BAYCLONE: BAYESIAN NONPARAMETRIC INFERENCE OF TUMOR SUBCLONES USING NGS DATA

SUBHAJIT SENGUPTA[1], JIN WANG[2], JUHEE LEE[3], PETER MÜLLER[4], KAMALAKAR GULUKOTA[5], ARUNAVA BANERJEE[6], YUAN JI[1,7,*]

[1] *Center for Biomedical Research Informatics, NorthShore University HealthSystem*
[2] *Department of Statistics, University of Illinois at Urbana-Champaign*
[3] *Department of Applied Mathematics and Statistics, University of California Santa Cruz*
[4] *Department of Mathematics, University of Texas Austin*
[5] *Center for Molecular Medicine, NorthShore University HealthSystem*
[6] *Department of Computer & Information Science & Engineering, University Of Florida*
[7] *Department of Health Studies, The University Of Chicago*

In this paper, we present a novel feature allocation model to describe tumor heterogeneity (TH) using next-generation sequencing (NGS) data. Taking a Bayesian approach, we extend the Indian buffet process (IBP) to define a class of nonparametric models, the categorical IBP (cIBP). A cIBP takes categorical values to denote homozygous or heterozygous genotypes at each SNV. We define a subclone as a vector of these categorical values, each corresponding to an SNV. Instead of partitioning somatic mutations into non-overlapping clusters with similar cellular prevalences, we took a different approach using feature allocation. Importantly, we do not assume somatic mutations with similar cellular prevalence must be from the same subclone and allow overlapping mutations shared across subclones. We argue that this is closer to the underlying theory of phylogenetic clonal expansion, as somatic mutations occurred in parent subclones should be shared across the parent and child subclones. Bayesian inference yields posterior probabilities of the number, genotypes, and proportions of subclones in a tumor sample, thereby providing point estimates as well as variabilities of the estimates for each subclone. We report results on both simulated and real data. BayClone is available at `http://health.bsd.uchicago.edu/yji/soft.html`.

Keywords: Categorical Indian buffet process; Heterozygosity; Latent feature model; NGS data; Random categorical matrices; Subclones; Tumor heterogeneity.

1. Introduction

1.1. *Background*

Tumorgenesis is a complex process.[1,2] A wide variety of genetic features that promotes tumors are involved in this process, including the acquisition of somatic mutations that allow tumor cells to gain advantages over time compared to normal cells. As such, a tumor is oftentimes heterogeneous consisting of multiple subclones with unique genomes, a phenomenon called tumor heterogeneity (TH). Multiple recent reviews[3–8] support the existence of subclones within tumors. Specifically, cancer cells undergo Darwinian-like clonal somatic evolution and tumor formation is dependent on acquisition of oncogenic mutations. In fact it has been found that individual tumors have a unique clonal architecture that is spatially and temporally evolving, which poses challenges as well as opportunities on individualized cancer treatment. We consider the differences in subclones arising from single nucleotide variations (SNVs), although

*Address for Correspondence: Research Institute, NorthShore University HealthSystem, 1001 University Place, Evanston, IL 60201, USA. Email: koaeraser@gmail.com

there can be other differences such as copy number variations. An SNV represents modification to a single DNA sequence. A scaffold of SNVs along the same haploid genome constitutes a *haplotype*. A pair of haplotypes gives rise to a subclonal genome.

Next-generation sequencing (NGS) experiments use massively parallel sequenced short reads to study long genomes. The short reads are mapped to the reference genome based on sequence similarities. Mapped reads are used to produce estimates of SNVs, small indels and copy number (CN) variations along the genome. In this paper we use the whole-genome sequencing (WGS) or whole-exome sequencing (WES) data to model the variant allele fraction (VAF) at an SNV, defined as the fraction of short reads that bear a variant sequence (compared to the reference genome). Innovatively, we infer subclones using scaffolds of SNVs, or haplotypes.

1.2. *Main idea*

Most multicellular organisms have two sets of chromosomes – they are called diploids. Diploid organisms have one copy of each gene (and therefore one allele) on each chromosome. At each locus, two alleles can be *homozygous* if they share the same genotypes, or *heterozygous* if they do not. In a recent paper[9] the authors use an Indian buffet process (IBP)[10] that assumes that SNVs are homozygous, where both alleles are either mutated or wild-type. However, biologically there are three possible allelic genotypes at an SNV: homozygous wild-type (no mutation on both alleles), heterozygous mutant (mutation on only one allele), or homozygous mutant (mutation on both alleles). Therefore, the IBP model is not sufficient to fully describe the subclonal genomes.

Our main idea is to extend IBP to categorical IBP that allows three values, 0, 0.5, and 1, to describe the corresponding genotypes at each SNV. Such an extension is mathematically nontrivial as we show later. More importantly, it allows for a principled and powerful statistical inference on TH. Different from existing methods based on Dirichlet processes,[11,12] IBP and cIBP allow one SNV to appear in multiple subclones. We argue that this is more realistic and agrees with the fundamental evolutionary theory of clonal expansion. In particular, somatic mutations occurred in early tumor development should be shared by child subclones.

