Bayesian Nonparametric Inference for Tumor Purity and Tumor Subclones

Juhee Lee

Department of Applied Mathematics and Statistics, University of California Santa Cruz,

Santa Cruz, California, U.S.A.

email: juheelee@soe.ucsc.edu

and

Subhajit Sengupta

Program for Computational Genomics and Medicine, Northshore University HealthSystem,

Evanston, Illinois, U.S.A.

email: subhajit06@gmail.com

and

Raymond Hovey

Program for Computational Genomics and Medicine, Northshore University HealthSystem,

Evanston, Illinois, U.S.A.

email: RHovey@northshore.org

and

Yuan Ji

Program for Computational Genomics and Medicine, Northshore University HealthSystem,

Evanston, Illinois, U.S.A. & Department of Health Studies, The University of Chicago, Chicago, Illinois, U.S.A.

email: YJi@northshore.org

SUMMARY: A tumor sample consists of heterogeneous cells, such as normal cells and subpopulations of tumor cells called tumor subclones. We present *BayClone3* as a nonparametric Bayesian method that extends Lee et al. (2016)

to reconstruct tumor subclones in the presence of normal cells. *BayClone3* separates tumor cells from normal cells in a sample and describes tumor cells with tumor subclones. Utilizing three random matrices ($\boldsymbol{L}$, $\boldsymbol{Z}$ and $\boldsymbol{w}$), tumor subclones are characterized with copy number aberrations ($\boldsymbol{L}$), variant allele counts ($\boldsymbol{Z}$) and cellular population frequencies ($\boldsymbol{w}$). We use two sets of data, read counts from next-generation sequencing experiments and locus-specific subclonal copy numbers of tumor cells provided by upstream bioinformatics tools to enhance inference on tumor purity and subclonal structure. Using simulation studies and a real data analysis, we demonstrate improved inference, especially for datasets with small sample sizes.

KEY WORDS: Categorical Indian buffet process; Feature allocation models; Markov chain Monte Carlo; Next-generation sequencing; Random matrices; Tumor purity, Tumor heterogeneity.

1. Introduction

Solid tumor samples are typically an admixture of tumor cells and normal cells. Normal cells contain germline DNA and tumor cells contain DNA with somatic mutations such as somatic copy number aberrations (SCNAs) and single nucleotide variants (SNVs). When somatic mutations occur only in a fraction of tumor cells, cell population becomes heterogeneous and subpopulations of tumor cells called subclones give rise to tumor heterogeneity (Nowell, 1976; Marusyk and Polyak, 2010; Bonavia et al., 2011; Greaves and Maley, 2012; Wang et al., 2014). Subclonal mutations arise and are accumulated over the entire life of a tumor, making tumor cells dynamically and spatially different in their genomes. Importantly, these genome variations in space and time often lead to drastically different diagnosis and prognosis within and between cancer patients and are a fundamental cause of treatment resistance (Zardavas et al., 2015; Marusyk et al., 2012; McGranahan and Swanton, 2015).

Figure 1 presents a stylized example of tumor purity and heterogeneity in two tumor samples. Panel (a) shows three genomic loci on the genomes for normal cells and two tumor subclones. Normal cells do not contain any somatic mutations on these loci while the tumor subclones possess common and private somatic mutations. The two tumor samples in (b) consist of the three subpopulations described in panel (a) with different population frequencies. Our goal is to recover the subclonal mutations and the population frequencies of the tumor subclones using next-generation sequencing (NGS) data.

Recent development has generated various computational methods and useful tools for subclonal reconstruction. In general, methods can be categorized into two approaches, indirect inference and direct inference. The indirect inference approach uses a mixture model to cluster SNVs based on variant allele frequencies (VAFs, the proportion of reads with variant allele at a locus). It later indirectly obtains subclonal genotypes based on its inferred clustering. Examples include PhyloSub (Jiao et al., 2014), PhyloWGS (Deshwar et al., 2015), PyClone (Roth et al., 2014), THetA (Oesper et al., 2013), SciClone (Miller et al., 2014), BitPhylogeny (Yuan et al., 2015). Among these, some models (e.g. SciClone and Clomial) focus only on SNVs at copy neutral regions with heterozygous mutations and some other models (e.g., PyClone and PhyloSub) consider both SNV and copy number aberration (CNA) and correct the VAFs for the CNA values. It is studied that subclonal genotype construction based on mutation clustering could be sensitive to the choices of hypeparameter values for mixture models. For the direct inference approach, the main aim is to directly reconstruct subclonal genotypes by inferring two quantities, a mutation assignment matrix for characterization of subclonal geneotypes and a vector (or a matrix for multiple samples) of population frequencies for subclones. Examples of this approach are CloneHD (Fischer et al., 2014), Clomial (Zare et al., 2014), TrAp (Strino et al., 2013), BayClone (Sengupta et al., 2015), BayClone2 (Lee et al., 2016). TrAp assumes a binary matrix for subclonal structure and directly uses matrix factorization to obtain a local optimal solution. It discusses the issue of model identifiability stating the need to have sufficient sample size for unique mathematical solutions. Our proposed method BayClone3 takes a direct inference approach by placing a prior distribution on unknown underlying subclonal structure. It provides a fully model-based and probabilistic inference on the underlying structure with uncertainty quantification. Another aspect that can be considered for subclonal reconstruction is to impose a phylogenetic tree structure among tumor subclones, for example, see PhyloSub, PhyloWGS and BitPhylogeny. As a model choice, BayClone3 does not assume a phylogenetic tree structure among subclones since not all subclones on the phylogenetic tree may be present in any given tumor sample. Instead, subclones in a tumor sample may only represent nodes on a subset of branches of the phylogenetic tree. With this notion, BayClone3 uses data to infer the underlying subclones without tree structure.

The aim of the proposed method is to simultaneously investigate tumor purity which refers to the fraction of tumor cells in a tumor sample, and reconstruct tumor subclones. In BayClone2 (Lee et al., 2016), we use NGS counts data and develop a Bayesian feature allocation model to infer tumor subclonal structure jointly with subclonal DNA copy numbers, subclonal variant allele counts and population frequencies of subclones. Here, we extend BayClone2 (Lee et al., 2016) and provide novel insights of tumor heterogeneity including tumor purity and subclonal structure. Since the accurate estimation of tumor purity is crucial to precise characterization of tumor subclones (Larson and Fridley, 2013; Bao et al., 2014), BayClone3 models tumor purity explicitly. Furthermore, BayClone3 incorporates locus-specific subclonal copy numbers of tumor cells inferred by upstream bioinformatics analysis such as analysis from Battenburg (Nik-Zainal et al., 2012) or FACET (Shen and Seshan, 2016) in addition to read counts directly from NGS experiments. This achieves significantly improved inference on the overall tumor subclonal structure as well as subclonal CNVs, especially when the number of samples is tiny or NGS data has low sequencing depth. We use matrices $\boldsymbol{L}$ and $\boldsymbol{Z}$ to represent subclonal copy numbers and variant allele counts. Figure 1(c) shows a representation of the subclonal structure in the two samples in panels (a) and (b). On the top, two matrices $\boldsymbol{L}$ and $\boldsymbol{Z}$ are used to describe subclonal copy numbers and the corresponding number of variant alleles. Columns and rows of the matrices correspond to tumor subclones and loci, respectively. On the bottom, a population frequency matrix called $\boldsymbol{w}$ shows the population frequencies of the tumor subclones and the normal cell fraction in both samples. BayClone3 assumes that elements in all three matrices are random, including the number C of columns in the $\boldsymbol{L}$ and $\boldsymbol{Z}$ matrices.

The remainder of the paper is organized as follows: Section 2 describes the proposed Bayesian feature allocation model. Section 3 introduces the posterior inference including a post processing step to prune the inferred but negligible or similar subclones. Sections 4 and 5 report simulation studies and an analysis for a TCGA dataset. The last section concludes the paper with a discussion.

2. Probability Model

2.1 *Sampling Model*

We use $t = 1, \ldots, T$ and $s = 1, \ldots, S$ to index tumor samples and genomic loci, respectively. We utilize two sets of data. One consists of the total number of reads mapped to locus s in sample t, $\boldsymbol{N} = [N_{st}]$ and the number of reads with a variant sequence (relative to normal cells) at that locus, $\boldsymbol{n} = [n_{st}]$. The other data contains the copy numbers of the tumor genome at locus s in sample t, $\boldsymbol{M} = [M_{st}]$, $M_{st} > 0$, provided from a separate upstream bioinformatics analysis.

We let $(1 - \mu_t)$, $0 < \mu_t < 1$ denote the proportion of tumor cells in sample t called tumor purity, and μ_t represents the proportion of normal cells. We assume that normal cells are copy number neutral with two copies of the genome at all loci. Different from normal cells, tumor cells may possess distinct copy numbers that may or may not equal two, leading to a potentially non-integer copy number at a locus when averaged across subclones. As such, let m_{st} represent the latent expected copy number of all tumor cells at locus s in sample t. Since m_{st} is not observed, we assume that a copy number analysis (e.g., Battenberg or FACET) is performed prior to our analysis and an inferred copy number M_{st} is made available for BayClone3. M_{st} could be either an integer or a non-integer. An integer value of M_{st} means that the copy number analysis concludes that all tumor cells have the same copy number at locus s in sample t, i.e., the tumor cell copy number is clonal. A non-integer value of M_{st} indicates that the tumor cell copy number is subclonal, implying that tumor subclones have different copy numbers. We introduce a locus-specific binary indicator λ_s which takes on the value 1 if M_{st} takes the same integer value in all samples, $\lambda_s = \prod_{t=1}^{T} \mathrm{I}\{M_{st} = \lfloor M_{st'} \rfloor\}$ for any t', where $\mathrm{I}(\cdot)$ is an indicator function and $\lfloor x \rfloor$ is the largest integer less than or equal to x. Note that λ_s is not a parameter. When the upstream analysis concludes that all samples have the same clonal copy number for tumor cells (i.e., $\lambda_s = 1$), we assume the tumor cell copy number at locus s is known and treat M_{st} as the known copy number. For such loci, we model only the total read counts N_{st} conditional on M_{st}. We assume

$$N_{st} \mid \mu_t, \phi_t, M_{st}, \lambda_s = 1 \overset{indep}{\sim} \text{Poi}\left(\frac{\phi_t}{2} \times \{\mu_t \times 2 + (1-\mu_t)M_{st}\}\right). \tag{1}$$

In modeling N_{st}, we adopt the idea that the number of total reads reflects the overall number of alleles (or copy number) after mixing normal cells and tumor cells. To see this, we denote ϕ_t as the expected number of reads (i.e., expected sequencing depth) when the copy number is 2. We define $Tot_{allele} = \{\mu_t \times 2 + (1-\mu_t)M_{st}\}$ in (1) as the expected copy number averaged across tumor and normal cells, with 2 being the normal cell copy number and M_{st} the average tumor cell copy number, weighted by μ_t and $(1-\mu_t)$, the fractions of normal and tumor cells, respectively. In the case where there is no copy number aberration at locus s in any tumor subclone (i.e. $M_{st} = 2$), the expected total read count in the Poisson model becomes ϕ_t.

