ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2016)
65, *Part* 4, *pp.* 547–563

Bayesian inference for intratumour heterogeneity in mutations and copy number variation

Juhee Lee,

University of California at Santa Cruz, USA

Peter Müller,

University of Texas at Austin, USA

Subhajit Sengupta and Kamalakar Gulukota

NorthShore University HealthSystem, Evanston, USA

and Yuan Ji

NorthShore University HealthSystem, Evanston, and University of Chicago, USA

[Received April 2015. Final revision September 2015]

Summary. Tissue samples from the same tumour are heterogeneous. They consist of different subclones that can be characterized by differences in DNA nucleotide sequences and copy numbers on multiple loci. Inference on tumour heterogeneity thus involves the identification of the subclonal copy number and single-nucleotide mutations at a selected set of loci. We carry out such inference on the basis of a Bayesian feature allocation model. We jointly model subclonal copy numbers and the corresponding allele sequences for the same loci, using three random matrices, **L**, **Z** and **w**, to represent subclonal copy numbers (**L**), the number of subclonal variant alleles (**Z**) and the cellular fractions (**w**) of subclones in one or more tumour samples respectively. The unknown number of subclones implies a random number of columns. More than one subclone indicates tumour heterogeneity. Using simulation studies and a real data analysis with next generation sequencing data, we demonstrate how posterior inference on the subclonal structure is enhanced with the joint modelling of both structure and sequencing variants on subclonal genomes. An R package is available from http://cran.r-project.org/web/packages/BayClone2/index.html.

Keywords: Categorical Indian buffet process; Feature allocation models; Markov chain Monte Carlo methods; Next generation sequencing; Random matrices; Subclone; Variant calling

1. Introduction

1.1. Biological background and motivation

Understanding tumour heterogeneity (TH) is critical for precise cancer prognosis. Not all tumour cells have the same genome and respond to the same treatment. TH arises when somatic mutations occur in only a fraction of tumour cells and results in the often observed spatial and temporal heterogeneity of tumour samples (Russnes *et al.*, 2011; Greaves and Maley, 2012; Frank and Nowak, 2003, 2004; Biesecker and Spinner, 2013; De, 2011; Bedard *et al.*, 2013;

Address for correspondence: Yuan Ji, Program of Computational Genomics and Medicine, NorthShore University HealthSystem, 1001 University Place, Evanston, IL 60201, USA.
E-mail: jiyuan@uchicago.edu