To start, note that each SNV can be associated with a non-negative number of subpopulations. Consider a finite number of S SNV loci and assume that an unknown number of C subclones are present. We introduce an $S \times C$ ternary matrix, $\boldsymbol{Z} = [z_{sc}]$ where each z_{sc} denotes the allelic variation at SNV site s for subclone c, $s = 1, 2, \cdots, S$; $c = 1, 2, \cdots, C$. Specifically, we let $z_{sc} \in \{0, 0.5, 1\}$ be a ternary random variable to denote three possible genotypes at the locus, homozygous wild-type ($z_{sc} = 0$), heterozygous variant ($z_{sc} = 0.5$), and homozygous variant ($z_{sc} = 1$); see Figure 1. Each sample is potentially an admixture of the subclones (columns of $\boldsymbol{Z}$), mixed in different proportions. Given $\boldsymbol{Z}$, we can denote the proportions of the C subclones by $\boldsymbol{w}_t = (w_{t0}, w_{t1}, \cdots, w_{tC})$ for sample t, where $0 < w_{tc} < 1$ for all c and $\sum_{c=0}^{C} w_{tc} = 1$. Therefore, the contribution of a subclone to the VAF at an SNV is $0 \times w_{tc}$, $0.5 \times w_{tc}$ or $1 \times w_{tc}$, if the subclone is homozygous wild-type, heterozygous or homozygous mutant at the SNV, respectively. We develop a latent feature model (Section 2.3) for the entire matrix $\boldsymbol{Z}$ to uncover the unknown subclones that constitute the tumor cells and given the data, we aim to

infer two quantities, $\boldsymbol{Z}$ and $\boldsymbol{w}$, by a Bayesian inference scheme.

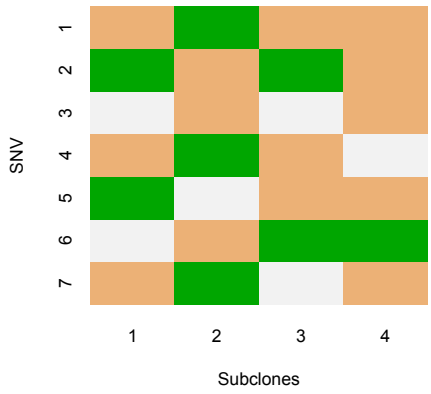


Fig. 1. Illustration of cIBP matrix $\boldsymbol{Z}$ for subclones in a tumor sample. Colored cells in green=1, brown=0.5, and white=0 represent homozygous variants, heterozygous variants, and homozygous wild-type, respectively.

As shown in Figure 1, a subclone is defined by a vector of categorical values in $\{0, 0.5, 1\}$ representing the genotypes at specific SNV location. For example, in Figure 1 there are seven different SNV locations and four subclones. SNV 5 takes values 1 in subclone one, 0 in subclone two, and 0.5 in subclones three and four. Therefore, the same mutation is shared by two subclones (three and four).

1.3. *Existing Methods*

Recent rapid development has generated useful tools for subclonal inference, notably represented by SciClone,[13] TrAp,[14] PhyloSub,[15] and Clomial,[16] among others. TrAp nicely described the issue of solution identifiability stating the need to have sufficient sample size for unique mathematical solutions. SciClone and Clomial assume a binary matrix, focusing SNVs at copy neutral regions with heterozygous mutations. PhyloSub carefully considered possible genotypes at SNVs accounting for potential copy number changes. BayClone differs from these methods in that it novelly proposes a categorical IBP model as a nonparametric approach, accounting for the three potential genotypes at each SNV. BayClone does not take the indirect route of estimating nested clusters in a tree structure, as in PhyloSub. Instead, it directly outputs subclones with overlapping SNVs. As a model choice, BayClone does not assume a phylogenetic tree structure for the inferred subclones since in any given tumor sample not all the subclones on the phylogenetic tree may be present. Instead, existing subclones may only represent nodes on a subset of branches of the phylogenetic tree. With this notion, BayClone does not assume a phylogenetic tree but rather use data to infer subclones with overlapping SNVs.

The remainder of the paper is organized as follows. In section 2, we elaborate on the proposed probability model. Section 3 describes model selection and posterior inference. In the following section, we report experimental results, one with simulated data and another by real-life data from an NGS experiment. In the final section we conclude with discussion and future work.

2. Probability Model

2.1. *Latent feature model with IBP*

In latent feature model, each data point is generated by a vector of latent feature values. In our case, each subclone (one column of $\boldsymbol{Z}$) is a latent feature vector and a data point is the observed VAF. The IBP model is used to define a prior on the space of binary matrices that indicate the presence of a particular feature for an object, with the number of columns in the matrix (corresponding to features) being potentially unbounded. The detailed construction of IBP can be found in Ref. [10]. We consider a constructive definition of IBP as follows. For each component z_{sc} in the binary matrix $\boldsymbol{Z}$, assume

$$\begin{aligned} z_{sc} \mid \pi_c &\sim Bern(\pi_c), \\ \pi_c \mid \alpha &\sim beta(\alpha/C, 1), \quad c = 1, \ldots, C, \end{aligned} \tag{1}$$

where $Bern(\pi_c)$ is the Bernoulli distribution and $\pi_c \in (0, 1)$ is the probability $Pr(z_{sc} = 1)$ *a priori*. Also, the marginal $p(\boldsymbol{Z}) = \prod_{c=1}^{C} p(\boldsymbol{Z}_c) = \int p(z_{sc} \mid \pi_c) p(\pi_c) d\pi_c$ factors assuming conditional independence, where $\boldsymbol{Z}_c$ is the c-th column vector. When $C \to \infty$, the marginal distribution of $\boldsymbol{Z}$ (as an equivalence class) exists and is called IBP. We extend the IBP model to a categorical setting, where each entry of the matrix is not necessarily 0 or 1, but a set of integers in $\{0, 1, \cdots, Q\}$ where Q is fixed *a priori*. We call the extended model categorical IBP (cIBP) and use it as a prior in exploring subclones of tumor samples. In upcoming discussion, SNVs correspond to objects (rows) and subclones correspond to feature (columns) in the $\boldsymbol{Z}$ matrix.