Recall that the event $\lambda_s = 0$ means that tumor copy number is subclonal, or samples have different clonal tumor copy numbers. For loci with $\lambda_s = 0$, we model both M_{st} and N_{st}. In specific, assuming conditional independence of M_{st} and N_{st} we let

$$\begin{aligned} N_{st} \mid \mu_t, \phi_t, m_{st}, \lambda_s = 0 &\overset{indep}{\sim} \text{Poi}\left(\frac{\phi_t}{2} \times \{\mu_t \times 2 + (1-\mu_t)m_{st}\}\right), \\ M_{st} \mid m_{st}, \kappa_t, \lambda_s = 0 &\overset{indep}{\sim} \text{Gamma}(\psi_{st}, \kappa_t), \end{aligned} \tag{2}$$

where $\text{Gamma}(a, b)$ denotes a gamma distribution with mean a/b, and $\psi_{st} = m_{st} \times \kappa_t$ (i.e., $\mathrm{E}(M_{st}) = m_{st}$). A main reason that we model M_{st} when $\lambda_s = 0$ is that the upstream bioinformatics caller indicates tumor cells are heterogeneous in their copy number at the locus and only provides inferred copy numbers averaged across tumor subclones M_{st}. Due to uncertainty in the estimation of M_{st}, we model M_{st} to infer subclone specific copy numbers which is essential for subclone construction. Similar to the case of $\lambda_s = 1$, we let $Tot_{allele} = \{\mu_t \times 2 + (1 - \mu_t) m_{st}\}$ for loci having $\lambda_s = 0$. For locus s where there is no copy number aberration in any tumor subclone, $\mathrm{E}(M_{st}) = m_{st} = 2$ and the expected total read count in the Poisson model becomes ϕ_t, and it has the same interpretation as that in (1).

Lastly, for the number of variant reads n_{st}, we assume

$$n_{st} \mid N_{st}, p_{st} \quad \overset{indep}{\sim} \quad \mathrm{Bin}(N_{st}, p_{st}). \tag{3}$$

Here p_{st} is the expected VAFs, i.e., the proportion of variant alleles in the tumor genomes. We will model p_{st} as a function of tumor purity, subclonal copy numbers and variant allele counts, shown next. Together, equations (1) – (3) define the sampling model for data $\{M_{st}, N_{st}, n_{st}\}$.

2.2 *Prior*

L** and **Z Let C denote the unknown number of tumor subclones in T samples. We introduce two $S \times C$ matrices, $\boldsymbol{L} = [\ell_{sc}]$ and $\boldsymbol{Z} = [z_{sc}]$, $s = 1, \ldots, S$ and $c = 1, \ldots, C$ to represent subclonal copy numbers and numbers of variant alleles along S loci for C subclones, respectively. A genomic representation of tumor subclone c is defined by two column vectors $\boldsymbol{\ell}_c$ and $\boldsymbol{z}_c$, where $\boldsymbol{\ell}_c = \{\ell_{sc},\ s = 1, \ldots, S\}$ and $\boldsymbol{z}_c = \{z_{sc},\ s = 1, \ldots, S\}$. Specifically, $\ell_{sc} \in \{0, 1, \ldots, Q\}$ represents the number of DNA alleles (no more than a prespecified integer Q) at locus s in subclone c and $z_{sc} \in \{0, \ldots, \ell_{sc}\}$ the number of alleles with a sequence mutation at that locus. For example, the event $\{\ell_{sc} = 3, z_{sc} = 1\}$ indicates that three copies of alleles reside at locus s in subclone c and among the three alleles one has a variant sequence that is different from that in normal cells. We introduce a probability vector $\boldsymbol{\pi}_c = (\pi_{c0}, \ldots, \pi_{cQ})$ with $0 < \pi_{cq} < 1$, $q = 1, \ldots, Q$ and $\sum_{q=0}^{Q} \pi_{cq} = 1$, and construct a prior $p(\boldsymbol{\pi}, \boldsymbol{L}, \boldsymbol{Z} \mid C)$ conditional on C. We later place a prior on C to infer the unknown number of tumor subclones. The joint prior of $\boldsymbol{\pi}$, $\boldsymbol{L}$ and $\boldsymbol{Z}$ is factored as $p(\boldsymbol{\pi} \mid C) p(\boldsymbol{L} \mid \boldsymbol{\pi}, C) p(\boldsymbol{Z} \mid \boldsymbol{L}, \boldsymbol{\pi}, C)$ using feature allocation priors (Sengupta et al., 2015; Lee et al., 2016); for $c = 1, \dots, C$ and $s = 1, \dots, S$,

$$\begin{aligned} \boldsymbol{\pi}_c \mid C &\overset{iid}{\sim} \text{Beta-Dirichlet}(\alpha/C, \beta, \gamma_0, \gamma_1, \gamma_3, \dots, \gamma_Q), && (4) \\ \ell_{sc} \mid \boldsymbol{\pi}_c &\overset{indep}{\sim} \text{Multinomial}(1, \boldsymbol{\pi}_c), && (5) \\ z_{sc} \mid \ell_{sc} &\overset{indep}{\sim} \text{Uniform}(0, \dots, \ell_{sc}). && (6) \end{aligned}$$

The Beta-Dirichlet distribution in (4) assumes that $\pi_{c2} \overset{iid}{\sim} \text{Be}(\beta, \alpha/C)$ and $\pi_{cq} = (1 - \pi_{c2})\tilde{\pi}_{cq}$, $q = 0, 1, 3, \dots, Q$ with $(\tilde{\pi}_{c0}, \tilde{\pi}_{c1}, \tilde{\pi}_{c3}, \dots, \tilde{\pi}_{cQ}) \overset{iid}{\sim} \text{Dir}(\gamma_0, \gamma_1, \gamma_3, \dots, \gamma_Q)$. By integrating out $\boldsymbol{\pi}_c$ and sending $C \to \infty$, the prior for $\boldsymbol{L}$ in (4) and (5) becomes the categorical Indian buffet process that defines a distribution over Q-nary random matrices with infinite dimension (Lee et al., 2016). In addition to C tumor subclones, we introduce an additional column in both matrices that represents a hypothetical subclone, which we call subclone 0, to account for the experimental noise and tumor cells with negligible abundance. Arbitrarily, we fix $\ell_{s0} = 2$ and $z_{s0} = 2$ for all loci s in subclone 0.

$\mathbf{m_{st}}$*,* $\mathbf{p_{st}}$ *and* $\boldsymbol{\mu}_\mathbf{t}$ Recall that the expected tumor cell copy number m_{st} is only defined for loci s having $\lambda_s = 0$. Following its biological meaning, m_{st} is obtained by averaging copy numbers across tumor subclones. In specific, we use $\boldsymbol{L}$ and $\boldsymbol{w}$ and define

$$m_{st} = \sum_{c=0}^{C} w_{tc}\ell_{sc}. \tag{7}$$

Here the $(C + 1)$-dimensional vector $\boldsymbol{w}_t = (w_{t0}, \dots, w_{tC})$ denotes population frequencies of C tumor subclones and the hypothetical background subclone, denoted as subclone 0, for sample t, where $0 < w_{tc} < 1$, $c = 0, \dots, C$ and $\sum_{c=0}^{C} w_{tc} = 1$. Another way to view $\boldsymbol{L}$ and m_{st} in (7) is that m_{st} is deconvoluted into subclonal copy numbers ℓ_{sc}, $c = 0, \dots, C$ with their weights w_{tc}.

We view the expected VAFs p_{st} as the probabilities of observing a short read that possesses a somatic mutation. The following generative model of sampling a variant read in a tumor sample is used to construct p_{st}; Recall that $\text{Tot}_{\text{allele}}$ is the average number of alleles across all cells at locus s in sample t. We define the expected number of variant alleles across C

tumor subclones as $\mathrm{Var}_{\text{allele}} = \sum_{c=1}^{C} w_{tc} z_{sc}$ since z_{sc} represents the number of variant alleles at locus s in subclone c. Next, accounting for tumor purity $(1 - \mu_t)$ and sequencing noise in the NGS experiment with subclone 0, the probability of observing a variant read p_{st} is given by

$$p_{st} = \frac{(1-\mu_t)(noise + \mathrm{Var}_{\text{allele}})}{\mathrm{Tot}_{\text{allele}}} = \frac{(1-\mu_t)(p_0 w_{t0} z_{s0} + \sum_{c=1}^{C} w_{tc} z_{sc})}{\mu_t \times 2 + (1-\mu_t) m_{st}}, \tag{8}$$

where p_0 denotes a tiny probability of observing a variant read due to noise in an experiment and is common for all samples. We let $0 < p_0 \ll 1$ and keep w_{t0} small to avoid any identifiability problem.

We use a beta distribution for $\mu_t \overset{iid}{\sim} \mathrm{Be}(a_\mu, b_\mu)$ and a Dirichlet distribution for $\boldsymbol{w}_t \overset{iid}{\sim} \mathrm{Dir}(d_0, d, \ldots, d)$ with $0 < d_0 \ll d$, $t = 1, \ldots, T$. Finally, we complete the model specification by placing priors for C, ϕ_t and p_0. Recall that ϕ_t accounts for different average read counts in T samples and p_0 is the proportion of variant reads due to noise. We assume $C \sim \mathrm{Geometric}(a)$ with mean $1/a$, $\phi_t \overset{indep}{\sim} \mathrm{Gamma}(a_\phi, b_\phi)$ and $p_0 \sim \mathrm{Be}(a_0, b_0)$.

