between different subclones. We do not use nonoverlapping SNV clusters as the building block for subclones. That is, instead of first estimating the SNV clusters and then constructing subclones based on clusters, we directly infer the subclonal structure based on haplotypes. We show in later examples the distinction between PyClone and our proposed method.

The remainder of the paper is organized as follows: Section 2 describes the proposed Bayesian feature allocation model and a model selection criterion to select the number of subclones. Section 3 describes simulation studies. Sections 4 and 5 report data analyses of in-house data sets to illustrate inter-tumor heterogeneity and intra-tumor heterogeneity, respectively. The last section concludes with a final discussion.

2. Probability model.

2.1. *Sampling model.* Let $\mathbf{n}$ denote an $S \times T$ matrix of count data from an NGS genome sequencing experiment, with n_{st} denoting the number of reads that bear a variant sequence at the location of SNV s for tissue sample t, $s = 1, \ldots, S$ and $t = 1, \ldots, T$. We assume a binomial sampling model. Let N_{st} denote the total number of reads in sample t that are mapped to the genomic location of SNV s. We assume

$$n_{st} \overset{\text{indep}}{\sim} \text{Bin}(N_{st}, p_{st}). \tag{1}$$

In Figure 3, $n_{st} = 3$ and $N_{st} = 5$. We do not model N_{st}, that is, we treat N_{st} as fixed, and only consider a sampling model for n_{st} conditional on N_{st} (modeling

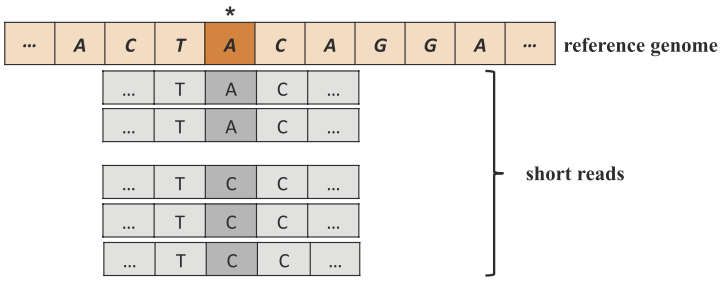


FIG. 3. *An illustration of the Binomial model. The illustration shows that* 5 *short reads are mapped to a position marked with * and among them three reads indicate variation at the position, that is,* $N_{st} = 5$ *and* $n_{st} = 3$.

N_{st} would not contribute any information on tumor heterogeneity based on SNVs). Conditional on N_{st}, the observed counts n_{st} are independent across s and t. The model in (1) is illustrated in Figure 3.

2.2. *Prior.* We build a prior probability model for p_{st} in two steps, using the notion that each sample is composed of a mixture of different haplotypes. And each haplotype, in turn, is characterized by the haplotypes consisting of the SNVs. Let w_{tc} denote the proportion of haplotype c in sample t and let $z_{sc} \in \{0, 1\}$ denote a latent indicator of the event that SNV s bears a variant sequence for haplotype c. Note that $z_{sc} = 1$ corresponds to a black block in Figure 2. Then p_{st} is written as a sum over C latent haplotypes

$$p_{st} = w_{t0}p_0 + \sum_{c=1}^{C} w_{tc}z_{sc} \equiv \varepsilon_{t0} + \sum_{c=1}^{C} w_{tc}z_{sc}. \tag{2}$$

The construction of the haplotypes, including the number of haplotypes, C, and the indicators z_{sc} are latent. The key term, $\sum_{c=1}^{C} w_{tc}z_{tc}$, indirectly infers haplotypes by explaining p_{st} as arising from sample t being composed of a mix of hypothetical haplotypes which do ($z_{sc} = 1$) or do not ($z_{sc} = 0$) include a mutation for SNV s. The indicators z_{sc} are collected in a $(S \times C)$ binary matrix $\mathbf{Z}$. The number of latent haplotypes, C, is unknown. Conditional on C, the binary matrix $\mathbf{Z}$ describes C latent tumor haplotypes present in the observed samples. Joint inference on C, Z and $\mathbf{w}_t$ explains tumor heterogeneity.

In addition, we introduce a background haplotype, labeled as haplotype $c = 0$, which includes all SNVs. The background haplotype accounts for experimental noise and haplotypes that appear with negligible abundance. Specifically, $\varepsilon_{t0} = w_{t0}p_0$ in (2) relates to this background haplotype, with p_0 being the relative frequency of observing a mutation at an SNV due to noise and artifact (we assume equal frequency for all SNVs) and w_{t0} being the proportion in sample t. The prior on w_{t0} is defined later. For p_0, we assume $p_0 \sim \text{Be}(a_{00}, b_{00})$ with $a_{00} \ll b_{00}$ to inform a small p_0 value a priori.

We start the prior construction with a prior for the number of haplotypes, C. We consider a geometric distribution, $C \sim \text{Geometric}(r)$ where $\text{E}(C) = 1/r$. Conditional on C, we use a feature allocation model for a binary matrix $\mathbf{Z}$. We first define the model for any given C and start with feature-specific selection probabilities,

$$\mu_c | C \overset{\text{i.i.d.}}{\sim} \text{Be}(\alpha/C, 1). \tag{3}$$

Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_C)$. The selection probabilities are used to define $p(\mathbf{Z}|\boldsymbol{\mu}, C)$ as

$$\text{p}(\mathbf{Z}|\boldsymbol{\mu}, C) = \prod_{s=1}^{S}\prod_{c=1}^{C} \mu_c^{z_{sc}}(1-\mu_c)^{(1-z_{sc})} = \prod_{c=1}^{C} \mu_c^{m_c}(1-\mu_c)^{S-m_c}, \tag{4}$$

where $m_c = \sum_{s=1}^{S} z_{sc}$ is the number of SNVs bearing variant sequences for haplotype c. A limit of the model, as $C \to \infty$, becomes a constructive definition of

the Indian buffet process (IBP) [Griffiths and Ghahramani (2005), Teh, Görür and Ghahramani (2007)]. The model is symmetric with respect to arbitrary indexing of the SNVs, simply because of the symmetry in (4) and (3). Note that $m_c = 0$ is possible with positive prior probability.