2.2. *Development of cIBP*

We discuss the development of the cIBP for a general case with an arbitrary Q. A straightforward extension of IBP in (1) would be to replace the underlying beta distribution of π_c with a Dirichlet distribution, and replace the Bernoulli distribution of z_{sc} with a multinomial distribution. However, as $C \to \infty$, Ref. [17] showed that the limiting distribution is degenerate. Instead, utilizing a Beta-Dirichlet distribution defined in Ref. [18] we propose a construction given C and Q: let $\{1, \ldots, Q\}$ be the possible values z_{sc} takes. Then we assume

$$\boldsymbol{\pi}_c \sim \text{Beta-Dirichlet}\,(\alpha/C, 1, \underbrace{\beta, \cdots, \beta}_{Q \text{ of them}}); \qquad z_{sc} \mid \boldsymbol{\pi}_c \sim \text{Multi}(1, \boldsymbol{\pi}_c). \tag{2}$$

Integrating out $\boldsymbol{\pi}_c$ in (2), the probability of a $(Q+1)$-nary matrix, $\boldsymbol{Z}$ is

$$\mathrm{p}(\boldsymbol{Z}) = \left(\frac{1}{\prod_{s=1}^{S} (s + \alpha/C)} \right)^{C} \prod_{c=1}^{C_+} \left(\frac{\alpha}{C} \cdot \frac{1}{Q} \right) \frac{(S - m_{c\cdot})!}{S!} \times \prod_{j=1}^{m_{c\cdot}-1} \left[\frac{(j + \alpha/C)}{(j + Q\beta)} \right] \frac{1}{\beta} \prod_{q=1}^{Q} \frac{\Gamma(\beta + m_{cq})}{\Gamma(\beta)},$$

where m_{cq} denotes the number of rows possessing value $q \in \{1, \ldots, Q\}$ in column c, i.e., $m_{cq} = \sum_{s=1}^{S} \mathbb{I}(z_{sc} = q)$ and $m_{c\cdot} = \sum_{q=1}^{Q} m_{cq}$. This gives birth to a random matrix with C columns, each entry taking a discrete value in a set of $(Q+1)$ values. It can be shown that the limiting distribution of $\boldsymbol{Z}$ (as an equivalent class) exists and is called the cIBP.[17]

2.3. *Sampling model*

Suppose there are T tumor samples in the data in which S SNVs are measured for each sample. Let N_{st} be the total number of reads mapped to SNV s in sample t, $s = 1, 2, \cdots, S$ and $t = 1, 2, \cdots, T$. Among N_{st} reads, assume n_{st} possess a variant sequence at the locus. We assume a binomial sampling model

$$n_{st} \overset{indep.}{\sim} \text{Binomial}\,(N_{st}, p_{st})\,, \tag{3}$$

where p_{st} is the expected proportion of variant reads.

We assume that the matrix $\boldsymbol{Z}$ follows a finite version of cIBP in (2), $\boldsymbol{Z} \sim \text{cIBP}_C(Q = 2, \alpha, [\beta_1, \beta_2])$. Recall that $\boldsymbol{w}_t = (w_{t0}, w_{t1}, \cdots, w_{tC})$ denotes the vector of subclonal weights. We assume $\boldsymbol{w_t}$ follows a Dirichlet prior given by,

$$\boldsymbol{w}_t \overset{indep.}{\sim} \text{Dirichlet}(a_0, a_1, \cdots, a_C).$$

As we have mentioned earlier that, each sample t potentially consists of several subclones with different proportions. Thus the variant reads must come from those subclones possessing variant alleles. In other words, parameters p_{st} can be modeled as a linear combination of variant alleles $z_{sc} \in \{0, 0.5, 1\}$ weighted by the proportions of subclones bearing the alleles. Remember that, $z_{sc} = 0$, 0.5 and 1 means that there is no mutation, heterozygous mutation and homozygous mutation at SNV position s for subclone c, respectively. Apparently, when a subclone bears no variant alleles, i.e., $z_{sc} = 0$, the contribution from that subclone to p_{st} should be zero. We assume the expected p_{st} is a result of mixing subclones with different proportions. Mathematically, given $\boldsymbol{Z}$ and $\boldsymbol{w}$ we assume

$$p_{st} = \sum_{c=1}^{C} w_{tc}\, z_{sc} \quad + \epsilon_{t0}. \tag{4}$$

Equation (4) is a key model assumption. It allows us to back out the unknown subclones from a decomposition of the expected VAF p_{st} as a weighted sum of latent genotype calls z_{sc} with weights w_{tc} being the proportions of subclones. Importantly, we assume these weights to be the same across all SNV's, $s = 1, \ldots, S$. In other words, the expected VAF is contributed by those subclones with variant genotypes, weighted by the subclone prevalences. Subclones without variant genotype on SNV s do not contribute to the VAF for s since all the short reads generated from those subclones are normal reads.

In (4) ϵ_{t0} is an error term defined as $\epsilon_{t0} = p_0 w_{t0}$, where $p_0 \sim \text{Beta}(\alpha_0, \beta_0)$. Importantly ϵ_{t0} is devised to capture experimental and data processing noise. Specifically, p_0 is the relative frequency of variant reads produced as error from upstream data processing and takes a small value close to zero; w_{t0} absorbs the noise left unaccounted for by $\{w_{t1}, \ldots, w_{tC}\}$. Ignoring the error term ϵ_{t0}, model (4) can also be considered as a non-negative matrix factorization (NMF) if it is written as a matrix format, in which p_{st} could be replaced by observed VAFs. Our proposed feature allocation models differ from (NMF) in that we take a probabilistic approach effectively accounting for the noise in the data and providing variabilities measures for the model parameters using probability inference. We discuss the inference below.