 0035–9254/16/65547

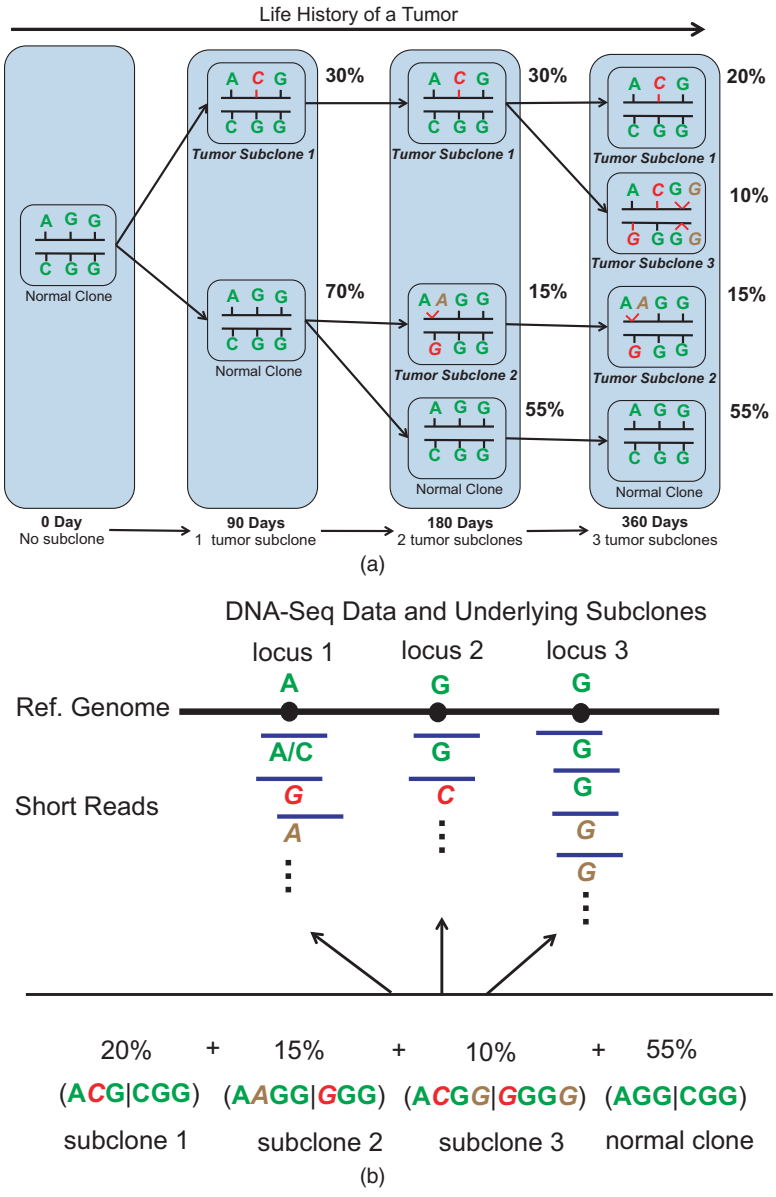

Fig. 1. (a) TH arising from clonal expansion (on days 90, 180 and 360, four somatic mutations (represented by red letters) and three somatic copy number gains (represented by brown letters) result in three tumour subclones) and (b) observed short reads (some with variants) are results of heterogeneous subclonal genomes (the formula at the bottom shows how subclonal alleles are mixed in proportions to produce short reads, which are mapped to different loci)

Navin *et al.*, 2011; Ding *et al.*, 2012). Fig. 1(a) illustrates this process with a hypothetical case in which the accumulation of sequence and copy number variants (CNVs) over the lifetime (360 days in this case) of a tumour gives rise to different subpopulations of tumour cells. Researchers have recently started to recognize the importance of TH and to realize the mistake of treating cancer by using a one size fits all approach. Instead, precision medicine now aims to develop

Table 1. Three matrices describing the subclonal structure in Fig. 1†

Locus	*Subclone 1*	*Subclone 2*	*Subclone 3*	*Normal clone*
(a) **L**				
1	2	3	2	2
2	2	2	2	2
3	2	3	4	2
(b) **Z**				
1	0	1	1	0
2	1	0	1	0
3	0	0	0	0
(c) **w**				
1	30%	0%	0%	70%
2	30%	15%	0%	55%
3	20%	15%	10%	55%

†**L** describes the subclonal copy numbers, **Z** describes the numbers of subclonal variant alleles and **w** describes the cellular fractions of subclones.

targeted treatment of individual tumours based on their molecular characteristics, including TH.

In Fig. 1(a), a tumour sample on day 360 (the last column in Fig. 1(a)) contains four subclones, including the normal clone without somatic mutations. The subclones are defined as distinct cell subtypes, each of which has a unique genome marked by single-nucleotide variants (SNVs) as sequence mutations and CNVs as structural mutations at one or more loci shown in Fig. 1(a). If the cells in this tumour sample are sequenced by using next generation deoxyribonucleic acid (DNA) sequencing, short reads that map to the three loci (Fig. 1(b)) will exhibit different sequences at the three loci and the number of reads at each locus will be informative about the copy number at the locus. For example, in Fig. 1(b), because of the copy number gains in subclones 2 and 3 (shown in Fig. 1(a)), additional reads with sequence A at locus 1 and additional reads with sequence G at locus 3 (both marked by brown letters) are expected. Also, because of sequence mutations in the subclones in Fig. 1(a), variant short reads will be generated harbouring mutational sequences at the loci. They are marked as the red letters in Fig. 1(b). Using next generation sequencing data we aim to recover the subclonal sequences at these loci and cellular fractions at the bottom of Fig. 1(b) that explain the true biology in Fig. 1(a). In particular, we aim to provide three matrices as shown in Table 1 to describe the subclonal genomes and sample heterogeneity. For illustration, Table 1 fills in the (biological) truth for these three matrices corresponding to the hypothetical TH that is described in Fig. 1(a). In an actual data analysis, all three matrices are latent and must be estimated.

1.2. Existing methods

Recent literature has introduced several useful tools for subclonal inference. These include, in particular, THetA (Oesper *et al.*, 2013), SciClone (Miller *et al.*, 2014), TrAp (Strino *et al.*, 2013), PhyloSub (Jiao *et al.*, 2014), PhyloWGS (Deshwar *et al.*, 2015), Clomial (Zare *et al.*, 2014) and CloneHD (Fischer *et al.*, 2014). THetA considers only subclonal copy numbers and is among the earliest methods for subclonal inference. TrAp emphasizes identifiability and sufficient sample size for unique mathematical solutions. SciClone and Clomial assume a binary matrix, focusing on SNVs at copy neutral regions with heterozygous mutations. PhyloSub and

PhyloWGS consider possible genotypes at SNVs accounting for potential copy number changes and phylogenetic constraints. CloneHD provides inference that is similar to our method but assumes the availability of data from matched normal samples. Also, CloneHD provides only point estimates of the subclonal copy numbers and subclonal mutations, lacking a description of uncertainty for the inferred subclones. In addition, PyClone (Roth *et al.*, 2014) and CHAT (Li and Li, 2014) adjust the estimation of subclonal cellular fractions for both CNVs and SNVs. However, they still stop short of directly inferring subclonal copy numbers or variant allele counts, which makes the method that is proposed in this paper distinct from the existing methods.