3. Posterior Inference

3.1 *Trans-dimensional MCMC Simulation*

Let $\boldsymbol{x} = (\boldsymbol{L}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{w}, \boldsymbol{\phi}, \boldsymbol{\pi}, p_0)$ denote all the unknown parameters besides C, where $\boldsymbol{\mu} = \{\mu_t\}$, $\boldsymbol{\phi} = \{\phi_t\}$, $\boldsymbol{w} = \{w_{tc}\}$ and $\boldsymbol{\pi} = \{\pi_{cq}\}$. Recall that C determines the dimensions of $\boldsymbol{L}$, $\boldsymbol{Z}$ and π. We generate a Monte Carlo sample of $(C^{(i)}, \boldsymbol{x}^{(i)}) \sim p(C, \boldsymbol{x} \mid \boldsymbol{n}, \boldsymbol{N}, \boldsymbol{M})$, $i = 1, \ldots, I$ through posterior Markov chain Monte Carlo (MCMC) simulation. We use Gibbs sampling transition probabilities to update $\boldsymbol{x}$ for fixed C. To update C, we use the trans-dimensional MCMC strategy proposed in Lee et al. (2015). The strategy is motivated by fractional Bayes factors (O'Hagan, 1995) and simplifies calculation of the acceptance probability of the Metropolis-Hastings algorithm for updating C. For efficient implementation, the posterior distribution for a given value of C is constructed using the full likelihood raised to the power b with

$0 < b < 1$ as a training dataset, that is, $p_1(\boldsymbol{x} \mid \boldsymbol{n}, \boldsymbol{N}, \boldsymbol{M}, C) \propto p(\boldsymbol{x} \mid C)\{p(\boldsymbol{n}, \boldsymbol{N}, \boldsymbol{M} \mid \boldsymbol{x})\}^b$. We then treat the observed data with the likelihood function $\{p(\boldsymbol{n}, \boldsymbol{N}, \boldsymbol{M} \mid \boldsymbol{x})\}^{1-b}$ as the testing set and use $p_1(\boldsymbol{x} \mid \boldsymbol{n}, \boldsymbol{N}, \boldsymbol{M}, C)$ as the prior for inference of the testing data and as the proposal distribution for the Metropolis-Hastings algorithm. Since the likelihood function is a product of binomial, Poisson, and gamma densities, the inference under the proposed procedure is equivalent to the original posterior inference without splitting. The choice of b is simple. Due to the overly informative likelihood from a typical NGS dataset, we use a close-to-1 value for b so that it downweights the likelihood of the testing data. We found that using b_{st} instead of common b yields improved mixing of the Markov chains. Specifically, random $b_{st} \overset{iid}{\sim} \text{Be}(900, 100)$ works well, as the mean of the beta distribution is 0.9 with a small variance ($< 9E - 5$). That is, effectively we use 90% of information for training and 10% for testing. For more details, see Lee et al. (2015, 2016).

3.2 *Post Processing*

It is known that read counts N_{st} and n_{st} are measured with noise. The noise may affect posterior inference more significantly for data with a large number S of loci, low sequencing depth and/or small sample sizes. For example, *BayClone3* sometimes produces subclones with negligible population frequencies w_{tc} for all t or similar columns in $\boldsymbol{L}$ and $\boldsymbol{Z}$. We propose a post-processing procedure that uses a posterior Monte Carlo sample and achieve parsimonious interpretation of subclonal structure.

Denote a posterior Monte Carlo sample, $\{(C^{(i)}, \boldsymbol{x}^{(i)}, p_{st}^{(i)}), i = 1, \ldots, I\}$ where I is the number of MCMC draws. For iteration i, we process in two steps, *removal* and *merge*. In the removal step, we remove tumor subclones with small population frequencies $(1 - \mu_t)w_{tc} < \epsilon_1$ in all tumor samples where ϵ_1 is prespecified. After removal, we set $w_{tc} = 0$ for the removed subclones and redistribute the original w_{tc} associated with them over the remaining subclones, $c' \neq c$, $1 \leqslant c' \leqslant C^{(i)}$ proportional to $w_{tc'}$ within a sample. We then proceed to the merge step to combine subclones that are similar to each other in $\boldsymbol{L}$ and $\boldsymbol{Z}$. Suppose that $\tilde{C}^{(i)}$ subclones are left after the removal and the subclones are renumbered from 1 to $\tilde{C}^{(i)}$. We consider a pair of tumor subclones c_1 and c_2 for merging, $c_1 \neq c_2$, $1 \leqslant c_1, c_2 \leqslant \tilde{C}^{(i)}$. In the merge step, we first propose to keep the subclone with the larger value of $\sum_{t=1}^{T} w_{tc}$, which is the aggregated population frequencies across all samples. Without loss of generality, we assume that c_1 is kept and c_2 is removed. We then reallocate the w's and define a new $\boldsymbol{w}'$. Set $w'_{tc_1} = w_{tc_1} + w_{tc_2}$, $w'_{tc_2} = 0$ and $w'_{tc} = w_{tc}$, $c \neq c_1, c_2$. Also set $\ell'_{sc_2} = 0$ and $z'_{sc_2} = 0$ and $\ell'_{sc} = \ell_{sc}$ and $z'_{sc} = z_{sc}$ for $c \neq c_2$. Recompute p'_{st} according to (8) using w'_{st}, ℓ'_{sc} and z'_{sc} for all s and t. By this point, we have obtained a new proposal of $\boldsymbol{w}'$, $\boldsymbol{L}'$, $\boldsymbol{Z}'$, and $\{p'_{st}\}$ to replace $\boldsymbol{w}$, $\boldsymbol{L}$, $\boldsymbol{Z}$, and $\{p_{st}\}$ of the current MCMC iteration. We accept this merge proposal based on a model fitting criterion. Specifically, the proposal is accepted if the $u-$th percentile of the absolute differences $|p_{st} - p'_{st}|$ is less than ϵ_2. For our analysis, $\epsilon_1 = 0.05$, $\epsilon_2 = 0.15$ and the third quartile with $u = 75\%$ are used.

3.3 *Posterior Summary*

We obtain a point estimate of $(C, \boldsymbol{x})$ to summarize the joint posterior distribution using the posterior Monte Carlo sample after the post-processing following Lee et al. (2015). The joint posterior $p(C, \boldsymbol{x} \mid \boldsymbol{n}, \boldsymbol{N}, \boldsymbol{M})$ is factorized as

$$p(C \mid \boldsymbol{n}, \boldsymbol{N}, \boldsymbol{M})\, p(\boldsymbol{L} \mid \boldsymbol{n}, \boldsymbol{N}, \boldsymbol{M}, C)\, p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{w} \mid \boldsymbol{n}, \boldsymbol{N}, \boldsymbol{M}, C, \boldsymbol{L})\, p(\boldsymbol{\phi}, \boldsymbol{\mu}, p_0 \mid \boldsymbol{n}, \boldsymbol{N}, \boldsymbol{M}, C).$$

We evaluate the marginal posterior $p(C \mid \boldsymbol{n}, \boldsymbol{N}, \boldsymbol{M})$ and determine the maximum a posteriori (MAP) estimate $C^\star$. To obtain a posterior point estimate of $\boldsymbol{L}$, $\boldsymbol{L}^\star$, we let $\mathcal{D}_{cc'}(\boldsymbol{L}, \boldsymbol{L}') = \sum_{s=1}^{S} |\ell_{sc} - \ell'_{sc'}|$ for any two $S \times C^\star$ matrices $\boldsymbol{L}$ and $\boldsymbol{L}'$, and define a distance between the matrices as $d(\boldsymbol{L}, \boldsymbol{L}') = \min_{\boldsymbol{\sigma}} \sum_{c=1}^{C^\star} \mathcal{D}_{c,\sigma_c}(\boldsymbol{L}, \boldsymbol{L}')$, where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_C)$ is a permutation of $\{1, \ldots, C^\star\}$. Note that the minimum is taken over all possible permutations. We let $\boldsymbol{L}^\star = \arg\min_{\boldsymbol{L}'} \int d(\boldsymbol{L}, \boldsymbol{L}')\, dp(\boldsymbol{L} \mid \boldsymbol{n}, \boldsymbol{N}, C^\star)$. We approximate $\boldsymbol{L}^\star$ using a posterior Monte Carlo sample, $\{\boldsymbol{L}^{(i)}, i = 1, \ldots, I\}$. We report posterior point estimates $\boldsymbol{Z}^\star$ and $\boldsymbol{w}^\star$ conditional on

$C^\star$ and $\boldsymbol{L}^\star$. Finally, we report $\boldsymbol{\mu}^\star$, $\boldsymbol{\phi}^\star$ and $p_0^\star$ as the posterior mean of $\boldsymbol{\mu}$, $\boldsymbol{\phi}$ and p_0 conditional on $C^\star$.