3. Model Selection and Posterior Inference

3.1. *MCMC simulation*

In order to infer the sampling parameters from the posterior distribution, we use Markov chain Monte Carlo (MCMC) simulations. The Gibbs sampling method is used to update z_{sc}, whereas the Metropolis-Hastings (MH) sampling is used to get the samples of w_{tc} and p_0. We omit detail except the one for sampling z_{sc}. Due to exchangeability, we let SNV s be the last customer. Let $\boldsymbol{z}_{-s,c}$ be the set of assignment of all other SNVs but SNV s for subclone c, m^-_{cq} the number of SNVs with level q, not including SNV s and $m^-_{c\cdot} = \sum_{q=1}^{Q} m^-_{cq}$. We obtain,

$$\mathrm{p}(z_{sc} = q \mid \boldsymbol{z}_{-s,c}, rest) \propto \left(\frac{m^-_{c\cdot}}{s}\right) \times \left(\frac{\beta_q + m^-_{cq}}{\beta^\star + m^-_{c\cdot}}\right) \prod_{t=1}^{T} \binom{N_{st}}{n_{st}} (p'_{st})^{n_{st}} (1 - p'_{st})^{(N_{st} - n_{st})}$$

for any c such that $m^-_{c\cdot} > 0$, where *rest* includes the data and current MCMC values for all the other parameters. Also, p'_{st} is value of p_{st} by plugging the current MCMC values and setting $z_{sc} = q$.

3.2. *Choice of C*

The number of subclones C in cIBP is unknown and must be estimated. We discuss a model selection to select the correct value for C. We use predictive densities as a selection criterion. Let $\boldsymbol{n}_{-st}$ denote the data removing n_{st}. Also denote the set of parameters for a given C by $\boldsymbol{\eta}^C$. The conditional predictive ordinate (CPO)[19] of n_{st} given $\boldsymbol{n}_{-st}$ is given by

$$CPO_{st} = p(n_{st}|\boldsymbol{n}_{-st}) = \int p(n_{st}|\boldsymbol{\eta}^C, \boldsymbol{n}_{-st}) p(\boldsymbol{\eta}^C|\boldsymbol{n}_{-st})\, d\boldsymbol{\eta}^C. \tag{5}$$

The Monte-Carlo estimate of (5) is the harmonic mean of the likelihood values[20] $p(n_{st} \mid \boldsymbol{\eta}^C_l)$,

$$\hat{p}(n_{st}|\boldsymbol{n}_{-st}) \approx \frac{1}{L^{-1} \sum_{l=1}^{L} p(n_{st} \mid \boldsymbol{\eta}^C_l)^{-1}} \tag{6}$$

where $\boldsymbol{\eta}^C_l$'s are MCMC draw's and L is the number of iterations. We take each data point out from $\boldsymbol{n}$ and compute average *log-pseudo-marginal likelihood* (LPML) over this set as $L^C = \sum_{n_{st} \in \boldsymbol{n}} \log[\hat{p}(n_{st}|\boldsymbol{n}_{-st})]$. For different values of C, we compare the values of L^C and choose that $\hat{C}$ which maximizes L^C.

3.3. *Estimate of Z*

The MCMC simulations generate posterior samples of the categorical matrix $\boldsymbol{Z}$ and other parameters. Directly taking sample average is not desirable since it will result in an estimated matrix with entries taking values outside the set $\{0, 0.5, 1\}$. Instead, we define a posterior point estimate of $\boldsymbol{Z}$ similar to that in Ref. [9], i.e.,

$$\hat{\mathbf{Z}} = \arg\min_{\mathbf{Z}'} \frac{1}{L} \sum_{l=1}^{L} d(\mathbf{Z}^{(l)}, \mathbf{Z}') \tag{7}$$

where $\mathbf{Z}^{(l)}, l = 1, \ldots, L$ are MCMC samples. The term $d(\mathbf{Z}^{(l)}, \mathbf{Z}')$ is a distance with the following definition. Note that the MCMC samples $\mathbf{Z}^{(l)}$ may have different labels for Z across iterations.

Therefore we introduce a permutation for comparing any two matrices. For two matrices Z and Z', let $D_{cc'}(\mathbf{Z}, \mathbf{Z}') = \sum_{s=1}^{S} |z_{sc} - z'_{sc'}|$ for two columns c and c'. We define a distance $d(\mathbf{Z}, \mathbf{Z}') = \min_{\zeta} \sum_{c=1}^{C} D_{c,\zeta_c}(\boldsymbol{Z}, \boldsymbol{Z}')$ where $\zeta_c, c = 1, \ldots, C$ is a permutation of $\{1, \ldots, C\}$. Having the permutation ζ_c resolves the potential label-switching issue in the MCMC samples.

4. Results

4.1. *Simulated Data*

We demonstrate the performance of BayClone with two sets of simulated data. First, we take a set of $S = 100$ SNV locations and consider $T = 30$ samples. The true number of latent subclones is $C = 4$ in this experiment. The true $\boldsymbol{Z}$ values are given in the left most panel of Figure 2. We generate the true proportion matrix $\boldsymbol{w}$ by setting $w_{t0} = 0.05$ to account for the