Most of these methods are based on either finite mixture models or Dirichlet process mixture models and aim to infer subclones on the basis of clusters of mutations. These methods implicitly assume that mutation clusters are direct results of clonal expansion that can be characterized by a phylogenetic tree and that all subclones in the tumour can be reconstructed by the observed mutation clusters. Although the assumption has a solid theoretical footing from an evolutionary biology perspective, we argue that the assumption may be violated because of limitations in experimental set-ups and data generation. For example, typically one or few biopsy samples are retrieved from a tumour, and the samples may or may not include all subclones on the hypothetical phylogenetic tree. When some subclones are not harvested from the biopsy, the assumption behind Dirichlet-process- and tree-based models might be violated. In contrast, our approach assumes that variant reads that have sequence or structural mutations can be generated only from subclones harbouring these mutations. Therefore, the variant allelic fractions (VAFs) are modelled as a mixture, mixing the subclones harbouring the mutations. Allowing overlapping mutations across subclones, we use a novel latent feature prior, the categorical Indian buffet process (CIBP) (Sengupta, 2013; Sengupta *et al.*, 2015) to model the subclonal copy numbers and subclonal variant allele counts. We build hierarchical Bayesian models to facilitate joint inference on all the unknown quantities simultaneously.

We use a Markov chain Monte Carlo (MCMC) scheme based on splitting the data into a small training set and a large test set developed in Lee *et al.* (2015). The approach exploits the nature of the next generation sequencing data and allows us to implement practicable transdimensional MCMC posterior simulation. See Section 2.3 and the on-line supplementary material for detail.

1.3. Model-based inference for tumour heterogeneity

We briefly summarize our modelling strategy. We propose a new class of Bayesian feature allocation models (Broderick *et al.*, 2013) to implement inference for the three matrices $\mathbf{L}$, $\mathbf{Z}$ and $\mathbf{w}$ that we use to characterize TH. We first construct an integer-valued random matrix $\mathbf{L}$ to characterize subclonal copy numbers. Each column corresponds to a subclone and each row corresponds to a genomic location. The number of columns C is unknown and random. The cth column $\mathbf{l}_c = (l_{1c}, \ldots, l_{Sc})$ records copy numbers across S loci for subclone c, $c = 1, \ldots, C$. For example, in Fig. 2, $l_{sc} = 3$ for $s = 1$ and $c = 2$ since subclone 2 has three alleles at locus 1. As a prior distribution for $\mathbf{L}$, $p(\mathbf{L})$, we use a finite version of a CIBP. Next, we introduce a second integer-valued matrix $\mathbf{Z}$ with dimensions matching $\mathbf{L}$ to record SNVs. Denote by $\mathbf{z}_c$ the cth column of $\mathbf{Z}$. The vector $\mathbf{z}_c = (z_{1c}, \ldots, z_{Sc})$, $z_{sc} \leqslant l_{sc}$, records the number of alleles, out of the $\mathbf{l}_c$ copies, that bear a mutant sequence that is different from the reference sequence at S loci in subclone c. For example, in Fig. 2, $z_{sc} = 1$ for $s = 2$ and $c = 1$, indicating that one allele bears a variant sequence. By definition, the number of variant alleles z_{sc} in a subclone cannot be larger than the copy number l_{sc} of the subclone, i.e. $z_{sc} \leqslant l_{sc}$. Jointly, the two random-integer vectors $\mathbf{l}_c$ and $\mathbf{z}_c$ describe a subclone and its genetic architecture at the corresponding loci. In addition, the ratio z_{sc}/l_{sc} represents the VAF at locus s in subclone c. To describe the VAFs

at different loci in the sample, we introduce the **w**-matrix that represents the cellular fractions of the subclones. And the sample VAF can be modelled as a mixture of the subclone VAFs, mixing weights being the cellular fractions. Each row $\mathbf{w}_t = (w_{t1}, \ldots, w_{tC})$ represents the cellular fractions of the C subclones in each sample. An important feature of the model proposed is that the number of subclones, modelled as the number of columns in the three matrices, is assumed random and is estimated in posterior inference.

We build on earlier work in Lee *et al.* (2015) and Sengupta *et al.* (2015) which implements inference for TH based on SNV data only, and therefore using a model with **Z** and **w** only, implicitly assuming that all loci are copy number neutral. An important implication of the different modelling approaches is that inference in Lee *et al.* (2015) characterized TH in terms of haplotypes rather than subclones, the difference being that subclones are pairs of haplotypes.