Next, we consider a prior distribution for abundances associated with the haplotypes defined by $\mathbf{Z}$. The haplotypes are common for all tumor samples, but the relative weights in the composition (2) are different across tissue samples. We assume Dirichlet priors for the relative weights w_{tc}, defined as follows. Let θ_{tc} denote an (unscaled) abundance level of haplotype c in tissue sample t. We assume $\theta_{tc}|C \overset{\text{i.i.d.}}{\sim} \text{Gamma}(a, 1)$ for $c = 1, \ldots, C$ and $\theta_{t0} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(a_0, 1)$. We then define

$$w_{tc} = \theta_{tc} \Big/ \sum_{c'=0}^{C} \theta_{tc'}$$

as the relative weight of haplotype c in sample t. This is equivalent to $\mathbf{w}_t|C \overset{\text{i.i.d.}}{\sim} \text{Dir}(a_0, a, \ldots, a)$ for $t = 1, \ldots, T$, where $\mathbf{w}_t = (w_{t0}, w_{t1}, \ldots, w_{tC})$.

Recall the binomial sampling likelihood (1) with success probability, p_{st}. Given C, $\mathbf{Z}$ and $\mathbf{w}$, we define p_{st} in (2). In words, p_{st} is determined by C, $\mathbf{Z}$ and $\mathbf{w}_t$ with the earlier describing the latent haplotypes and the latter specifying the relative abundance of each haplotype in sample t.

2.3. *Posterior simulation.* Let $\mathbf{x} = (\mathbf{Z}, \boldsymbol{\theta}, p_0)$, where $\boldsymbol{\theta} = \{\theta_{tc}\}$. Markov chain Monte Carlo (MCMC) posterior simulation proceeds by sequentially drawing random numbers for the parameters in $\mathbf{x}$. Given C, such MCMC simulation is straightforward. Specifically, Gibbs sampling transition probabilities are used to update z_{sc}, and Metropolis–Hastings transition probabilities are used to update $\boldsymbol{\theta}$ and p_0. It is possible to improve the mixing of the Markov chain by updating all columns in row s of the matrix $\mathbf{Z}$ jointly by means of a Metropolis–Hastings transition probability that proposes changes in the entire row vector $\mathbf{z}_s$.

The construction of transition probabilities that involve a change of C is more challenging, since the dimension of $\mathbf{Z}$ and $\boldsymbol{\theta}$ changes as C varies. We use a reversible jump (RJ) MCMC algorithm for posterior simulation [Green (1995)]. We first define a proposal distribution $q(C, \widetilde{C})$ for C, and then introduce a proposal distribution $q(\widetilde{\mathbf{x}}|\widetilde{C})$ for $\mathbf{x}$ conditional on the proposed $\widetilde{C}$. The latter potentially involves a change in dimension of the parameter vector. We found that high posterior correlation of $\mathbf{Z}$ and $\mathbf{w}$ conditional on C greatly complicated the construction of a practicable RJ scheme. To overcome this, we use an approach similar to Casella and Moreno (2006). We split the data into a minimal training set $(\mathbf{n}', \mathbf{N}')$ with $n'_{st} = b_{st} n_{st}$, $N'_{st} = b_{st} N_{st}$, and a test data set, $(\mathbf{n}'', \mathbf{N}'')$ with $n''_{st} = (1 - b_{st}) n_{st}$ etc. In the implementation we use b_{st} generated from $\text{Be}(25, 975)$. Let $p_1(\mathbf{x}|C) = p(\mathbf{x}|\mathbf{n}', C)$ denote the posterior distribution under C

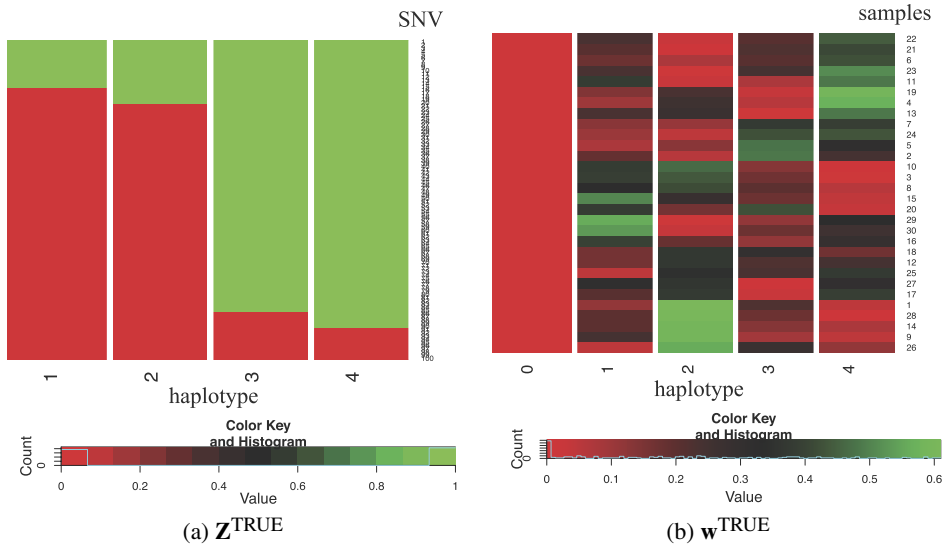
using the training sample. We use p_1 in two instances. First, we replace the original prior $p(\mathbf{x}|C)$ by $p_1(\mathbf{x}|C)$ and, second, we also use it as proposal distribution $q(\widetilde{\mathbf{x}}|\widetilde{C}) = p_1(\widetilde{\mathbf{x}}|\widetilde{C})$. The test data is then used to evaluate the acceptance probability. The strategy can be characterized as model comparison by fractional Bayes factors [O'Hagan (1995)] and is related to a similar approach proposed in Casella and Moreno (2006) for model comparison with intrinsic Bayes factors. Both are originally proposed for model comparison with noninformative priors, but can be modified to facilitate MCMC across models as we need it here.

We summarize the joint posterior distribution, $p(C, \mathbf{Z}, \mathbf{w}, p_0|\mathbf{n})$, by factorizing it as $p(C|\mathbf{n})p(\mathbf{Z}|\mathbf{n}, C)p(\mathbf{w}, p_0|\mathbf{n}, C, \mathbf{Z})$. Based on the available posterior Monte Carlo sample, we (approximately) evaluate the marginal posterior distribution for C and determine the maximum a posteriori (MAP) estimate $C^\star$. We then estimate $\mathbf{Z}$ conditional on $C^\star$ as follows: For any two matrices, $\mathbf{Z}$ and $\mathbf{Z}'$, $1 \le c, c' \le C^\star$, let $\mathcal{D}_{cc'}(\mathbf{Z}, \mathbf{Z}') = \sum_{s=1}^{S} |z_{sc} - z'_{sc'}|$. We then define a distance $d(\mathbf{Z}, \mathbf{Z}') = \min \sum_{c=1}^{C^\star} \mathcal{D}_{c,\pi_c}(\mathbf{Z}, \mathbf{Z}')$, where π_c is a permutation of $\{1, \ldots, C^\star\}$ and the minimum is over all possible permutations. A posterior point estimate for $\mathbf{Z}$ is defined as