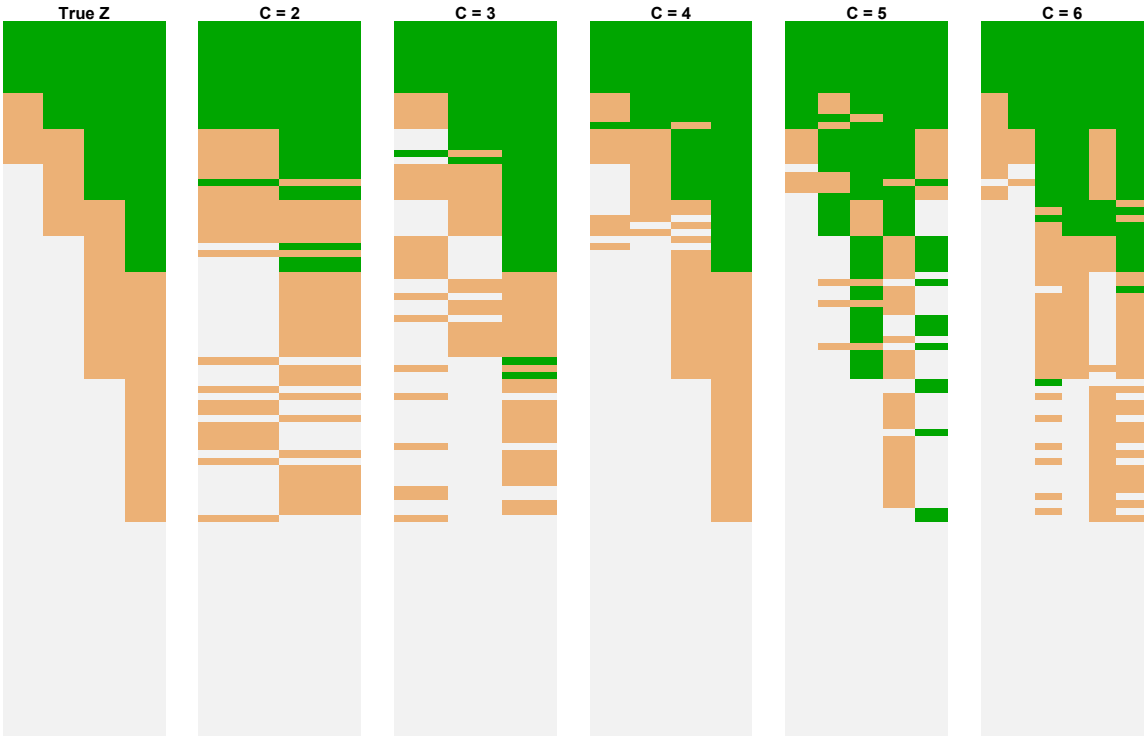
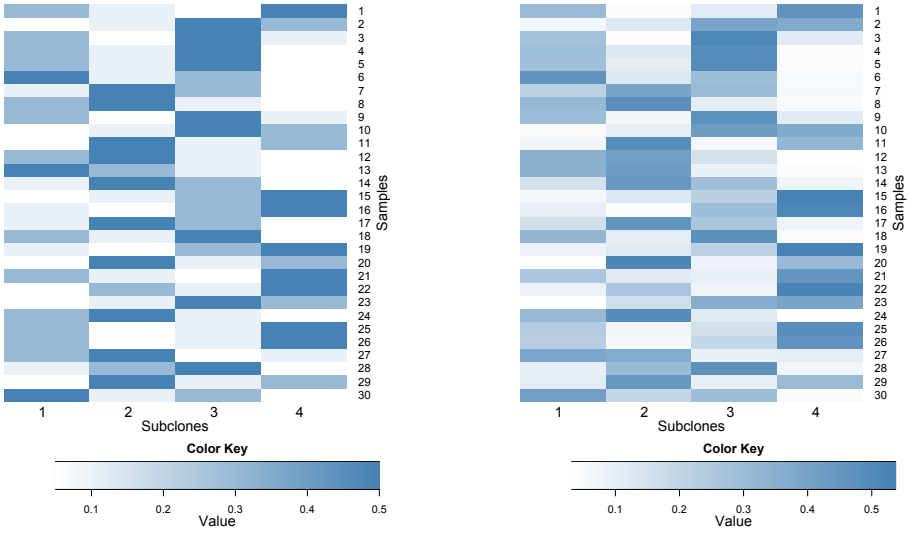


Fig. 2. True $\boldsymbol{Z}$ and estimate $\hat{\boldsymbol{Z}}$ in (7) with green standing for homozygous mutation i.e. $z_{sc} = 1$, brown for heterozygous mutation i.e. $z_{sc} = 0.5$ and white for homozygous wild type i.e. $z_{sc} = 0$. The model with $C = 4$ fits the data the best.

background noise in sample t, and the rest w_{tc}'s from the permutations of $(0.5, 0.3, 0.1, 0.05)$ (where $c = 1, 2, 3, 4$). We take the true p_0 as 0.01 and fix $N_{st} = 50$ for all $s = 1, 2, \cdots, 100$ and $t = 1, 2, \cdots, 30$. Finally we generate n_{st} from Binomial(N_{st}, p_{st}). Hyperparameters are set up as follows: for $\boldsymbol{w}_t$: $a_0 = a_1 = a_2 = \cdots = a_C = 1$, for $\boldsymbol{\pi}_c$: $\alpha = 1$, $\beta_1 = \beta_2 = 2$, and for p_0: $\alpha_0 = 1$, $\beta_0 = 100$. Given C, we randomly initialize the binary matrix $\boldsymbol{Z}$ and draw the initial p_0 from the specified prior. The initial $\boldsymbol{w}_t$ are generated by drawing gamma random variables from the prior $\boldsymbol{\theta}_t \sim \text{Gamma}(a_0, a_1, \ldots, a_C)$, and then normalizing them. That is, $w_{tc} = \theta_{tc}/(\sum_{k=0}^{C} \theta_{tk})$.



(a) True $\boldsymbol{w}$ (b) Estimate $\hat{\boldsymbol{w}}$

Fig. 3. True $\boldsymbol{w}$ and estimated proportions $\hat{\boldsymbol{w}}$ for $\hat{C} = 4$ with simulated data.