The remainder of the paper is organized as follows: Section 2 describes the Bayesian feature allocation model proposed. Section 3 describes simulation studies. Section 4 reports a data analysis for an in-house data set to illustrate intratumour heterogeneity. The last section concludes with a final discussion.

The program that was used to analyse the data can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Probability model

2.1. Sampling model

Suppose that T ($\geqslant 1$) samples have been sequenced in a next generation sequencing experiment. These samples are assumed to be from the same patient, obtained either at different time points or different spatial locations within the tumour. Suppose that we have collected read mapping data on S loci for the T samples by using bioinformatics pipelines such as the Burrows–Wheeler aligner BWA (Li and Durbin, 2009), SAMtools (Li *et al.*, 2009), and the 'Genome analysis toolkit' GATK (McKenna *et al.*, 2010). For inference on TH, we may restrict the choice of loci to be somatic point mutations. Let **N** and **n** denote $S \times T$ matrices of these counts, N_{st} and n_{st} ($\leqslant N_{st}$) representing the total number of reads and the number of reads that bear a mutated sequence respectively at locus s for tissue sample t, $s = 1, \ldots, S$ and $t = 1, \ldots, T$. Following Klambauer *et al.* (2012), we assume a Poisson sampling model for N_{st}:

$$N_{st} | \phi_t, M_{st} \overset{\text{indep}}{\sim} \text{Poi}(\phi_t M_{st}/2). \tag{1}$$

Here, M_{st} is the sample copy number that represents an average copy number across subclones. We shall formally define and model M_{st} by using subclonal copy numbers **L** next. The factor ϕ_t is the expected number of reads in sample t if there were no CNV, i.e. the copy number equals 2.

Conditionally on N_{st} we assume a binomial sampling model for n_{st}:

$$n_{st} | N_{st}, p_{st} \overset{\text{indep}}{\sim} \text{Bin}(N_{st}, p_{st}). \tag{2}$$

Here p_{st} is the probability of observing a read with a variant sequence. It is interpreted as the expected VAFs in the sample. In the following discussion we shall represent M_{st} and p_{st} in terms of the underlying matrices **L**, **Z** and **w**.

2.2. Prior

2.2.1. Construction of M_{st}

Let C denote the unknown number of subclones in T samples. We first relate M_{st} to CNV at

locus s in sample t. We construct a prior model for M_{st} by using the notion that each sample is composed of a mixture of C subclones. Let w_{tc} denote the proportion of subclone c, $c = 1, \ldots, C$, in sample t and let $l_{sc} \in \{0, 1, 2, \ldots, Q\}$ denote the number of copies at locus s in subclone c where Q is a prespecified maximum number of copies. The event $l_{sc} = 2$ means no CNV at locus s in subclone c, $l_{sc} = 1$ indicates one copy loss and $l_{sc} = 3$ indicates one copy gain. Later we shall extend the prior model with a prior on $\mathbf{L} = (l_{sc})$. But for the moment we condition on l_{sc} and assume that it is known. Then the mean number of copies for sample t can be expressed as the weighted sum of the number of copies over C latent subclones where the weight w_{tc} denotes the cellular fractions of subclone c in sample t. We assume that

$$M_{st} = l_{s0} w_{t0} + \sum_{c=1}^{C} w_{tc} l_{sc}. \tag{3}$$

The sum $\Sigma_{c=1}^{C} w_{tc} l_{sc}$ reflects the key assumption of decomposing the sample copy number into a weighted average of subclonal copy numbers. The first term $l_{s0} w_{t0}$ accounts for noise that can arise from upstream bioinformatics analysis. For example, a small number of short reads may be erroneously mapped to locus s because of ambiguity in the human reference genome or because of base calling error. As such, we use $l_{s0} w_{t0}$ to denote the expected copy number from a hypothetical background subclone to account for potential noise and artefacts in the data, labelled as subclone $c = 0$. Arbitrarily we assume no CNVs at any locations for the background subclone, i.e. $l_{s0} = 2$ for all s.