4. Simulation Studies

4.1 *Simulation 1*

Simulation Setup We assess the performance of BayClone3 through simulation studies. We assume $S = 1,000$ loci and vary $T = 1$ or 3 samples. As the simulation truth, we assume two tumor subclones $C^{\text{TRUE}} = 2$. The heatmap in Figure 2(a) shows the simulation truth, $\boldsymbol{L}^{\text{TRUE}}$ and $\boldsymbol{Z}^{\text{TRUE}}$ where the rows and columns in the heatmap correspond to loci and subclones, respectively. Red, dark red, black, green and light green colors in the heatmaps represent 0, 1, 2, 3 and 4, respectively. For example, dark red color in $\boldsymbol{L}$ indicates copy number neutral and the same color in $\boldsymbol{Z}$ means two variant copies. We generate $\boldsymbol{w}_t^{\text{TRUE}}$ from $\text{Dir}(1, 30, 70)$ for $T = 1$. In cases with $T = 3$ samples, we permute $(30, 70)$ for each t and generate $\boldsymbol{w}_t^{\text{TRUE}}$ independently from the Dirichlet distribution with the permuted concentration parameters. $\boldsymbol{w}_t^{\text{TRUE}}$ is illustrated in Supplementary Figure 1. We set the true fraction of normal cells μ_t^{TRUE} to be one of the values $\{0.1, 0.4, 0.7\}$ for all tumor samples in a simulated dataset, resulting in six different simulation settings with $T = 1$ or 3 and $\mu_t^{\text{TRUE}} \in \{0.1, 0.4, 0.7\}$. As the true normal cell fraction μ_t^{TRUE} increases, the tumor cell fraction decreases and the reconstruction of tumor subclones gets more challenging. We generate the expected sequencing depth $\phi_t^{\text{TRUE}} \overset{iid}{\sim} \text{Gamma}(75, 3)$ (mean 25). With the assumed $\boldsymbol{L}^{\text{TRUE}}$, $\boldsymbol{Z}^{\text{TRUE}}$, $\boldsymbol{w}^{\text{TRUE}}$, μ_t^{TRUE} and ϕ_t^{TRUE}, we compute m_{st}^{TRUE} and p_{st}^{TRUE} using (7) and (8), respectively. We let $M_{st} = m_{st}^{\text{TRUE}}$ for the loci with the same clonal CNV in all samples, and for the other loci, we generate $M_{st} \overset{indep}{\sim} \text{Gamma}(m_{st}^{\text{TRUE}} \times \kappa_t, \kappa_t)$ with $\kappa_t = 32$. We finally generate $N_{st} \overset{indep}{\sim} \text{Poi}(\phi_t^{\text{TRUE}}\{2\mu_t^{\text{TRUE}} + (1 - \mu_t^{\text{TRUE}})m_{st}^{\text{TRUE}}\}/2)$ and $n_{st} \overset{indep}{\sim} \text{Bin}(N_{st}, p_{st}^{\text{TRUE}})$.

Results To fit the proposed model, we fix the hyperparameters as $r = 0.2$, $\alpha = 2$, $\beta = 10$, $\gamma_q = 0.5$ for $q = 0, 1, 3, 4 (= Q)$, $d_0 = 0.1$, $d = 1$, $a_{00} = 30$ and $b_{00} = 500$. For the prior on ϕ_t, we let $b_\phi = 20$ and specify a_ϕ such that the prior mean $\mathrm{E}(\phi) = a_\phi/b_\phi$ equals the sample median of N_{st} among loci having $M_{st} = 2$. To calibrate hyperparameters for the prior of μ_t, we first get an estimate $\tilde{\mu}_t$ based on the observed VAF using a Dirichlet process mixture model implemented in the R package DPpackage (Jara et al., 2011; Jara, 2007), and center the prior around $\tilde{\mu}_t$ by letting $a_\mu = \tilde{\mu}_t \times S \times \mathrm{E}(\phi_t)$ and $b_\mu = (1 - \tilde{\mu}_t) \times S \times \mathrm{E}(\phi_t)$. We find that BayClone3 better performs with these informative priors for ϕ_t and μ_t. For each value of C, we use read count data to initialize $\boldsymbol{Z}$ and $\boldsymbol{L}$. We initialize w_{tc} and p_0 with prior draws. We generate $b_{st} \overset{iid}{\sim} \mathrm{Be}(900, 100)$ to construct the training set and run the MCMC simulation over 7,000 iterations, discarding the first 3,000 iterations as initial burn-in. The post-processing is implemented with $\epsilon_1 = 0.05$, $\epsilon_2 = 0.15$ and $u = 0.75$.

The posterior distributions of C are illustrated for each of the six simulation settings in Supplementary Figure 2. Figure 2(b)–(g) shows heatmaps of the point estimates $\boldsymbol{L}^\star$ and $\boldsymbol{Z}^\star$ conditional on $C^\star$. In all three cases where $T = 3$ and one case where $T = 1$ and $\mu^{\text{TRUE}} = 0.1$, BayClone3 infers $C^\star = 2$ subclones which agrees with the truth. Comparing (b) and (e)–(g) of Figure 2 to (a), the estimated $\boldsymbol{L}^\star$ and $\boldsymbol{Z}^\star$ are close to their simulation truth as well. For the remaining two cases ($T = 1, \mu^{\text{TRUE}} = 0.4$; and $T = 1, \mu^{\text{TRUE}} = 0.7$), BayClone3 estimates $C^\star = 1$ and identifies the sublone with larger w_{tc}^{TRUE} (i.e., subclone 2 of $\boldsymbol{L}^{\text{TRUE}}$ and $\boldsymbol{Z}^{\text{TRUE}}$). The normal cell fraction μ_t is estimated with high accuracy (for example, $\mu_t^\star = 0.109$, 0.416 and 0.696 for $T = 1$). Supplementary Figure 2 illustrates $\boldsymbol{w}_t^\star$ and $\mu_t^\star$. Comparing with their simulation truth in Supplementary Figure 1, we observe small discrepancy between $w_{tc}^\star$ and w_{tc}^{TRUE} and the estimated population frequencies $w_{tc}^\star(1 - \mu_t^\star)$ is closer to their true value. Overall, even with a single sample and large normal cell contamination BayClone3 recovers the true tumor subclonal structure reasonably well.

Comparison We compare BayClone3 to two methods, BayClone2 (Lee et al., 2016) and PyClone (Roth et al., 2014), one of the most popular methods for tumor heterogeneity study. BayClone2 uses the feature allocation model in (4)–(6) using read count data only. It produces inference on $\boldsymbol{L}$, $\boldsymbol{Z}$ and $\boldsymbol{w}$ but not tumor purity $(1-\mu_t)$ (See Eq (5) of Lee et al. (2016)). For BayClone2, hyperparameter values are set in a similar way and post-processing and posterior summary are performed the same as BayClone3. Figure 3 (b)–(g) show the heatmaps of $\boldsymbol{L}^\star$ and $\boldsymbol{Z}^\star$ under BayClone2. Supplementary Figures 3 and 4 illustrate the posterior distributions of C and $\boldsymbol{w}_t^\star$, respectively. BayClone2 does not attempt to estimate tumor purity and tends to accommodate normal cells as subclones in $\boldsymbol{L}^\star$ and $\boldsymbol{Z}^\star$ (e.g. subclones 1 and 2 in Figure 3(b)). Estimates of ℓ_{st} and z_{st} under BayClone2 tend to have extreme values, either close to zero or close to the maximum number allowed in the model (e.g., red or green colors at many loci in subclone 3 in Figure 3(b)), resulting in poor inference. This shows that using the additional information about the copy numbers M_{st} of tumor cells enhances the inference on $\boldsymbol{L}$ and subsequently on the other parameters including $\boldsymbol{Z}$ and $\boldsymbol{w}$. As a numerical summary, Table 1 presents the sum of absolute differences (SADs) between estimated $(\boldsymbol{L}^\star, \boldsymbol{Z}^\star)$ and the true $(\boldsymbol{L}^{\text{TRUE}}, \boldsymbol{Z}^{\text{TRUE}})$ evaluated under the two methods, where $SAD = \frac{1}{C^\star}\sum_{c=1}^{C^\star} \min_{c'=1,\ldots,C^{\text{TRUE}}}\{\sum_{s=1}^{S}(|\ell_{sc}^\star - \ell_{sc'}^{\text{TRUE}}| + |z_{sc}^\star - z_{sc'}^{\text{TRUE}}|)\}$. See the left two columns for SADs in Simulation 1 with $\text{E}(\phi_t^{\text{TRUE}}) = 25$. Having $(\boldsymbol{\ell}_c^\star, \boldsymbol{z}_c^\star)$ very different from the true subclones produces a large value of SAD. It is clear that BayClone3 provides improved inference in the simulation compared to BayClone2. PyClone is also applied for comparison. PyClone uses a Dirichlet process mixture to model the read counts data. It defines variant allelic prevalence that is similar to our p_{st} in (8), produces clustering of loci, and estimates variant allelic prevalence at each locus. Figure 4 illustrates a heatmap of variant allelic prevalence estimates and p_{st}^{TRUE}. The values in the heatmap are standardized in each column and the loci are rearranged according to the inferred cluster memberships for easy comparison. When multiple samples are available, it provides better estimates of variant allelic prevalence. It seems that their estimates are close to the true p_{st}^{TRUE} for cases with low tumor purity. Since PyClone is a mixture model approach, it does not attempt to construct subclones.

We further examine how the inference changes as the coverage ϕ_t^{TRUE} increases. The same $\boldsymbol{L}^{\text{TRUE}}$, $\boldsymbol{Z}^{\text{TRUE}}$ and $\boldsymbol{w}_t^{\text{TRUE}}$ are used, but we generate $\phi_t^{\text{TRUE}} \overset{iid}{\sim} \text{Gamma}(150, 3)$. The posterior inference under BayClone3 is summarized in Supplementary Figures 6–8. Overall the inference is improved with larger ϕ_t^{TRUE}. For example, $\boldsymbol{L}^{\star}$ and $\boldsymbol{Z}^{\star}$ are closer to the truth even with large normal cell contamination μ_t^{TRUE}. Inference under BayClone2 is summarized in Supplementary Figures 9-11. Its inference is improved but still worse than that under BayClone3. In particular, it tends to take either a very large value or small value for ℓ_{st} or z_{st}. SADs are summarized in Table 1. Posterior estimates of variant allelic prevalence under PyClone are shown in Supplementary Figure 12. From comparison to p_{st}^{TRUE}, PyClone's estimates are also improved.

4.2 *Simulation 2*

An additional simulation study that assumes three tumor subclones in the simulation truth (more subclones than in Simulation 1) is reported in Section 3 of Supplementary Material. In Simulation 2, sublones get smaller w_{tc}^{TRUE} than those in Simulation 1, which makes Simulation 2 more challenging. BayClone3 generates reasonable inference on the true subclonal structure. For small T and/or large μ_t^{TRUE}, BayClone3 tends to recover subclones with large w_{tc}^{TRUE} and its performance is improved when more samples are available ($T = 3$) or when the sequencing depth is increased ($\text{E}(\phi_t^{\text{TRUE}}) = 50$). Comparison to BayClone2 and PyClone is also performed and illustrated in Supplementary Figures 17-20. Similar to Simulation 1, BayClone 3 performs better than BayClone2. See also Table 1 for SADs. BayClone2 reports very poor inference especially for $T = 1$, larger μ_t^{TRUE} or small ϕ_t^{TRUE}.