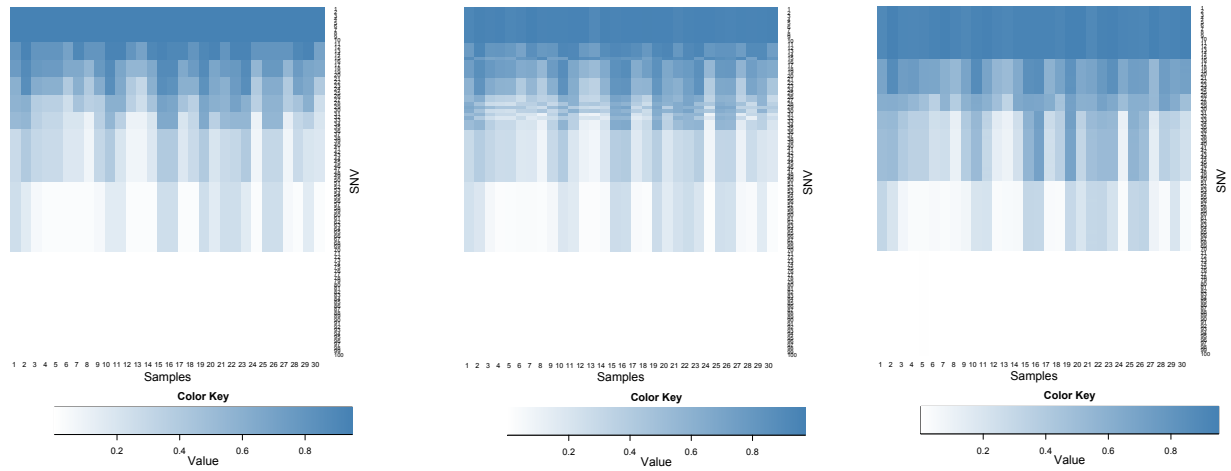
For MCMC simulations, we run 4,000 iterations, discard the initial 2,000 as burn in, and take one sample every 5th iteration. The Markov chain converges quickly. Table 1 presents the average LPML for various values of C. As we can see, L^C is maximized at $\hat{C} = 4$, which is

Table 1. LPML L^C for C values. The simulation truth is 4.

C	2	3	4	5	6
L^C	-9144.4	-6664.1	**-4992.869**	-5218.707	-5034.129

the true C. We find the estimate $\hat{\boldsymbol{Z}}$ in (7) based on the posterior samples drawn from MCMC simulations. In Figure 2, we compare the truth with estimates $\hat{\boldsymbol{Z}}$ for different values of C. Also we plot the true $\boldsymbol{w}$ and estimate $\hat{\boldsymbol{w}}$ across all the samples using $\hat{C} = 4$ in Figure 3. They have almost identical values. As model checking we computed the difference between the true p_{st} and the posterior mean $\hat{p}_{st}$, for different model C. When $\hat{C} = 4$, the difference of $\hat{p}_{st}$ and true p_{st} is the smallest (results not shown). Also the posterior mean of p_0 is 0.0107 for the correct value of C, which is very close to the simulation truth $p_0 = 0.01$. All the other parameters in the model were closely estimated under the Bayesian model as well.

Lastly, we compare the simulation results with PyClone,[12] which uses Dirichlet process to partition SNVs into mutation clusters. In Figure 4, we plot the true $\boldsymbol{p} = [p_{st}]$, estimate $\hat{\boldsymbol{p}}$ by our model and cellular prevalences inferred by PyClone, which is equivalent to $\hat{\boldsymbol{p}}$ in our models. PyClone estimates six SNV clusters. Also, the L1-norm, $\sum_{s,t} |p_{st} - \hat{p}_{st}|$ equals 35.24 for our method, compared to 132.04 for PyClone. The fitting is worse than our model when $C = 4$. We note that the comparison to PyClone is not to choose a superior method as PyClone does not directly provide estimates on subclones and their cellularities. Instead, it only aims to infer subclonal frequencies of SNVs.



(a) True $\boldsymbol{p}$ (b) Estimate $\hat{\boldsymbol{p}}$ (c) Cellular prevalence by PyClone

Fig. 4. True and estimated (by our model) expected VAF and cellular prevalence inferred by PyClone

As a second simulation, we use the same setting but reduce the numbers of SNVs and samples to $S = 70$ and $T = 7$. We find that again $\hat{C} = 4$ according to the LPML criterion and in Figure 5 we find that $\hat{\boldsymbol{Z}}$ is closest to the true Z when $\hat{C} = 4$. Estimate of $\boldsymbol{w}$ when $\hat{C} = 4$ is close to the truth (results not shown) and so are the other model parameters. In summary, the model performs well with only $T = 7$ samples.

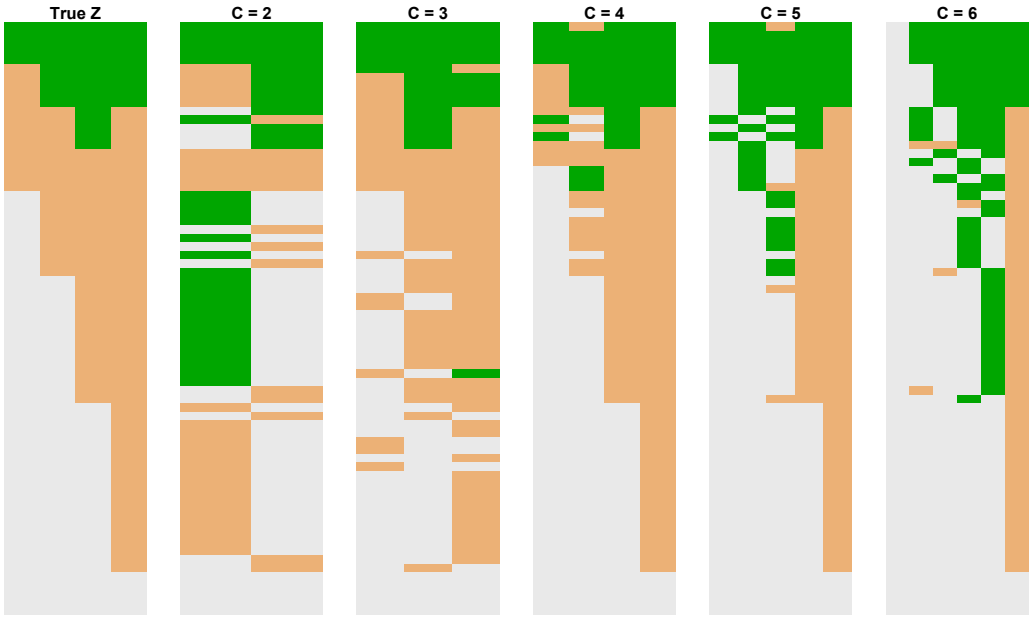


Fig. 5. True $\boldsymbol{Z}$ and estimate $\hat{\boldsymbol{Z}}$ in (7) with green standing for homozygous mutation i.e. $z_{sc} = 1$, brown for heterozygous mutation i.e. $z_{sc} = 0.5$ and white for homozygous wild type i.e. $z_{sc} = 0$. The model with $C = 4$ fits the data with $S = 70$ SNVs and $T = 7$ samples the best.