$$\mathbf{Z}^\star_C = \arg\min_{\mathbf{Z}'} \int d(\mathbf{Z}, \mathbf{Z}')\, dp(\mathbf{Z}|\mathbf{n}, C^\star) \approx \arg\min_{\mathbf{Z}'} \frac{1}{L} \sum_{\ell=1}^{L} d(\mathbf{Z}^{(\ell)}, \mathbf{Z}'),$$

for a posterior Monte Carlo sample, $\{\mathbf{Z}^{(\ell)}, \ell = 1, \ldots, L\}$. Finally, we report posterior point estimates $\mathbf{w}^\star$ and $p_0^\star$ for $\mathbf{w}$ and p_0 conditional on $C^\star$ and $\mathbf{Z}^\star_C$.

3. Simulation. We validated the proposed model in a simulation study. We simulated a set of $S = 100$ SNVs with $T = 30$ samples. In the simulation truth, we assumed four latent haplotypes ($C^{\mathrm{TRUE}} = 4$) as well as a background haplotype ($c = 0$) with all SNVs bearing variant sequences. Haplotype $c = 1$ has variant sequences for the first 15 SNV positions, haplotype 2 for the first 20 SNV positions, haplotype 3 for the first 85 positions and haplotype 4 for the first 90 positions. In other words, SNVs 1–15 bear variant sequences for all four haplotypes, SNVs 16–20 for haplotypes 2–4, SNVs 21–85 for haplotypes 3–4, SNVs 86–90 for haplotype 4 only and SNVs 91–100 for none of the haplotypes, as shown in Figure 4(a). The green color in panel (a) implies presence ($z_{sc} = 1$) of a variant sequence at SNV s for haplotype c and the red color shows absence ($z_{sc} = 0$), for $c = 1, \ldots, 4$ and $s = 1, \ldots, 100$. We then generated $\mathbf{w}_t^{\mathrm{TRUE}}$ as follows. We let $\mathbf{a}^{\mathrm{TRUE}} = (8, 6, 3, 1)$ and for each t randomly permuted $\mathbf{a}^{\mathrm{TRUE}}$. Let $\mathbf{a}_\pi^{\mathrm{TRUE}}$ denote a random permutation of $\mathbf{a}^{\mathrm{TRUE}}$. We generated $\mathbf{w}^{\mathrm{TRUE}} \sim \mathrm{Dir}(0.2, \mathbf{a}_\pi^{\mathrm{TRUE}})$. That is, the first parameter of the Dirichlet prior for the $(C^{\mathrm{TRUE}} + 1)$-dimensional weight vector was 0.2, and the remaining parameters were a permutation of $\mathbf{a}^{\mathrm{TRUE}}$. Using the assumed $\mathbf{Z}^{\mathrm{TRUE}}$ and $\mathbf{w}^{\mathrm{TRUE}}$ and letting $p_0^{\mathrm{TRUE}} = 0.01$ and $N_{st} = 50$ for all t and s, we generated $n_{st} \sim \mathrm{Bin}(N_{st}, p_{st}^{\mathrm{TRUE}})$ with $p_{st}^{\mathrm{TRUE}} = p_0^{\mathrm{TRUE}} w_{t0}^{\mathrm{TRUE}} + \sum_{c=1}^{C} w_{tc}^{\mathrm{TRUE}} z_{sc}^{\mathrm{TRUE}}$. The weights $\mathbf{w}^{\mathrm{TRUE}}$ are shown



(a) $\mathbf{Z}^{\text{TRUE}}$ (b) $\mathbf{w}^{\text{TRUE}}$

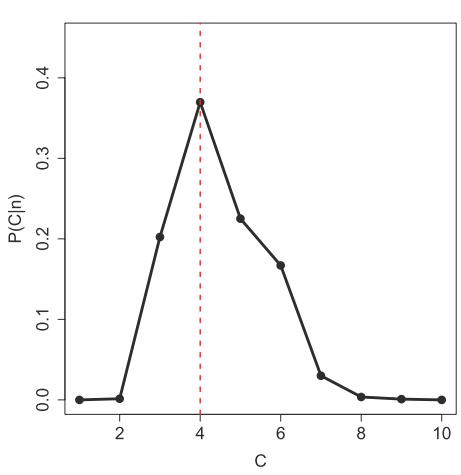
FIG. 4. *Heatmaps of* $\mathbf{Z}^{\text{TRUE}}$ *and* $\mathbf{w}^{\text{TRUE}}$ *in the simulation.*

in Figure 4(b). Similar to the heatmap of $\mathbf{Z}^{\text{TRUE}}$, the green color in panel (b) represents high abundance of a haplotype in a sample and the red color low abundance for $c = 0, \ldots, 4$ and $t = 1, \ldots, 30$. For haplotype 0 the heatmap plots $w_{t0} p_0$. The samples in rows are rearranged for better display.

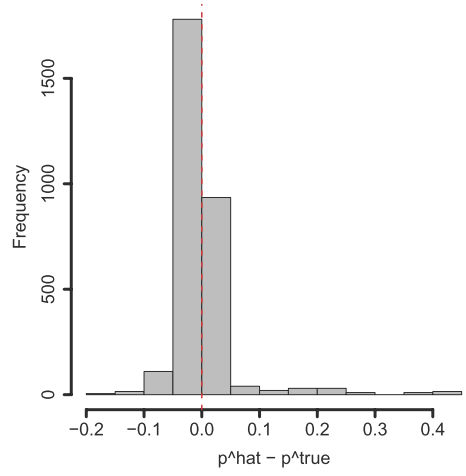
To fit the proposed model, we took $r = 0.2$, $\alpha = 3$, $a_0 = 0.5$, $a = 0.5$, $a_{00} = 1$ and $b_{00} = 100$. For each value of C, we initialized $\mathbf{Z}$ using the observed sample proportions. We generated initial values for $\boldsymbol{\theta}_{tc}$ and p_0 by prior draws. We generated $b_{st} \overset{\text{i.i.d.}}{\sim} \text{Be}(25, 975)$ to construct the minimal training set and ran the MCMC simulation over 25,000 iterations, discarding the first 10,000 iterations as initial burn-in.