2.2.2. *Prior on* $\mathbf{L}$

We develop a feature allocation prior for a latent random matrix of copy numbers, $\mathbf{L} = (l_{sc})$, $c = 1, \ldots, C$ and $s = 1, \ldots, S$. We first construct a prior $p(\mathbf{L}|C)$ conditionally on C. Let $\pi_{cq} = p(l_{sc} = q)$, and $\boldsymbol{\pi}_c = (\pi_{c0}, \pi_{c1}, \ldots, \pi_{cQ})$, with $\Sigma_{q=0}^{Q} \pi_{cq} = 1$. As a prior distribution of $\boldsymbol{\pi}_c$, we use a beta–Dirichlet distribution that was developed in Kim *et al.* (2012). Let $\pi_{c2} = p(l_{sc} = 2)$ and $\widetilde{\boldsymbol{\pi}} = (\widetilde{\pi}_{c0}, \widetilde{\pi}_{c1}, \widetilde{\pi}_{c3}, \ldots, \widetilde{\pi}_{cQ})$, with $\widetilde{\pi}_{cq} = \pi_{cq}/(1 - \pi_{c2})$. Conditionally on C, we assume $\pi_{c2} \sim^{\text{IID}} \text{Be}(\beta, \alpha/C)$ and $\widetilde{\boldsymbol{\pi}}_c \sim^{\text{IID}} \text{Dir}(\gamma_0, \gamma_1, \gamma_3, \ldots, \gamma_Q)$. Assuming *a priori* independence among subclones, we write $\boldsymbol{\pi}_c \sim^{\text{IID}} \text{Be-Dir}(\alpha/C, \beta, \gamma_0, \gamma_1, \gamma_3, \ldots, \gamma_Q)$. Similarly to Griffiths and Ghahramani's (2006) constructive definition of the IBP, a limit of the described model $p(\mathbf{L})$ becomes a definition of a CIBP under the following construction (Sengupta, 2013; Sengupta *et al.*, 2015). Let $\beta = 1$ and $C \to \infty$ and drop all columns l_c with all 2s. Next, equivalence classes of Q-nary matrices can be defined by a left ordering similarly to Griffiths and Ghahramani (2006). The marginal limiting distribution of a particular left-ordered equivalent class of $\mathbf{L}$ then defines a CIBP as $C \to \infty$ (Sengupta, 2013). Note that a Dirichlet prior instead of the beta–Dirichlet model would not do. Sengupta (2013) showed that $p(\mathbf{L})$ with a Dirichlet distribution on $\boldsymbol{\pi}_c$ becomes degenerate as $C \to \infty$.

2.2.3. *Prior on* $\mathbf{Z}$ *and construction of* p_{st}

To model the expected VAF of the sample, p_{st}, we construct another feature allocation model linking p_{st} with l_{sc}. We introduce an $S \times C$ matrix $\mathbf{Z}$ whose entries $z_{sc} \in \{0, \ldots, l_{sc}\}$ denote the number $z_{sc} \leqslant l_{sc}$ of alleles bearing a variant sequence among the total of l_{sc} copies at locus s in subclone c. Assume that

$$z_{sc} \,|\, l_{sc} \sim \text{Unif}(0, 1, \ldots, l_{sc}), \tag{4}$$

where Unif$(\cdot)$ indicates a (discrete) uniform distribution, implying in particular for $z_{sc}=0$ when $l_{sc}=0$.

Next, we write p_{st} in equation (2) as a ratio between the expected number of variant alleles and the expected sample copy number. In particular, the expected number of variant alleles is a weighted sum of subclonal variant allele counts over $C+1$ latent subclones including the background subclone, and the expected VAF is defined as

$$p_{st}=\frac{p_0 z_{s0} w_{t0}+\sum_{c=1}^{C} w_{tc} z_{sc}}{M_{st}}. \tag{5}$$

Similarly to the previous argument for assumption (3), the term $\Sigma_{c=1}^{C} w_{tc} z_{sc}$ in equation (5) reflects the assumption that the sample level variant allele count is a weighted average of subclonal variant allele counts. The first term of the numerator, $p_0 z_{s0} w_{t0}$, describes a hypothetical background subclone that accounts for experimental noise such as base calling errors. We let $z_{s0}=2$ for all s denote the number of variant alleles in the background subclone. We add a global parameter p_0 to account for artefacts and experimental noise that would produce variant reads even if no subclones were to have variant alleles. Since p_0 does not depend on s or t, it can be estimated by pooling data from all loci and samples and does not affect the identifiability of the model. Note that model (5) for p_{st} is different from those in Lee *et al.* (2015) and Sengupta *et al.* (2015) in that it accounts for the mean number of copies of locus s in sample t. We consider $p_0 \sim \text{Be}(a_{00}, b_{00})$ with $a_{00} \ll b_{00}$ to inform a small p_0 value *a priori*.

2.2.4. *Prior for* $\mathbf{w}$

Next, we introduce a prior distribution for the weights w_{tc} in equations (3) and (5). The subclones are common for all tumour samples, but the relative weights w_{tc} vary across tumour samples. We assume independent Dirichlet priors as follows. Let θ_{tc} denote an (unscaled) abundance level of subclone c in tissue sample t. We assume that $\theta_{tc}|C \sim^{\text{IID}} \text{gamma}(d,1)$ for $c=1,\ldots,C$ and $\theta_{t0} \sim^{\text{IID}} \text{gamma}(d_0,1)$. We then define $w_{tc}=\theta_{tc}/\Sigma_{c'=0}^{C}\theta_{tc'}$, as the relative weight of subclone c in sample t. This is equivalent to $\mathbf{w}_t|C \sim^{\text{IID}} \text{Dir}(d_0,d,\ldots,d)$ for $t=1,\ldots,T$. Using $d_0<d$ implies that the background subclone takes a smaller proportion in a sample.