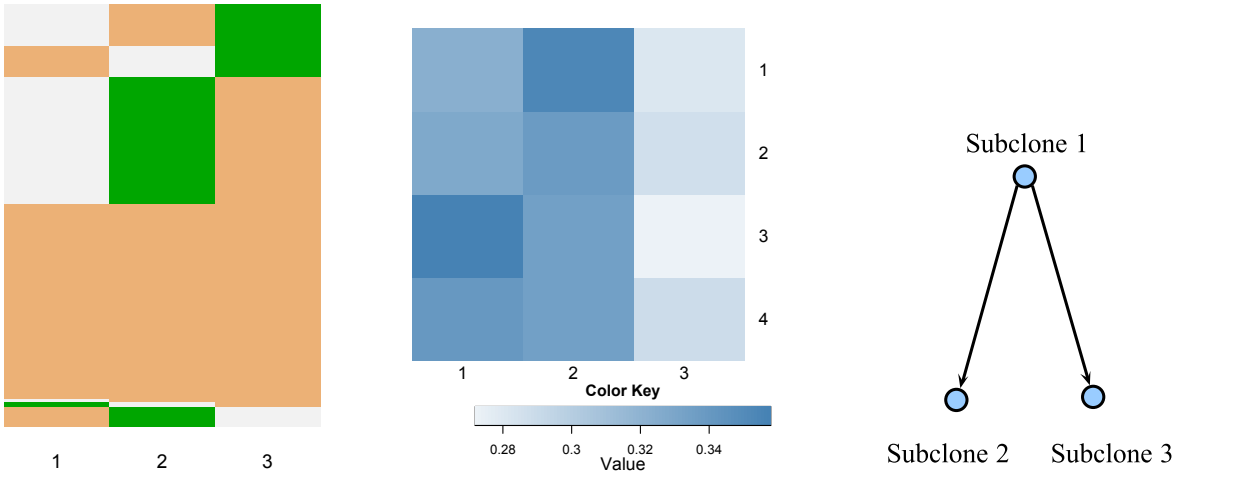
4.2. *Intra-Tumor Lung Cancer Samples*

We record whole-exome sequencing for four surgically dissected tumor samples taken from a single patient diagnosed with lung adenocarcinoma. A portion of the resected tumor is flash frozen and another portion is formalin fixed and paraffin embedded (FFPE). Two different specimens are taken from the frozen portion of the resected tumor and another two from the FFPE portion. Genomic DNA is extracted from all four specimens and an exome capture is done using Agilent SureSelect v5+UTR probe kit. The exome library is then sequenced in paired-end fashion on an Illumina HiSeq 2000 platform. Only two specimens are sequenced on each to ensure a high depth of coverage. We map the reads to the human genome (version HG19)[21] using BWA[22] and called variants using GATK.[23] Post-mapping, the mean coverage of the samples is around 100 fold.

We restrict our attention to the SNVs that (i) exhibit significant coverage in all our samples (total number of mapped reads N_{st} are ranged in $[100, 240]$) and (ii) have reasonable chance of mutation (the empirical fractions n_{st}/N_{st} in $[0.25, 0.75]$). This filtering left us with $12,387$ SNV's. We then randomly select $S = 150$ for computational purposes. In summary, using the above notations, the data record the read counts (N_{st}) and mutant allele read counts (n_{st}) of $S = 150$ SNVs from $T = 4$ tumor samples.

The large values for N_{st} make the binomial likelihood very informative. For the prior specification, we adopt the same hyperparameters in the simulation study. We ran MCMC for $6,000$ iterations, discarding the first $3,000$ iterations as initial burn-in and thinning by 3. We consider $C = 2, 3, 4, 5, 6$ and use LPML to select the best C. LPML values for $C = 2, 3, 4, 5$ and 6 are $-1991.87, \mathbf{-1991.82}, -1992.64, -1993.57$ and -1994.67, respectively. So the LPML is maximized at $\hat{C} = 3$ implying that three distinct subclones are present. Conditioning on $\hat{C} = 3$, the estimate $\hat{\boldsymbol{Z}}$ is shown in Figure 6(a). The proportions of the three subclones in each of the four samples are plotted in Figure 6(b). A phylogenetic tree is hypothesized in Figure 6(c). In particular, subclone 1 appears to be the parent giving birth to two branching child subclones 2 and 3. Comparing columns in Figure 6(a), we hypothesize that subclones 2 and 3 arise by acquiring additional somatic mutations in the top portion of the SNV regions where subclone 1 shows "white" color, i.e., homozygous wild type. The three subclones share the same genotype in the middle and lower half of the SNVs (the large chunk of "brown" bars in Figure 6(a)), suggesting that these could be either somatic mutations acquired in the parent subclone 1, or germline mutations. All four tumor samples have similar proportions of the subclones, showing lack of geographical heterogeneity although each sample is mosaic. This is expected since the four tumor samples were dissected from regions that were close by on the original lung tumor.

Clinically, our analysis provides valuable information for treatment considerations. Since each tumor sample is mosaic consisting of three subclones, detailed mutational annotation could be conducted to seek potential biomarker mutations for targeted therapy. Also, combinational drugs could be considered if possible to specifically target each subclone. Since the four tumor samples possess similar proportions of subclones, the tumor appears to be homogeneous spatially. The results from our subclonal analysis could be used as a future reference should the disease progress or relapse. For example, future subclonal analysis could be



(a) Estimate $\hat{\boldsymbol{Z}}$ (b) Subclone proportions $\hat{\boldsymbol{w}}$ (c) An estimated lineage for subclones

Fig. 6. Subclone structures, proportions and a possible lineage for the lung cancer data.

compared to the existing one to understand the temporal genetic changes.