Figure 5(a) reports the posterior distribution of C in which the dashed vertical line represents the true value $C^{\text{TRUE}} = 4$. The posterior mode $C^\star = 4$ recovers the truth. We then find the posterior point estimates of $\mathbf{Z}$, $\mathbf{w}$ and p_0 conditional on $C^\star$ as described in Section 2.3. We compared p_{st}^{TRUE} with the posterior estimates $\hat{p}_{st} = p_0^\star w_{t0}^\star + \sum_{c=1}^{C^\star} w_{tc}^\star z_{sc}^\star$. Figure 5(b) shows the histogram of the errors $(\hat{p}_{st} - p_{st}^{\text{TRUE}})$. Fitting appears to be great as $(\hat{p}_{st} - p_{st})$ centers at 0. Figure 5(c) and (d) show heatmaps for $\mathbf{Z}_C^\star$ and $\mathbf{w}_C^\star$ (given $C^\star = 4$). The estimate $\mathbf{Z}_C^\star$ in Figure 5(c) places SNV 86–90 into haplotypes 3 and 4. The latter are two identical haplotypes. This may be because w_{t3}^{TRUE} for haplotype $c = 3$ is small for almost all the samples. The weights for the two dominant subclones, $w_{tc}^\star$ for $c = 1, 2$, are close to the simulation truth, and $w_{tc}^\star$ for $c = 3, 4$ are closer to the average of w_{tc}^{TRUE}, $c = 3, 4$.

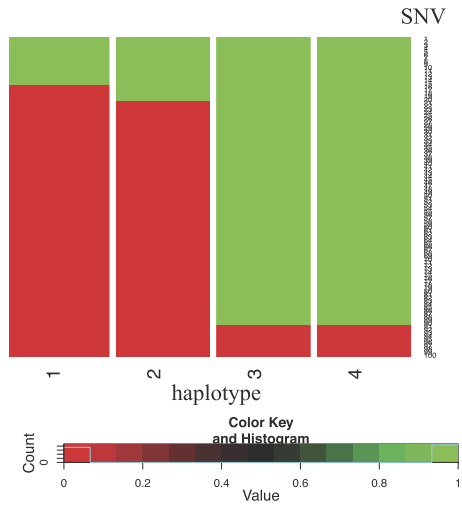
For comparison, we implemented PyClone [Roth et al. (2014)] with the same simulated data. We used the infinite beta-binomial mixture model in PyClone assuming that the copy number at mutation positions is known. Figure 6(a) shows



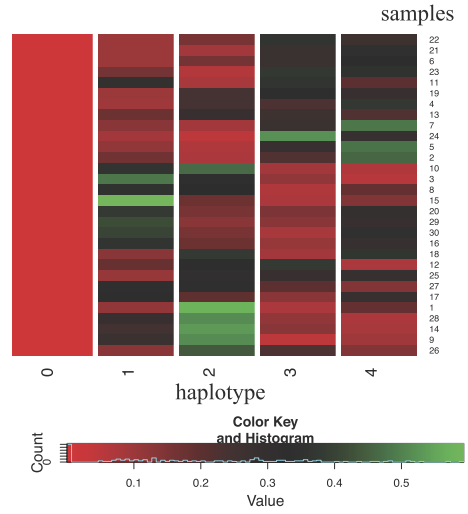
(a) Posterior distribution of C



(b) Histogram of $\hat{p}_{st} - p_{st}^{\text{TRUE}}$



(c) Heatmap of $\mathbf{Z}_C^\star$ with $C^\star = 4$

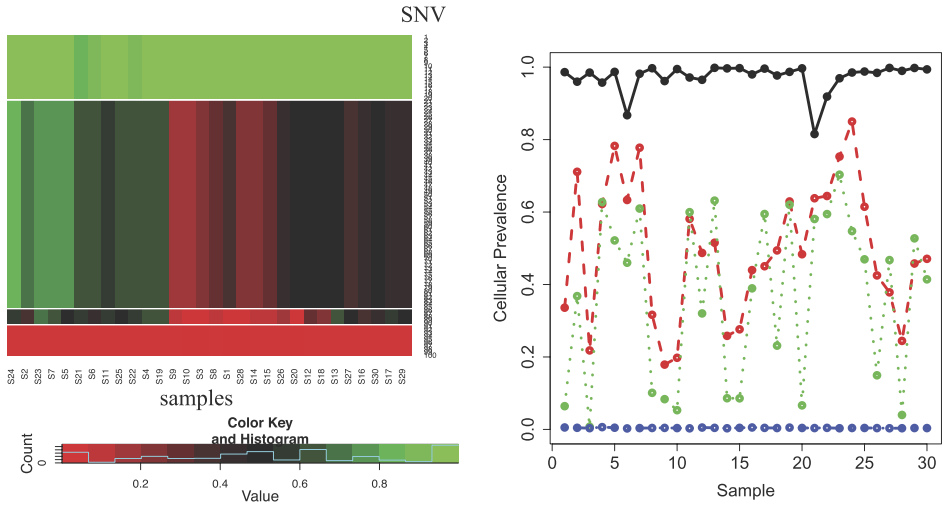


(d) Heatmap of $\mathbf{w}_C^\star$ with $C^\star = 4$ and $\mathbf{Z}_C^\star$

FIG. 5. *Posterior inference for the simulated data.*

the estimated variant allelic prevalence for each mutation *for each sample* under PyClone. Columns are samples and rows are SNVs. The white horizontal lines separate the estimated SNV clusters. PyClone identified four clusters of SNVs: cluster 1 with SNV 1–20, cluster 2 with SNV 21–85, cluster 3 with SNV 86–90 in cluster 3, and cluster 4 with SNV 91–100.

The estimated cluster 1 includes the SNVs that under the simulation truth appear in all the true haplotypes or in true haplotypes 2–3; cluster 2 includes the SNVs from true haplotypes 3–4; cluster 3 includes SNVs from true haplotype 4; and cluster 4 includes SNVs that appear in none of the true haplotypes. Figure 6(b)



(a) Heatmap of cellular prevalence (b) Mean prevalence for each cluster

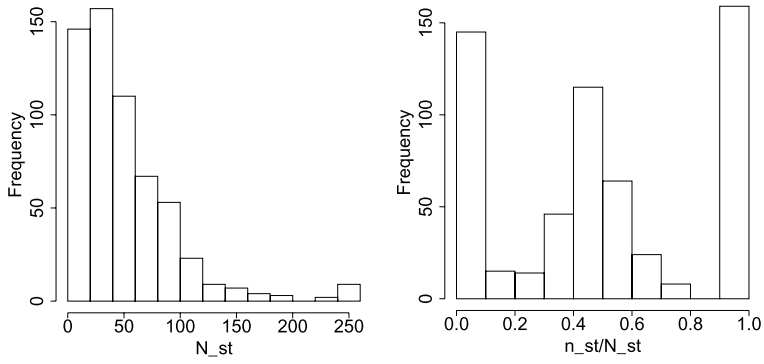
FIG. 6. *Estimated cellular prevalence of four SNV clusters over samples by PyClone for the simulated data.*

shows estimated mean cellular prevalence of each cluster across the 30 samples. In summary, the reconstruction under PyClone is reasonable, but stops short of recovering the true subclones, which cannot possibly be represented as the assumed nonoverlapping clusters.