5. TCGA Data

We use two real tumor samples ($T = 2$) in The Cancer Genome Atlas (TCGA) from the same patient to show the performance of BayClone3 on real data. The two samples are obtained from the primary and metastatic tumors, respectively, and sequenced using whole genome sequencing (WGS) for analysis. The average sequencing depth is around 45X. VCF (variant call format) files for both samples are downloaded from the TCGA website (https://tcga-data.nci.nih.gov/tcga/). Multiple somatic variant callers are used in TCGA for generating the somatic variant callset and we choose the output callset generated by VARSCAN (Koboldt et al., 2009). After that, we only keep "high-confidence" calls in order to ensure the quality of the SNV calls. From the VCF files for each SNV location, the number of reads N_{st} and the number of variant reads n_{st} are recorded. We use the Battenberg algorithm (Nik-Zainal et al., 2012) to call tumor cell copy numbers for each sample. We obtain a set of loci that are common in both samples and record the read counts (N_{st} and n_{st}) for them along with their respective tumor copy numbers (M_{st}). A total of $S = 2,767$ SNVs are called in both samples, and we removed SNVs that have unusually large M_{st} or N_{st} compared to other SNVs. We also removed the SNVs for which VAF is zero in both samples. This filtering leaves us with $S = 1{,}443$ loci. The data after the preprocessing is illustrated in Figure 5. We run a multi-sample analysis with the three methods including BayClone3.

We use hyperparameter values similar to those used in the simulations for BayClone3 and BayClone2. We run MCMC and conduct the post-processing with the same values of ϵ_1, ϵ_2 and u. The posterior inference under BayClone3 is summarized in Figure 6(a)–(c) and (g). Panel (a) of the figure indicates two subclones ($C^\star = 2$) and panel (b) shows the estimated subclonal copy numbers $\boldsymbol{L}^\star$ and mutations $\boldsymbol{Z}^\star$. Panel (c) shows the estimated population frequencies in the two tumor samples, where the solid and dotted lines represent $\boldsymbol{w}_t^\star$ and $\boldsymbol{w}_t^\star(1-\mu_t^\star)$, respectively, in different colors for the two samples (Sample 1 in black and Sample

2 in red). The inference implies that both tumor samples have a mix of normal cells and tumor cells. Their tumor purity estimates are similar, $1-\mu_t^\star = 0.849$ and 0.876, respectively. More interestingly, population frequency estimates $\boldsymbol{w}_t^\star$ indicate that the subclonal structure in the samples is different. The sample from the primary tumor, Sample 1 contains two subclones with population frequencies $\boldsymbol{w}_1^\star = 0.742$ and 0.254, respectively. In contrast, the metastatic tumor sample, Sample 2 is mainly dominated by subclone 1 that has $w_{21}^\star = 0.924$. One possible explanation is that during metastasis, one of the two subclones disseminated to a new location of the patient body and formed metastases. It can also be seen from Figure 6(b) that subclone 1 has many copy number losses and potentially loss of heterozygosity events, which needs to be further annotated in downstream analysis. We also compare $\hat{p}_{st}$ under BayClone3 (the first two columns of the heatmap in Figure 6(g)) to the observed VAFs n_{st}/N_{st} (the last two columns in the heatmap). The rows are rearranged for easy comparison. The heatmap shows that BayClone3 fits empirical VAFs well.

For comparison, BayClone2 and PyClone are applied to the same dataset and their inferences are summarized in the two bottom rows of Figure 6. The rows of $\boldsymbol{L}^\star$ and $\boldsymbol{Z}^\star$ in panel (e) are arranged in the same order of the rows in panel (b). BayClone2 yields $C^\star = 2$ and both samples have very similar population frequencies as shown in panel (f), leading to implications different from those under BayClone3. BayClone2 indicates that the samples have subclonal mutations and their subclonal structure is homogeneous. In addition, it is observed that $\boldsymbol{L}^\star$ under BayClone2 does not well coincide with M_{st} produced from Battenberg in Figure 5(c) and (f). Especially, subclone 2 under BayClone2 has large population frequency estimates in both samples, $w_{t2}^\star = 0.751$ and 0.767, respectively, and it has two copies loss at many loci (red color in column 2 of $\boldsymbol{L}^\star$), resulting in its tumor cell copy number estimates smaller than 2. However, many loci in Sample 1 has $M_{st} > 2$. From panel (g), BayClone2 provides reasonable estimates of p_{st} compared to empirical VAF. PyClone clusters the 1,443 loci into 76 groups, two of which have 1321 and 46 loci, respectively, and the others are mostly singletons. A heatmap of their variant allelic prevalence estimates is shown in panel (g) with the observed VAF n_{st}/N_{st}.

6. Conclusion and Discussion

The proposed model infers tumor purity and tumor subclonal structure from NGS data. Subclones are described with DNA copy numbers and variant allele counts. BayClone3 can be applied to data sets with multiple samples by allowing sample-specific population frequencies of the inferred subclones. Through intensive simulation studies, we showed that accurate separation of normal cells from a mixture of normal and tumor cells is very important by itself and further for inference on tumor subclones. We also showed that the inference on tumor heterogeneity (TH) is significantly enhanced by incorporating subclonal copy number calls from bioinformatics tools. There are easily accessible bioinformatics methods that efficiently infer copy number aberrations of tumor cells over the entire genome, such as Battenberg or FACET. Typically the sample size for TH studies is small and the sequencing depth of NGS data is around 30X-35X for WGS in practice. The improvement by incorporating the additional data is crucial, especially for datasets with small sample sizes and/or low sequencing depth as shown in the simulation studies.

The major motivation for accurate estimation of heterogeneity in tumor is personalized medicine. Currently the heterogeneity is inferred mostly based on SNV and CNA data. Other structural variants (SVs) such as deletion and duplication may enable more accurate VAF estimation, which is a possible key element for characterizing the heterogeneity (Fan et al., 2014). Apart from these DNA level genomic changes there is evidence on heterogeneous epigenetic abnormalities in cancer (Brocks et al., 2014). The proposed BayClone3 can be easily extended to incorporate them through taking additional DNA methylation data. Finally, in the era of big data, efficient computational methods are more needed to handle millions of

SNVs. Linear response Variational Bayes (Giordano et al., 2015) or MAD-Bayes (Broderick et al., 2013; Xu et al., 2015) methods can be considered as an alternative computational strategy to tackle the problem.

Acknowledgments

Yuan Jis research was partially supported by NIH R01 CA132897-7.

Supplementary Materials

Supplementray Material, referenced in Section 4, is available with this paper at the Biometrics website on Wiley Online Library.

References

Bao, L., Pu, M., and Messer, K. (2014). Abscn-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. *Bioinformatics* page btt759.

Bonavia, R., Cavenee, W. K., Furnari, F. B., et al. (2011). Heterogeneity maintenance in glioblastoma: a social network. *Cancer research* **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 reports* **8,** 798–806.

Broderick, T., Kulis, B., and Jordan, M. (2013). Mad-bayes: Map-based asymptotic derivations from bayes. In *Proceedings of The 30th International Conference on Machine Learning*, pages 226–234.

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 biology* **16,** 1.

Fan, X., Zhou, W., Chong, Z., Nakhleh, L., and Chen, K. (2014). Towards accurate characterization of clonal heterogeneity based on structural variation. *BMC bioinformatics* **15,** 1.

Fischer, A., Vázquez-García, I., Illingworth, C. J., and Mustonen, V. (2014). High-Definition Reconstruction of Clonal Composition in Cancer. *Cell Reports* **7,** 1740–1752.

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*, pages 1441–1449.

Greaves, M. and Maley, C. C. (2012). Clonal evolution in cancer. *Nature* **481,** 306–313.

Jara, A. (2007). Applied Bayesian non- and semi-parametric inference using DPpackage. *R News* **7,** 17–26.

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011). DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software* **40,** 1–30.

Jiao, W., Vembu, S., Deshwar, A. G., Stein, L., and Morris, Q. (2014). Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC bioinformatics* **15,** 1.

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., Weinstock, G. M., Wilson, R. K., and Ding, L. (2009). Varscan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25,** 2283–2285.

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., Ji, Y., et al. (2015). A bayesian feature allocation model for tumor heterogeneity. *The Annals of Applied Statistics* **9,** 621–639.

Lee, J., Müller, P., Sengupta, S., Gulukota, K., and Ji, Y. (2016). Bayesian inference for intratumour heterogeneity in mutations and copy number variation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65,** 547–563.

Marusyk, A., Almendro, V., and Polyak, K. (2012). Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer* **12,** 323–334.

Marusyk, A. and Polyak, K. (2010). Tumor heterogeneity: causes and consequences. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* **1805,** 105–117.

McGranahan, N. and Swanton, C. (2015). Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer cell* **27,** 15–26.

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 computational biology* **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.

Oesper, L., Mahmoody, A., and Raphael, B. J. (2013). Theta: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome Biol* **14,** R80.

O'Hagan, A. (1995). Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B* **57,** 99–138.

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* .

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)*, volume 20, pages 467–478.

Shen, R. and Seshan, V. E. (2016). Facets: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput dna sequencing. *Nucleic acids research* page gkw520.

Strino, F., Parisi, F., Micsinai, M., and Kluger, Y. (2013). Trap: a tree approach for fingerprinting subclonal tumor composition. *Nucleic acids research* **41,** e165–e165.

Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., et al. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512,** 155–160.

Xu, Y., Müller, P., Yuan, Y., Gulukota, K., and Ji, Y. (2015). Mad bayes for tumor heterogeneity?feature allocation with exponential family sampling. *Journal of the American Statistical Association* **110,** 503–514.

Yuan, K., Sakoparnig, T., Markowetz, F., and Beerenwinkel, N. (2015). Bitphylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. *Genome biology* **16,** 1.