5. Discussion and future work

One of the major motivations to detect the heterogeneity in tumors is personalized medicines.[24] Measure of heterogeneity can be useful as a prognosis marker.[25] Using NGS data to study the co-existence of genetically different subpopulations across tumors and within a tumor can shed light on cancer development. The main feature of BayClone is the model-based and principled inference on subclonal genomes for a set of SNVs, which directly genotypes subclones and the associated variabilities. Although not shown, posterior variances are easily obtained using MCMC samples for the $\boldsymbol{Z}$ and $\boldsymbol{w}$ matrices in our examples. More importantly, the feature allocation model, cIBP, reflects the underlying evolutionary biology of clonal expansion, such as the population "infinite sites assumption" in which mutations occurred in parental subclones are passed on to all offsprings. However, we do not explicitly enforce this assumption in modeling the SNVs, fearing that some subclones might not be present in the tumor samples as they lose to other subclones in the fitness selection. Instead, we explicitly model overlapping SNVs across subclones and let the data fit to dictates the subclonal genotypes. This is a distinction from clustering-based approaches in the existing literature.

There can be a number of possible extensions to the current model. First, the number of SNVs examined in this paper was relatively limited (about 150). Other than computational complexity, there is no limitation on extending the current model to analyze a large set of SNVs. We have begun to investigate efficient computational algorithms to take on a large number of SNVs, see Ref. [26].

As another important extension, we are considering joint modeling SNVs and copy number variations (CNVs) using linked feature allocation models. Briefly, we could consider a sampling model for the total read counts N_{st} to estimate the sample copy numbers, conditional on which a couple of feature allocation models can be linked for estimating subclonal copy numbers and

DNA sequences.

References

1. R. A. Weinberg, *The biology of cancer* (Garland Science New York, 2007).
2. N. Navin, A. Krasnitz, L. Rodgers, K. Cook, J. Meth, J. Kendall, M. Riggs, Y. Eberling, J. Troge, V. Grubor *et al.*, *Genome research* **20**, 68 (2010).
3. N. D. Marjanovic, R. A. Weinberg and C. L. Chaffer, *Clinical chemistry* **59**, 168 (2013).
4. V. Almendro, A. Marusyk and K. Polyak, *Annual Review of Pathology: Mechanisms of Disease* **8**, 277 (2013).
5. K. Polyak, *The Journal of clinical investigation* **121**, p. 3786 (2011).
6. J. Stingl and C. Caldas, *Nature Reviews Cancer* **7**, 791 (2007).
7. M. Shackleton, E. Quintana, E. R. Fearon and S. J. Morrison, *Cell* **138**, 822 (2009).
8. D. L. Dexter, H. M. Kowalski, B. A. Blazar, Z. Fligiel, R. Vogel and G. H. Heppner, *Cancer Research* **38**, 3174 (1978).
9. J. Lee, P. Müller, Y. Ji and K. Gulukota, *A Bayesian Feature Allocation Model for Tumor Heterogeneity*, tech. rep., UC Santa Cruz (2013).
10. T. L. Griffiths and Z. Ghahramani, *Journal of Machine Learning Research* **12**, 1185 (2011).
11. S. Nik-Zainal, P. Van Loo, D. C. Wedge, L. B. Alexandrov, C. D. Greenman, K. W. Lau, K. Raine, D. Jones, J. Marshall, M. Ramakrishna *et al.*, *Cell* **149**, 994 (2012).
12. A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté and S. P. Shah, *Nature methods* (2014).
13. C. A. Miller, B. S. White, N. D. Dees, M. Griffith, J. S. Welch, O. L. Griffith, R. Vij, M. H. Tomasson, T. A. Graubert, M. J. Walter *et al.*, *PLoS computational biology* **10**, p. e1003665 (2014).
14. F. Strino, F. Parisi, M. Micsinai and Y. Kluger, *Nucleic acids research* **41**, e165 (2013).
15. W. Jiao, S. Vembu, A. G. Deshwar, L. Stein and Q. Morris, *BMC bioinformatics* **15**, p. 35 (2014).
16. H. Zare, J. Wang, A. Hu, K. Weber, J. Smith, D. Nickerson, C. Song, D. Witten, C. A. Blau and W. S. Noble, *PLoS computational biology* **10**, p. e1003703 (2014).
17. S. Sengupta, J. Ho and A. Banerjee, *Two Models Involving Bayesian Nonparametric Techniques*, tech. rep., University Of Florida (2013).
18. Y. Kim, L. James and R. Weissbach, *Biometrika* **99**, 127 (2012).
19. A. E. Gelfand, Model determination using sampling-based methods, in *Markov chain Monte Carlo in practice*, (Springer, 1996) pp. 145–161.
20. L. Held, B. Schrödle and H. Rue, Posterior and cross-validatory predictive checks: a comparison of mcmc and inla, in *Statistical Modelling and Regression Structures*, (Springer, 2010) pp. 91–110.
21. D. M. Church, V. A. Schneider, T. Graves, K. Auger, F. Cunningham, N. Bouk, H.-C. Chen, R. Agarwala, W. M. McLaren, G. R. Ritchie *et al.*, *PLoS Biology* **9**, p. e1001091 (2011).
22. H. Li and R. Durbin, *Bioinformatics* **25**, 1754 (2009).
23. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly *et al.*, *Genome research* **20**, 1297 (2010).
24. D. L. Longo, *N Engl J Med* **366**, 956 (2012).
25. A. Marusyk, V. Almendro and K. Polyak, *Nature Reviews Cancer* **12**, 323 (2012).
26. Y. Xu, P. Muller, Y. Yuan, Y. Ji and K. Gulukota, *arXiv preprint arXiv:1402.5090* (2014).