Finally, we carried out another simulation to the sensitivity of the proposed inference to different assumptions on experimental noise. In particular, we considered experimental noise that varies across SNVs, as it could arise from potential bias or errors in data preprocessing, including sequencing bias, mapping bias, errors in variant calling etc. Details of the simulation study are reported in the supplementary material [Lee et al. (2015)]. Briefly summarized, in the simulation truth we replaced the error term ε_{t0} in (2) by an SNV-specific term ε_{ts}. But we continued to fit the model with the common ε_{t0}, as in (2). We still find reasonable posterior inference.

For details, refer to the supplemental material.

4. Pancreatic cancer data. We analyzed NGS data obtained from exome sequencing of five samples of pancreatic ductal adenocarcinoma (PDAC) patients at NorthShore hospital. PDAC is a particularly aggressive tumor with median survival of less than a year. We extracted genomic DNA from each tissue and constructed an exome library from these DNA using Agilent SureSelect capture probes. The exome library was then sequenced in paired-end fashion on an Illumina HiSeq 2000 platform. About 60 million reads—each 100 bases long—were obtained. Since the SureSelect exome was about 50 Mega bases, raw (premapping) coverage was about 120-fold. We then mapped the reads to the human



(a) N_{st} (b) n_{st}/N_{st}

FIG. 7. *Pancreatic cancer data*: *The left panel shows a histogram of the total number of mapped reads*, N_{st}, *and the right panel shows a histogram of the empirical fractions*, n_{st}/N_{st}.

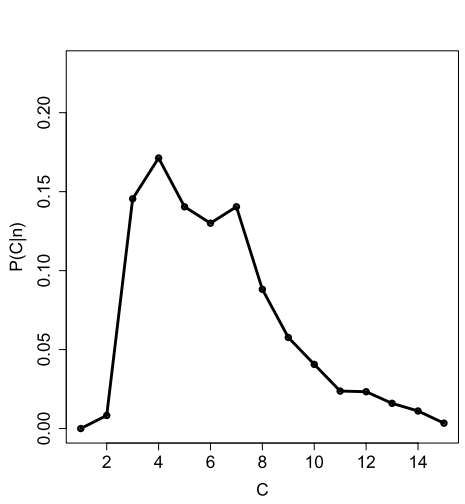
genome (version HG19) [Church et al. (2011)] using BWA [Li and Durbin (2009)] and called variants using GATK [McKenna et al. (2010)]. Post-mapping, the mean coverage of the samples was between 60 and 70 fold.

A total of nearly 115,000 SNVs and small indels were called within the exome coordinates. We restricted our attention to SNVs (i) that occur within genes that are annotated to be related to PDAC in the KEGG pathways database [Kanehisa et al. (2010)], (ii) that make a difference to the protein translated from the gene, and (iii) that exhibit significant coverage in all samples. This filtering left us with $S = 118$ SNVs.

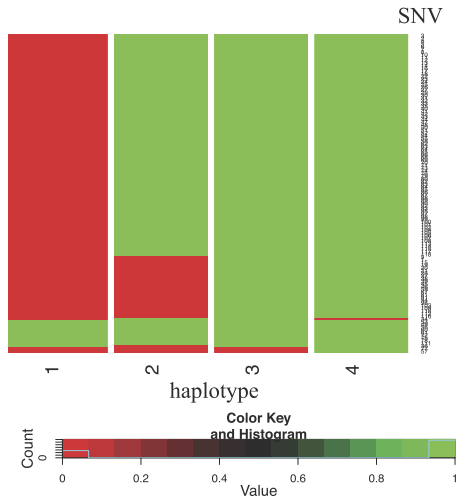
In summary, using the earlier introduced notation, the data record the read counts (N_{st}) and mutant allele read counts (n_{st}) of $S = 118$ SNVs from $T = 5$ tumor samples. Figure 7 shows a summary of the data. The large N_{st} values make the binomial likelihood very informative. For the prior specification, we let $r = 0.2$, $\alpha = 1$, $a = 1$, $a_0 = 1$, $a_{00} = 5$ and $b_{00} = 95$. We generated b_{st} from $\text{Be}(25, 975)$ for the minimal training set. We ran a MCMC posterior simulation over 35,000 iterations, discarding an initial transient of 10,000 iterations. Figure 8(a) shows the marginal posterior distribution for C. The posterior mode is $C^\star = 4$.

The posterior point estimate of $\mathbf{Z}$ conditional on $C^\star$ is shown in Figure 8(b) and the corresponding posterior point estimate of $\mathbf{w}$ in Figure 8(c). Here, green represents a variant sequence and red represents a reference sequence. We find that each sample has two or three two dominant haplotypes, that is, two green columns for each row in the heatmap. Haplotypes 2, 3, 4 are shared by different sets of the five samples. For example, sample 2 has a large-scaled abundance level for haplotypes 1, 2 and 3. Sample 4 is mainly dominated by haplotypes 1 and 3.

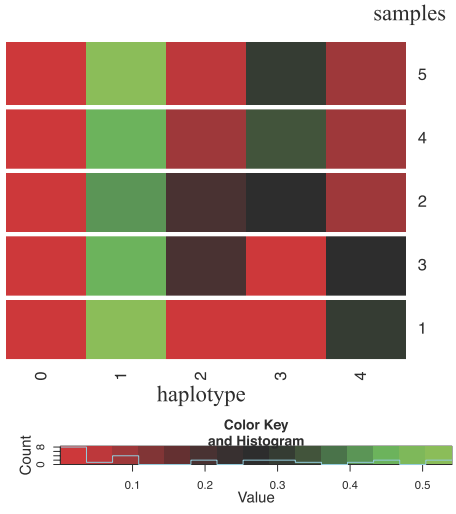
These results indicate that while tumors are unique, there are haplotypes that do recur across different patients. The results also clearly show that each tumor (in this data) is made of more than one haplotype: usually two or three dominant haplotypes and other minor types. To our knowledge, this is the first attempt to