Zardavas, D., Irrthum, A., Swanton, C., and Piccart, M. (2015). Clinical management of breast cancer heterogeneity. *Nature reviews Clinical oncology* **12,** 381–394.

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 computational biology* **10,** e1003703.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

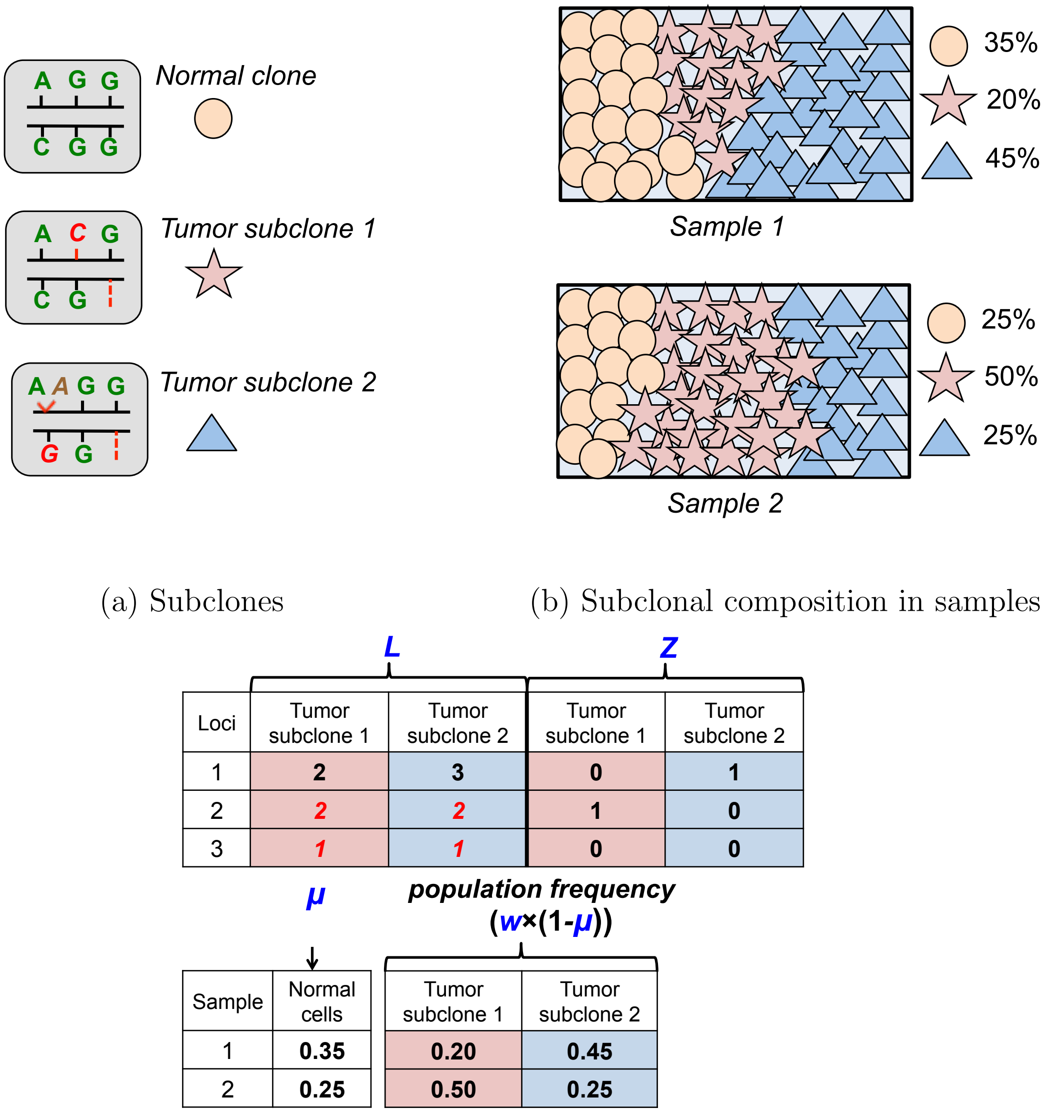
[Figure 4 about here.]

[Table 1 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

Received October 2007. *Revised February* 2008. *Accepted March* 2008.



(a) Subclones

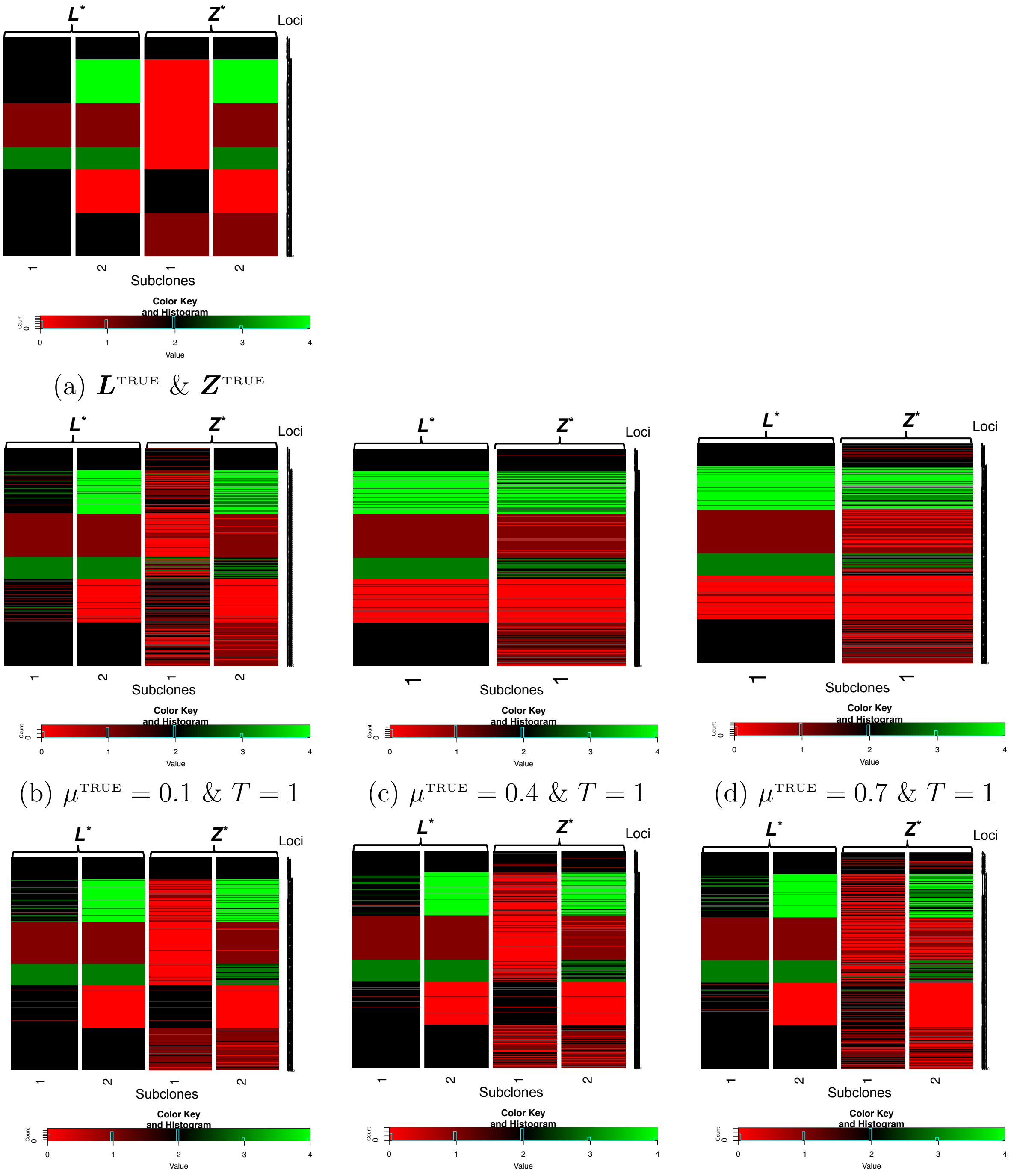
(b) Subclonal composition in samples

Loci	Tumor subclone 1	Tumor subclone 2	Tumor subclone 1	Tumor subclone 2
1	**2**	**3**	**0**	**1**
2	***2***	***2***	**1**	**0**
3	***1***	***1***	**0**	**0**

Sample	Normal cells	Tumor subclone 1	Tumor subclone 2
1	**0.35**	**0.20**	**0.45**
2	**0.25**	**0.50**	**0.25**

(c) Mathematical representation of the subclonal structure

Figure 1. (a) Genetic characterization of hypothetical subclones, (b) hypothetical subclonal composition in two samples and (c) a mathematical representation of the underlying subclonal structure in (a) and (b).