Finally, we complete the model construction with a prior on the unknown number of latent subclones C. We use a geometric distribution, $C \sim \text{Geom}(r)$ where $E(C)=1/r$. Conditionally on C, the two latent matrices, $\mathbf{L}$ and $\mathbf{Z}$ describe C latent tumour subclones. Joint inference on C, $\mathbf{L}$, $\mathbf{Z}$ and $\mathbf{w}_t$ explains TH.

The construction of the subclones, including the number of subclones, C, the subclonal copy number l_{sc} and the number z_{sc} of mutant copies are latent. The subclones are not directly observed. They are defined only as the components of the assumed mixture that gives rise to the observed CNV and VAFs. The key terms, $\Sigma_{c=1}^{C} w_{tc} l_{sc}$ in equation (3) and $\Sigma_{c=1}^{C} w_{tc} z_{tc}/M_{st}$ in equation (5) allow us to infer subclones indirectly by explaining M_{st} and p_{st} as arising from sample t being composed of a mix of hypothetical subclones.

Lastly, we take account of different average read counts in T samples through ϕ_t, which represents the expected read count with two copies in sample t. We assume $\phi_t \sim^{\text{indep}} \text{gamma}(a_t,b_t)$ where $E(\phi_t)=a_t/b_t$.

2.3. Posterior simulation

Let $\mathbf{x}=(\mathbf{L},\mathbf{Z},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\pi},p_0)$ denote all unknown parameters, where $\boldsymbol{\theta}=\{\theta_{tc}\}$ and $\boldsymbol{\pi}=\{\pi_{cq}\}$. We implement inference via posterior MCMC simulation, i.e. by generating a Monte Carlo sample

of $\mathbf{x}_i \sim p(\mathbf{x}|\mathbf{n}, \mathbf{N})$, $i = 1, \ldots, I$. MCMC posterior simulation proceeds by repeatedly using transition probabilities that update a subset of parameters at a time. See, for example Brooks *et al.* (2011) for a recent review.

For fixed C such MCMC simulation is straightforward. Gibbs sampling transition probabilities are used to update l_{sc}, z_{sc}, π_{cq} and ϕ_t and Metropolis–Hastings transition probabilities are used to update $\boldsymbol{\theta}$ and p_0. The construction of transition probabilities that involves a change of C is more difficult, since the dimension of $\mathbf{L}$, $\mathbf{Z}$, $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ changes as C varies. We use the approach that was proposed in Lee *et al.* (2015) for posterior simulation in a similar model. We split the data into a small training set $(\mathbf{n}', \mathbf{N}')$ with $n'_{st} = b_{st} n_{st}$ and $N'_{st} = b_{st} N_{st}$, and a test data set $(\mathbf{n}'', \mathbf{N}'')$ with $n''_{st} = (1 - b_{st}) n_{st}$ and $N''_{st} = (1 - b_{st}) N_{st}$. In the implementation we use $b_{st} \sim \mathrm{Be}(25, 975)$ for the analyses in the following sections. We found that using a random b_{st} worked better than a fixed fraction b across all samples and loci and that a moderate change in the beta distribution generating b_{st} did not significantly affect the resulting posterior inference. Let $p_1(\mathbf{x}|C) = p(\mathbf{x}|\mathbf{N}', \mathbf{n}', C)$ denote the posterior distribution conditional on a fixed C, under the training sample. We use p_1 in two instances. First, we replace the original prior $p(\mathbf{x}|C)$ by $p_1(\mathbf{x}|C)$ and, second, we use $p_1(\cdot)$ as proposal distribution $q(\tilde{\mathbf{x}}|\tilde{C}) = p_1(\tilde{\mathbf{x}}|\tilde{C})$ in a reversible jump style transition probability where $\tilde{C}$ is a proposed value of C. The test data are then used to evaluate the acceptance probability. The critical advantage of using the same $p_1(\cdot)$ as prior and proposal distribution is that the normalization constant cancels out in the Metropolis–Hastings acceptance probability. Note that this MCMC method does not change inference on $\mathbf{x}$, but inference on C becomes an approximation of $p(C|\mathbf{N}, \mathbf{n})$ with the original prior of $\mathbf{x}$. The details of the MCMC posterior simulation are described in the on-line supplementary material. We note that the computation takes approximately 26 min on 3.33-GHz central processor unit for the lung cancer data set in Section 4 for 10000 iterations. We examined mixing and convergence in MCMC chains for the simulation studies and the lung cancer data analysis that are presented in later sections. After discarding 6000 iterations for burn-in, the chains moved well and we found no evidence for lack of convergence.