(a) $p(C|\text{data})$



(b) $\mathbf{Z}^\star$ with $C^\star = 4$

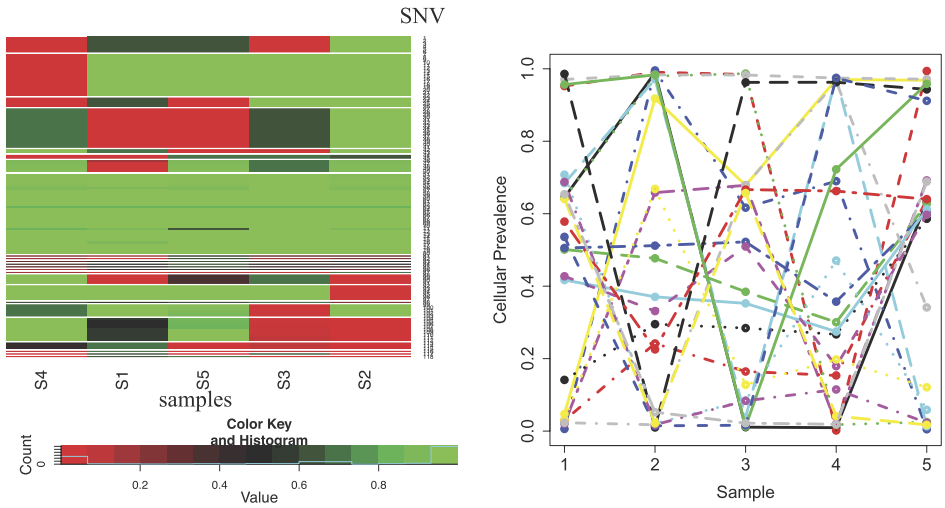


(c) $\mathbf{w}^\star$ with $C^\star = 4$ and $\mathbf{Z}^\star$

FIG. 8. *Pancreatic cancer data*: *The posterior distribution of* C *in* (a), *the heatmaps of* $\mathbf{Z}^\star$ *and* $\mathbf{w}^\star$ *with* $C^\star = 4$ *in* (b) *and* (c), *respectively. Note that for* $c = 0$, $p_0^\star z_{i0}^\star$ *is illustrated in the first column of panel* (c).

analyze the internal clonal composition of multiple PDAC tumor samples based on NGS data.

For comparison, we also evaluated tumor heterogeneity for the same pancreatic cancer data using PyClone [Roth et al. (2014)]. The results are shown in Figure 9. The posterior estimated clustering includes 24 SNV clusters, shown in panel (a). The estimated mean cellular prevalences of each cluster across the five samples are



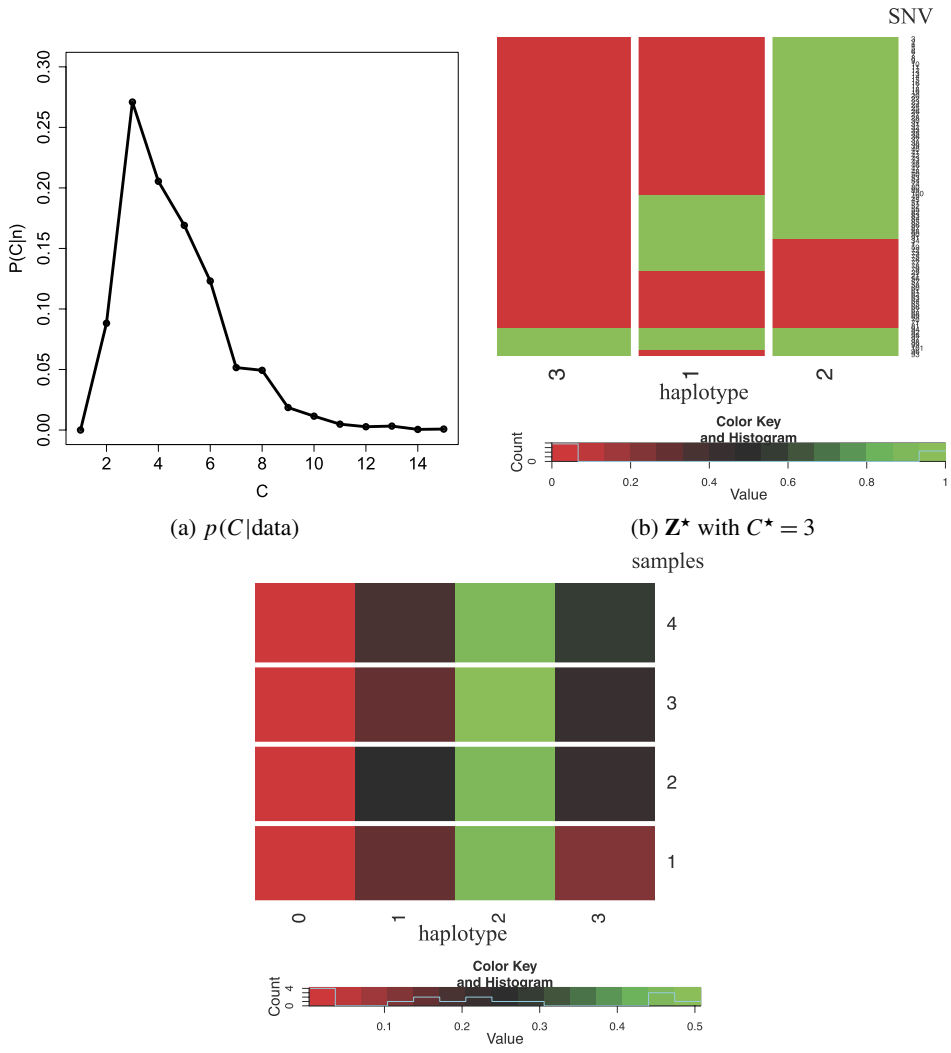
(a) Heatmap of cellular prevalence (b) Mean prevalence for each cluster

FIG. 9. *Pancreatic cancer data*: *Estimated cellular prevalence of SNV clusters over samples by PyClone*.

shown in (b). The estimated mean cellular prevalences vary substantially across samples.

5. Lung cancer data. We record whole-exome sequencing for four surgically dissected tumor samples taken from the same patient with lung cancer. The same bioinformatics preprocessing and analysis were carried out as in the previous pancreatic cancer example. We obtained SNVs and filtered them based on criteria similar to the previous example, leaving us in the end with $S = 101$ SNVs for the four intra-tumor samples.