(a) $\boldsymbol{L}^{\text{TRUE}}$ & $\boldsymbol{Z}^{\text{TRUE}}$

(b) $\mu^{\text{TRUE}} = 0.1$ & $T = 1$

(c) $\mu^{\text{TRUE}} = 0.4$ & $T = 1$

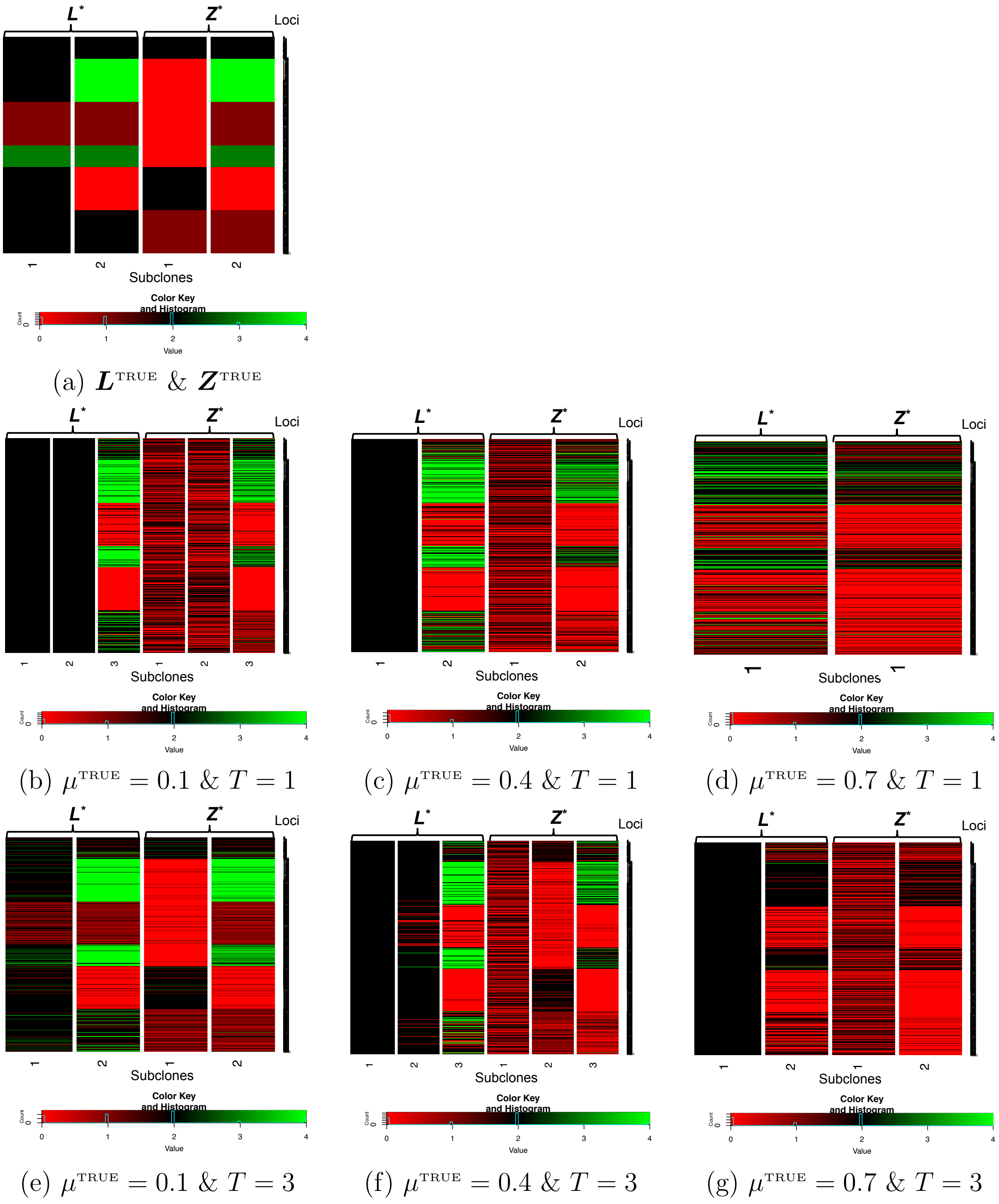
(d) $\mu^{\text{TRUE}} = 0.7$ & $T = 1$

(e) $\mu^{\text{TRUE}} = 0.1$ & $T = 3$

(f) $\mu^{\text{TRUE}} = 0.4$ & $T = 3$

(g) $\mu^{\text{TRUE}} = 0.7$ & $T = 3$

Figure 2. [BayClone3 Results for Simulation 1] (a) Heatmap of $\boldsymbol{L}^{\text{TRUE}}$ and $\boldsymbol{Z}^{\text{TRUE}}$. (b)–(g) Heatmaps of posterior point estimates of $\boldsymbol{L}$ and $\boldsymbol{Z}$, $\boldsymbol{L}^{\star}$ and $\boldsymbol{Z}^{\star}$ with μ_t^{TRUE} =0.1(left), 0.4(middle) and 0.7(right). The second and third rows are for $T = 1$, and 3, respectively.



(a) $\boldsymbol{L}^{\text{TRUE}}$ & $\boldsymbol{Z}^{\text{TRUE}}$

(b) $\mu^{\text{TRUE}} = 0.1$ & $T = 1$

(c) $\mu^{\text{TRUE}} = 0.4$ & $T = 1$

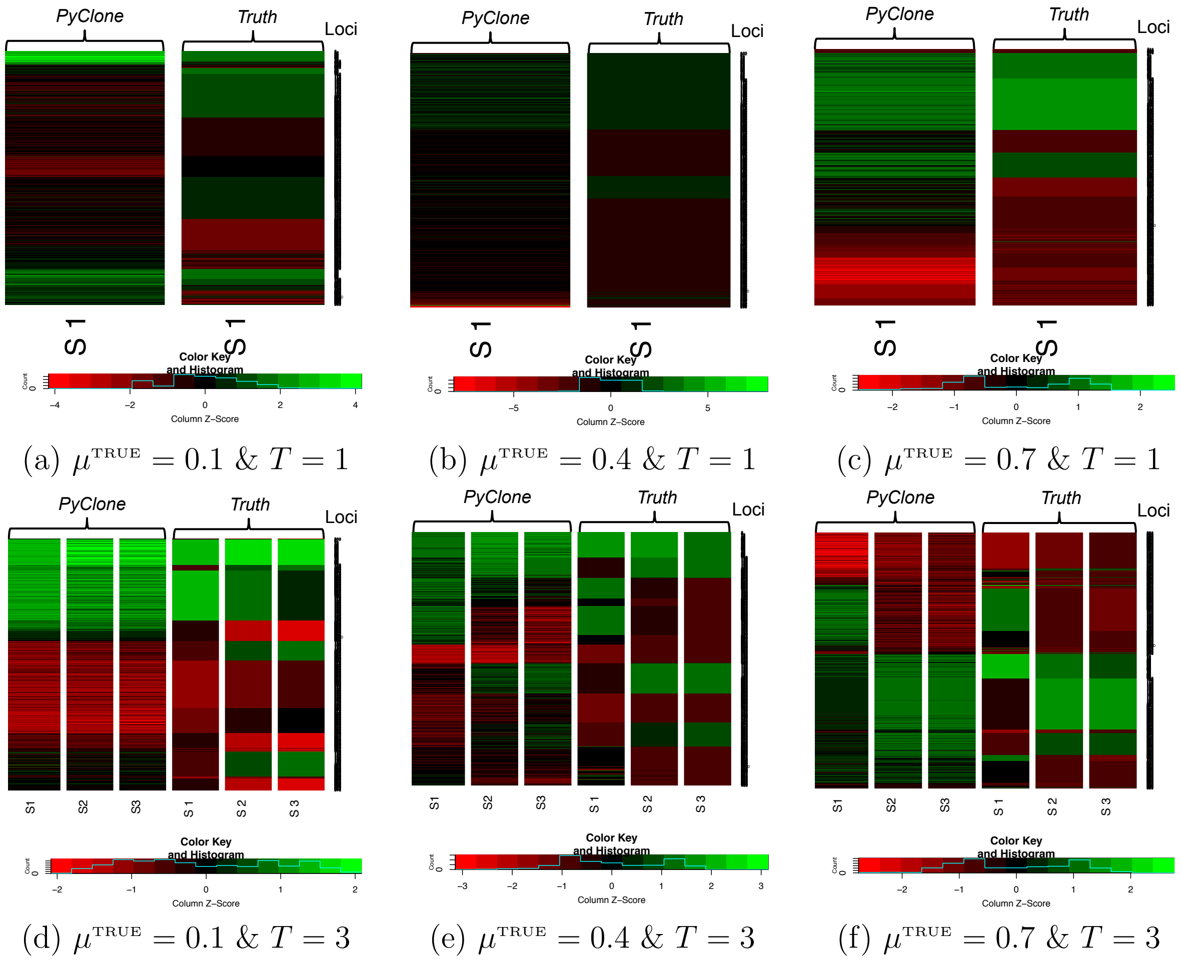
(d) $\mu^{\text{TRUE}} = 0.7$ & $T = 1$

(e) $\mu^{\text{TRUE}} = 0.1$ & $T = 3$

(f) $\mu^{\text{TRUE}} = 0.4$ & $T = 3$

(g) $\mu^{\text{TRUE}} = 0.7$ & $T = 3$

Figure 3. [BayClone2 Results for Simulation 1] (a) Heatmap of $\boldsymbol{L}^{\text{TRUE}}$ and $\boldsymbol{Z}^{\text{TRUE}}$. (b)–(g) Heatmaps of posterior point estimates of $\boldsymbol{L}$ and $\boldsymbol{Z}$, $\boldsymbol{L}^{\star}$ and $\boldsymbol{Z}^{\star}$ with μ_t^{TRUE} =0.1(left), 0.4(middle) and 0.7(right) and $T = 1$(second row) and 3(third row). The simulation truth is shown in (a) for each comparison.