We summarize the joint posterior by factorizing it as

$$\begin{aligned} p(C, \mathbf{L}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\phi}, \mathbf{w}, p_0|\mathbf{n}, \mathbf{N}) = {} & p(C|\mathbf{n}, \mathbf{N})\, p(\mathbf{L}|\mathbf{n}, \mathbf{N}, C)\, p(\mathbf{Z}, \boldsymbol{\pi}|\mathbf{n}, \mathbf{N}, C, \mathbf{L}) \\ & \times p(\mathbf{w}|\mathbf{L}, \mathbf{Z}, \mathbf{n}, C)\, p(\boldsymbol{\phi}, p_0|\mathbf{n}, \mathbf{N}, C). \end{aligned}$$

Using the posterior Monte Carlo sample we (approximately) evaluate the marginal posterior $p(C|\mathbf{n}, \mathbf{N})$ and determine the maximum *a posteriori* estimate C^*. We follow Lee *et al.* (2015) to define $\mathbf{L}^*$ conditionally on C^*. For any two $S \times C^*$ matrices $\mathbf{L}$ and $\mathbf{L}'$, $1 \leqslant c, c' \leqslant C^*$, let $\mathcal{D}_{cc'}(\mathbf{L}, \mathbf{L}') = \Sigma_{s=1}^{S} |l_{sc} - l'_{sc'}|$. We then define a distance $d(\mathbf{L}, \mathbf{L}') = \min_{\boldsymbol{\sigma}} \Sigma_{c=1}^{C^*} \mathcal{D}_{c,\sigma_c}(\mathbf{L}, \mathbf{L}')$, where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_C)$ is a permutation of $\{1, \ldots, C^*\}$ and the minimum is over all possible permutations. Let

$$\mathbf{L}^* = \arg\min_{\mathbf{L}'} \int d(\mathbf{L}, \mathbf{L}')\, \mathrm{d}\, p(\mathbf{L}|\mathbf{n}, \mathbf{N}, C^*) \approx \arg\min_{\mathbf{L}'} \sum_{i=1}^{I} d(\mathbf{L}^{(i)}, \mathbf{L}'),$$

for a posterior Monte Carlo sample $\{\mathbf{L}^{(i)}, i = 1, \ldots, I\}$, i.e. we report the matrix $\mathbf{L}^*$ that minimizes the posterior expected distance $d(\mathbf{L}, \mathbf{L}^*)$. We report posterior point estimates $\mathbf{Z}^*$, $\mathbf{w}^*$ and $\boldsymbol{\pi}^*$ conditional on C^* and $\mathbf{L}^*$. Finally, we report $\boldsymbol{\phi}^*$ and p_0^* as the posterior mean of $\boldsymbol{\phi}$ and p_0 conditional on C^*. For the lung cancer data in Section 4, searching $\mathbf{L}^*$ from 4918 Monte Carlo samples of $\mathbf{L}$ conditional on $C^* = 2$ takes approximately 10 min on 3.33-GHz central processor unit. The algorithm works fairly fast for a moderate value of the chosen C.