We estimated the proposed Bayesian feature allocation model with the same hyperparameters as in the previous PDAC data analysis. Figure 10 summarizes the inference results. Panel (a) shows the marginal posterior distribution for C, identifying a posterior mode at $C^\star = 3$, that is, three distinct haplotypes. Panel (b) shows the posterior point estimate, $\mathbf{Z}^\star_C$, conditional on $C^\star = 4$. The figure shows which SNVs are included for each of the three haplotypes. Haplotype 3 contains the smallest number of mutations (green bars), implying that haplotype 3 might be the parental tumor cells. Haplotypes 1 and 2 are descendants of haplotype 3 with additional somatic mutations. Phylogenetically, a simple lineage can be hypothesized, with haplotype 3 as the parent of haplotype 1 and/or haplotype 2. Haplotype 2 possesses a large number of new somatic mutations, potentially representing a type of aggressive tumor cell. Panel (c) presents the posterior point estimate of $\mathbf{w}$, $\mathbf{w}^\star$ with $C^\star$ and $\mathbf{Z}^\star$. Examining haplotypes 1–3, we found that, interestingly, all four tumor samples share similar values of $\mathbf{w}^\star_t$, implying a lack of spatial hetero-



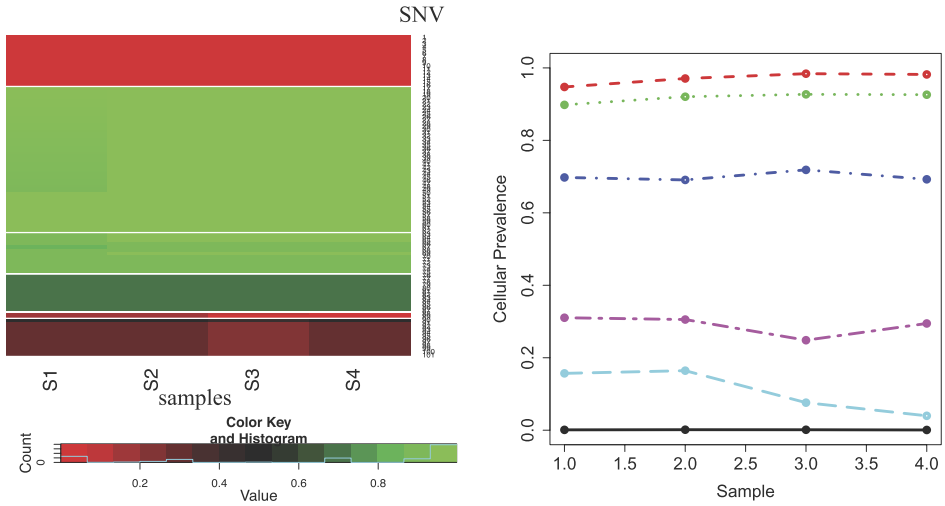
(a) $p(C|\text{data})$ (b) $\mathbf{Z}^\star$ with $C^\star = 3$

(c) $\mathbf{w}^\star$ with $C^\star = 3$ and $\mathbf{Z}^\star$

FIG. 10. *Lung cancer data*: *The posterior distribution of* C *in* (a), *the heatmaps of* $\mathbf{Z}^\star$ *and* $\mathbf{w}^\star$ *with* $C^\star = 3$ *in* (b) *and* (c), *respectively. Note that for* $c = 0$, $p_0^\star z_{t0}^\star$ *is illustrated in the first column of panel* (c).

geneity across the tumor samples. In other words, these samples all possess the inferred three tumor haplotypes in panel (b) with a similar composition.

Again, for comparison we also used PyClone [Roth et al. (2014)] with the lung cancer data. The results are shown in Figure 11. The estimated clustering identified six clusters of mutations. The mean prevalences within a mutation cluster are similar across samples.



(a) Heatmap of cellular prevalence (b) Mean prevalence for each cluster

FIG. 11. *Lung cancer data*: *Estimated cellular prevalence of SNV clusters over samples by PyClone.*

6. Conclusions. Tumors are heterogeneous tissues. The traditional way to identify this heterogeneity has been to sequence multiple samples from the tumor. Using such data to study the coexistence of genetically different subpopulations across tumors and within a tumor can shed light on cancer development. Identifying subpopulations within a tumor can lead to clinically important insights. For example, Landau et al. (2013) found that a chemotherapy affects subclonal heterogeneity in chronic lymphocytic leukemia. They also observed that the presence of a certain subpopulation may adversely affect clinical outcome.

We have proposed a model-based approach based on a feature allocation model. The feature allocation model allows us to impute inference about different components of tumor tissues based on NGS data. The identified components are not necessarily unique because there might be other possible solutions which can lead to the same hypothetical mutation frequencies. Instead of reporting a single solution, the proposed approach provides a full probabilistic description of all possible solutions as a coherent posterior probability model over C, $\mathbf{Z}$ and $\mathbf{w}$.

A number of extensions are possible for the present model. First, the number of SNVs examined in this paper was relatively limited (about 100), although the total number of SNVs that were found in the whole exome of a tissue is on the order of about 50,000. Other than computational complexity, there is no bar in principle on expanding the model to analyze the whole SNV complement of the exome. It could also be instructive to quantify the cellular diversity of the tumor based on findings from various regions of the exome.

Another important extension of the model is in the basic representation of subclones and haplotypes. The current model uses a binary matrix to record whether a

variant sequence for an SNV is present or absent in a haplotype. A variation of the model could instead record for each subclone whether an SNV is absent ($z_{sc} = 0$), heterozygous ($z_{sc} = 1$) or homozygous ($z_{sc} = 2$). That is, $\mathbf{Z}$ would become a trinary matrix. Other extensions of the model are to consider each SNV position to have four possible bases, A, C, G, T, to introduce dependence among mutations or to formally model the noise in variant calling. Each of these extensions is currently in development. For example, incorporating explicit error probabilities in variant calls is possible. Similar to our previous work [Ji et al. (2011)], we could replace the binomial likelihood (1) in the proposed model with a Bernoulli likelihood, for each read, where the probability associated with a read depends on quality scores of base calling and read mapping. We will consider this extension as future work.

Tumor genome sequencing projects have typically looked for specific genes to be mutated or not. The inherent assumption here, so far unproven, is that the overall effect of carcinogenesis could be explained by a handful of changes in a small number of genes. Our model takes the opposite approach and allows us to examine the whole genome (or exome) and, by considering VAF patterns, to construct reasonable models for the tissue. We believe this holistic approach to the analysis might provide more robust conclusions and biomarkers than the gene-by-gene approach.