(a) $\mu^{\text{TRUE}} = 0.1$ & $T = 1$

(b) $\mu^{\text{TRUE}} = 0.4$ & $T = 1$

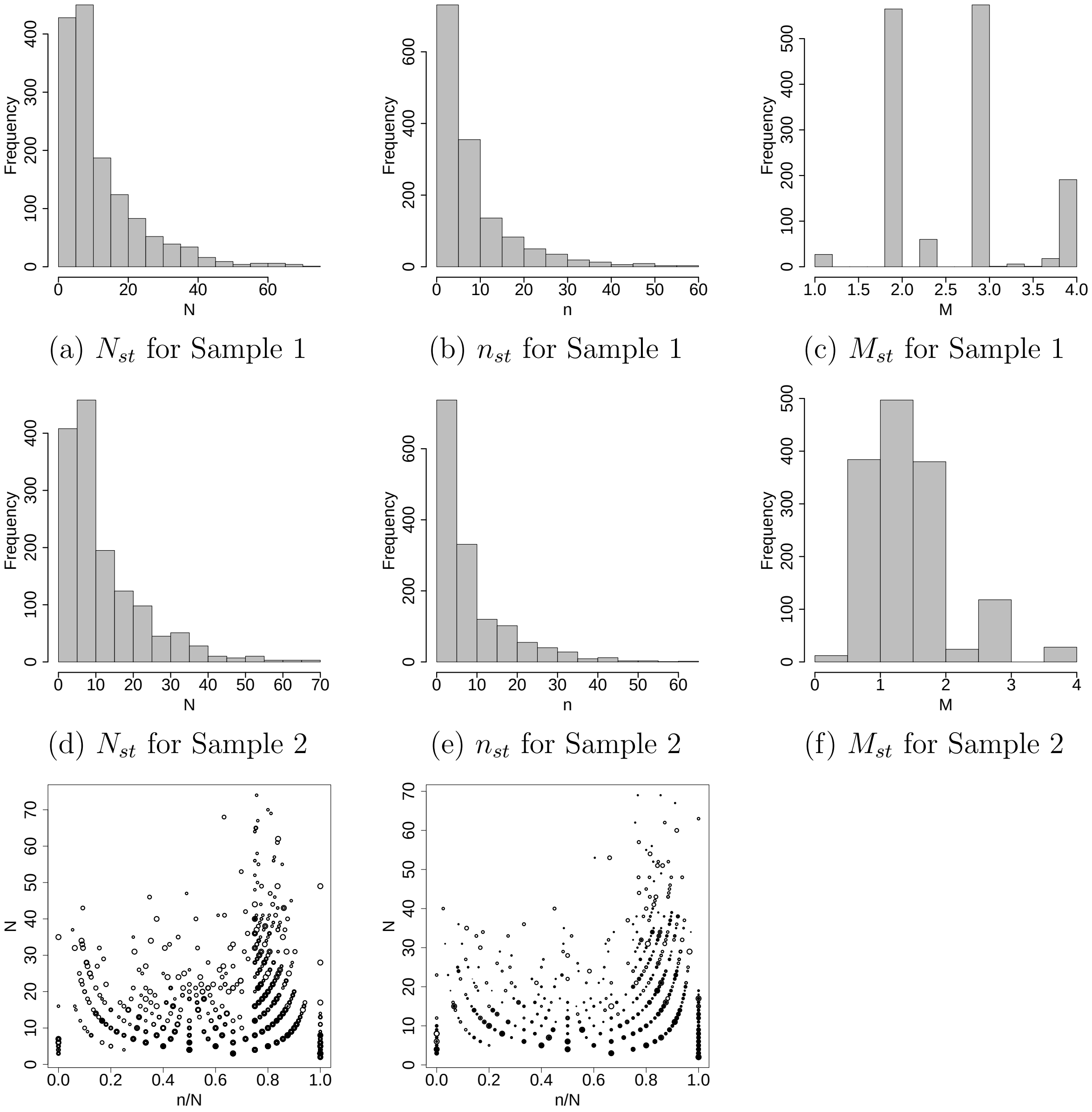
(c) $\mu^{\text{TRUE}} = 0.7$ & $T = 1$

(d) $\mu^{\text{TRUE}} = 0.1$ & $T = 3$

(e) $\mu^{\text{TRUE}} = 0.4$ & $T = 3$

(f) $\mu^{\text{TRUE}} = 0.7$ & $T = 3$

Figure 4. [PyClone Results for Simulation 1] Heatmaps of posterior estimates of variant allelic prevalence and p_{st}^{TRUE} with μ_t^{TRUE} =0.1(left), 0.4(middle) and 0.7(right) and $T = 1$(top) and 3(bottom).



(a) N_{st} for Sample 1

(b) n_{st} for Sample 1

(c) M_{st} for Sample 1

(d) N_{st} for Sample 2

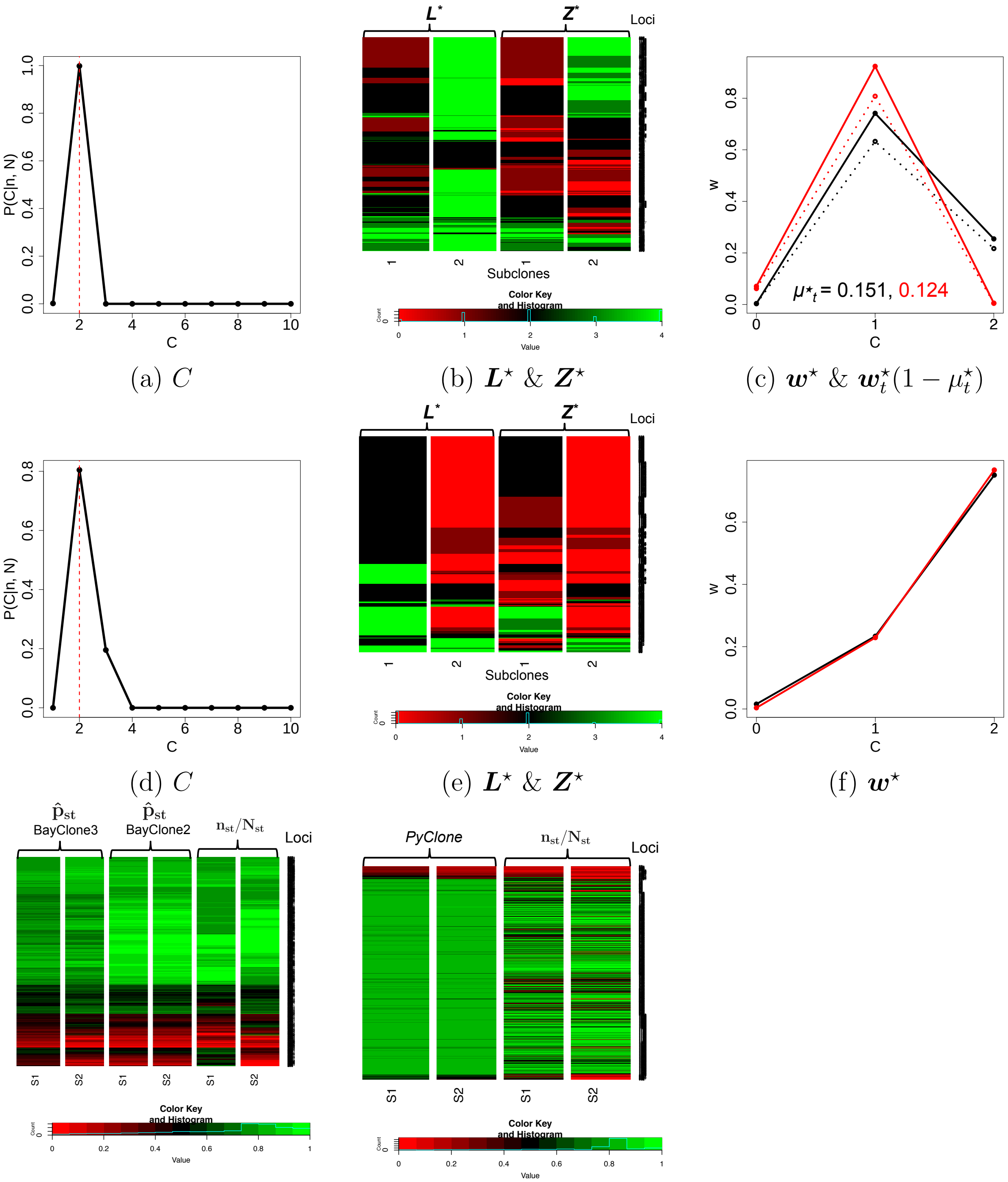
(e) n_{st} for Sample 2

(f) M_{st} for Sample 2

(g) N_{st} vs n_{st}/N_{st}, $t = 1$

(h) N_{st} vs n_{st}/N_{st}, $t = 2$

Figure 5. [TCGA Data] Histograms of N_{st}, n_{st} and M_{st} in (a)–(f) for the TCGA data. The top and middle rows are for Samples 1 and 2, respectively. Scatterplots of N_{st} vs n_{st}/N_{st} are shown in (g) and (h).



(a) C (b) $\boldsymbol{L}^\star$ & $\boldsymbol{Z}^\star$ (c) $\boldsymbol{w}^\star$ & $\boldsymbol{w}_t^\star(1-\mu_t^\star)$

(d) C (e) $\boldsymbol{L}^\star$ & $\boldsymbol{Z}^\star$ (f) $\boldsymbol{w}^\star$

(g) $\hat{p}_{st}$ under BayClone3 (h) Variant allelic prevalence under PyClone

Figure 6. [Results for TCGA Data] (a)–(c) have plots of the posterior inference under BayClone3. In (c) $\boldsymbol{w}_t^\star$ and $\boldsymbol{w}_t^\star(1-\mu_t^\star)$ are represented in solid lines and dotted lines, respectively, and two samples in black (Sample 1) and red (Sample 2) colors, respectively. (d)–(f) plots of posterior inference under BayClone2. Panel (g) shows heatmaps of $\hat{p}_{st}$'s under BayClone3 and BayClone2 and the empirical VAFs n_{st}/N_{st}. Panel (h) has heatmaps of posterior estimates of variant allelic prevalence under PyClone and the empirical VAFs n_{st}/N_{st}.

Simulation Setting			**(A)** Simulation 1		**(B)** Simulation 2	
E(ϕ_t^{TRUE})	T	μ_t^{TRUE}	BayClone3	BayClone2	BayClone3	BayClone2
25	1	0.1	752.0	1541.0	360.0	1352.0
		0.4	475.0	1900.0	443.0	2122.8
		0.7	707.0	3397.0	963.0	1813.0
	3	0.1	168.0	459.5	658.8	2301.5
		0.4	299.0	1302.3	502.5	2358.0
		0.7	678.0	2690.5	1085.5	2145.5
50	1	0.1	651.0	1366.7	235.0	391.0
		0.4	861.5	1547.5	310.0	1592.0
		0.7	486.0	2354.0	481.0	2134.0
	3	0.1	90.0	286.5	805.5	2982.0
		0.4	168.0	1044.0	438.0	1796.3
		0.7	360.0	1916.0	1095.0	2024.0

Table 1

[Simulations 1 & 2] Sums of absolute differences between $(\boldsymbol{L}^\star, \boldsymbol{Z}^\star)$ *and* $(\boldsymbol{L}^{TRUE}, \boldsymbol{Z}^{TRUE})$ *are summarized under 12 different simulation settings for each of Simulations 1 & 2. Here* $E(\phi_t^{TRUE})$ *is the expected sequencing depth, i.e., the expected number of short reads mapped to a locus,* T *is the number of tumor samples, and* μ_t^{TRUE} *is the proportion of normal cells in the simulation. Larger values mean larger discrepancy averaged over* $C^\star$ *subclones.*