SUPPLEMENTARY MATERIAL

Supplement to "A Bayesian feature allocation model for tumor heterogeneity" (DOI: 10.1214/15-AOAS817SUPP; .pdf). The supplementary material includes the second simulation study.

REFERENCES

BRODERICK, T., PITMAN, J. and JORDAN, M. I. (2013). Feature allocations, probability functions, and paintboxes. *Bayesian Anal.* **8** 801–836. MR3150470

BRODERICK, T., JORDAN, M. I. and PITMAN, J. (2013). Clusters and features from combinatorial stochastic processes. *Statist. Sci.* **28** 289–312.

CASELLA, G. and MORENO, E. (2006). Objective Bayesian variable selection. *J. Amer. Statist. Assoc.* **101** 157–167. MR2268035

CHURCH, D. M., SCHNEIDER, V. A., GRAVES, T., AUGER, K., CUNNINGHAM, F., BOUK, N., CHEN, H.-C., AGARWALA, R., MCLAREN, W. M., RITCHIE, G. R. S. et al. (2011). Modernizing reference genome assemblies. *PLoS Biol.* **9** e1001091.

ENGLE, L. J., SIMPSON, C. L. and LANDERS, J. E. (2006). Using high-throughput SNP technologies to study cancer. *Oncogene* **25** 1594–1601.

ERICHSEN, H. and CHANOCK, S. (2004). SNPs in cancer research and treatment. *British Journal of Cancer* **90** 747–751.

GERLINGER, M., ROWAN, A. J., HORSWELL, S., LARKIN, J., ENDESFELDER, D., GRONROOS, E., MARTINEZ, P., MATTHEWS, N., STEWART, A., TARPEY, P., VARELA, I., PHILLIMORE, B., BEGUM, S., MCDONALD, N. Q., BUTLER, A., JONES, D., RAINE, K., LATIMER, C., SANTOS, C. R., NOHADANI, M., EKLUND, A. C., SPENCER-DENE, B., CLARK, G., PICKERING, L., STAMP, G., GORE, M., SZALLASI, Z., DOWNWARD, J., FUTREAL, P. A. and SWANTON, C. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366** 883–892.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

GRIFFITHS, T. and GHAHRAMANI, Z. (2005). Infinite latent feature models and the Indian buffet process. Technical Report 2005-001, Gatsby Computational Neuroscience Unit, 2005.

JI, Y., XU, Y., ZHANG, Q., TSUI, K.-W., YUAN, Y., NORRIS, C. JR., LIANG, S. and LIANG, H. (2011). BM-map: Bayesian mapping of multireads for next-generation sequencing data. *Biometrics* **67** 1215–1224. MR2872372

KANEHISA, M., GOTO, S., FURUMICHI, M., TANABE, M. and HIRAKAWA, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*. **38** D355–D360.

LANDAU, D. A., CARTER, S. L., STOJANOV, P., MCKENNA, A., STEVENSON, K., LAWRENCE, M. S., SOUGNEZ, C., STEWART, C., SIVACHENKO, A., WANG, L., WAN, Y., ZHANG, W., SHUKLA, S. A., VARTANOV, A., FERNANDES, S. M., SAKSENA, G., CIBULSKIS, K., TESAR, B., GABRIEL, S., HACOHEN, N., MEYERSON, M., LANDER, E. S., NEUBERG, D., BROWN, J. R., GETZ, G. and WU, C. J. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152** 714–726.

LARSON, N. B. and FRIDLEY, B. L. (2013). PurBayes: Estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics* **29** 1888–1889.

LEE, J., MÜLLER, P., GULUKOTA, K. and JI, Y. (2015). Supplement to "A Bayesian feature allocation model for tumor heterogeneity." DOI:10.1214/15-AOAS817SUPP.

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

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. and 1000 GENOME PROJECT DATA PROCESSING SUBGROUP (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25** 2078–2079.

MARUSYK, A. and POLYAK, K. (2010). Tumor heterogeneity: Causes and consequences. *Biochim. Biophys. Acta*. **1085** 1.

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

NAVIN, N., KRASNITZ, A., RODGERS, L., COOK, K., METH, J., KENDALL, J., RIGGS, M., EBERLING, Y., TROGE, J., GRUBOR, V. et al. (2010). Inferring tumor progression from genomic heterogeneity. *Genome Res*. **20** 68–80.

NG, P. C. and KIRKNESS, E. F. (2010). Whole genome sequencing. In *Genetic Variation* 215–226. Springer, New York.

O'HAGAN, A. (1995). Fractional Bayes factors for model comparison. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **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. *Nature Methods* **11** 396–398.

RUSSNES, H. G., NAVIN, N., HICKS, J. and BORRESEN-DALE, A.-L. (2011). Insight into the heterogeneity of breast cancer through next-generation sequencing. *J. Clin. Invest*. **121** 3810–3818.

SU, X., ZHANG, L., ZHANG, J., MERIC-BERNSTAM, F. and WEINSTEIN, J. N. (2012). PurityEst: Estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics* **28** 2265–2266.

TEH, Y. W., GÖRÜR, D. and GHAHRAMANI, Z. (2007). Stick-breaking construction for the Indian buffet process. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, Vol. 11. The Society for Artificial Intelligence and Statistics, NJ.

WERSTO, R. P., LIBLIT, R. L., DEITCH, D. and KOSS, L. G. (1991). Variability in DNA measurements in multiple tumor samples of human colonic carcinoma. *Cancer* **67** 106–115.

WHEELER, D. A., SRINIVASAN, M., EGHOLM, M., SHEN, Y., CHEN, L., MCGUIRE, A., HE, W., CHEN, Y.-J., MAKHIJANI, V., ROTH, G. T. et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452** 872–876.

J. LEE
APPLIED MATHEMATICS & STATISTICS
BASKIN SCHOOL OF ENGINEERING
UNIVERSITY OF CALIFORNIA
1156 HIGH STREET
MAIL STOP SOE2
SANTA CRUZ, CALIFORNIA 95064
USA

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

K. GULUKOTA
NORTHSHORE UNIVERSITY HEALTHSYSTEM
1001 UNIVERSITY PLACE
EVANSTON, ILLINOIS 60201
USA

Y. JI
NORTHSHORE UNIVERSITY HEALTHSYSTEM
1001 UNIVERSITY PLACE
EVANSTON, ILLINOIS 60201
USA
AND
DEPARTMENT OF PUBLIC HEALTH SCIENCES
UNIVERSITY OF CHICAGO
5841 SOUTH MARYLAND AVE MC2000
CHICAGO, ILLINOIS 60637
USA
E-MAIL: jiyuan@uchciago.edu