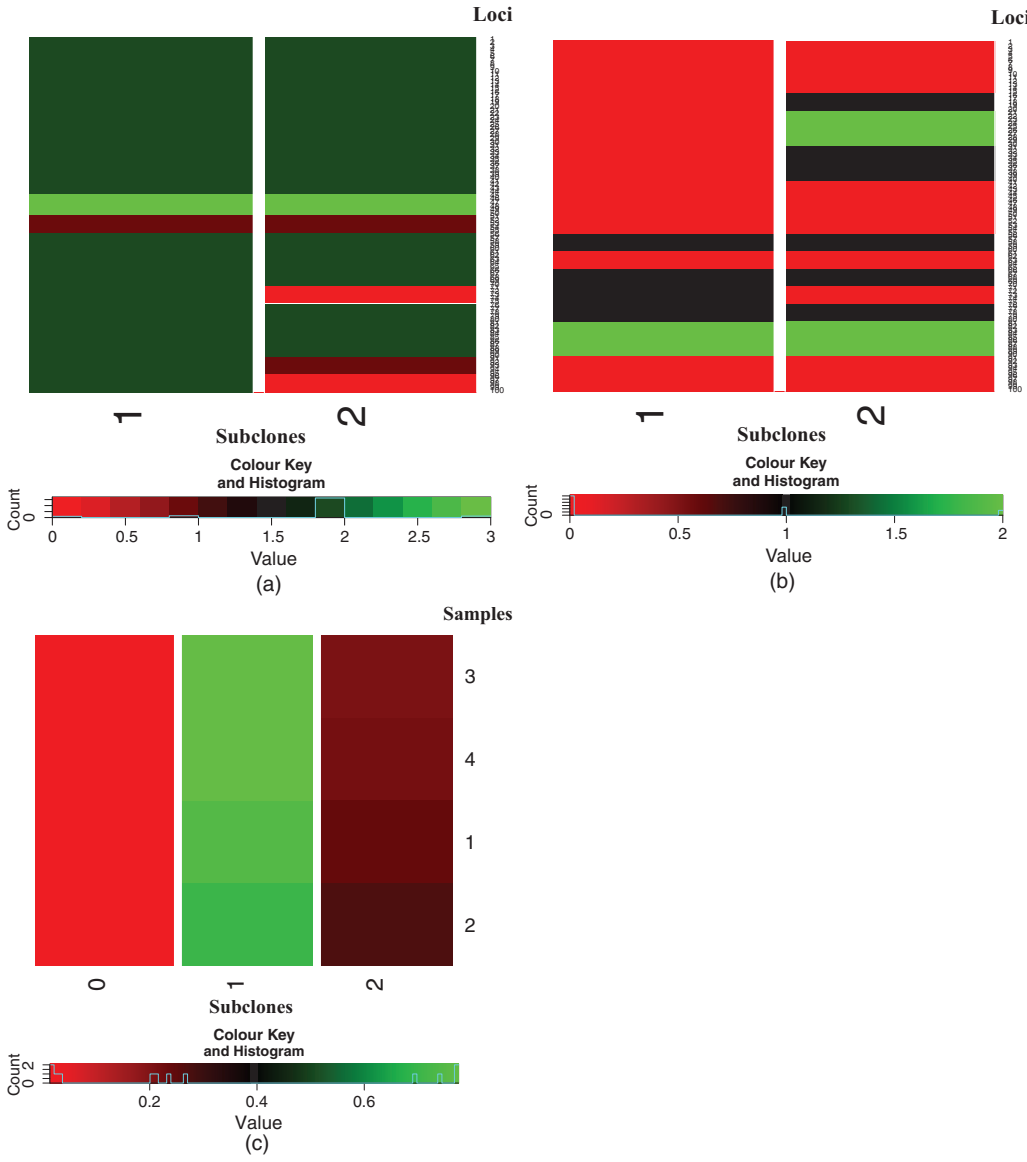

Fig. 2. Simulation truth: (a) $\mathbf{L}^{\text{TRUE}}$; (b) $\mathbf{Z}^{\text{TRUE}}$; (c) $\mathbf{w}^{\text{TRUE}}$

3. Simulation

3.1. Simulation truth

We assess the proposed model via simulation. We generate hypothetical read counts for a set of $S = 100$ loci in $T = 4$ hypothetical samples. In the simulation truth, we assume two latent subclones ($C^{\text{TRUE}} = 2$) as well as a background subclone ($c = 0$) with all SNVs bearing variant sequences with two copies. We use $Q = 3$. The simulation truth $\mathbf{L}^{\text{TRUE}}$ is shown in Fig. 2(a) where green in the panels indicates a copy gain ($l_{sc} = 3$) and red indicates two-copy loss ($l_{sc} = 0$). Fig. 2(b) shows the simulation truth $\mathbf{Z}^{\text{TRUE}}$. Similarly to $\mathbf{L}^{\text{TRUE}}$, green indicates three copies

with SNV and red indicates zero copies with SNV. We generate $\phi_t^{\mathrm{TRUE}} \sim^{\mathrm{IID}} \mathrm{gamma}(600, 3)$, $t = 1, \ldots, 4$, and then generate $\mathbf{w}^{\mathrm{TRUE}} \sim^{\mathrm{IID}} \mathrm{Dir}(0.4, 30.0, 10.0)$. The weights $\mathbf{w}^{\mathrm{TRUE}}$ are shown in Fig. 2(c). Similarly to the other heat maps, green in Fig. 2(c) represents high abundance of a subclone in a sample and red shows low abundance. On average, subclone 1 takes w_{tc} close to 0.75 for all the samples, with little heterogeneity across samples. Using the assumed $\mathbf{L}^{\mathrm{TRUE}}$, $\mathbf{Z}^{\mathrm{TRUE}}$ and $\mathbf{w}^{\mathrm{TRUE}}$ and letting $p_0^{\mathrm{TRUE}} = 0.05$, we generate $N_{st} \sim \mathrm{Poi}(\phi_t^{\mathrm{TRUE}} M_{st}^{\mathrm{TRUE}}/2)$ and $n_{st} \sim \mathrm{Bin}(N_{st}, p_{st}^{\mathrm{TRUE}})$.

3.2. Posterior inference

To fit the model proposed, we fix the hyperparameters as $r = 0.2$, $\alpha = 2$, $\gamma_q = 0.5$ for $q = 0, 1, 3$ $(= Q)$, $d_0 = 0.5$, $d = 1$, $a_{00} = 0.3$ and $b_{00} = 5$. For the prior on ϕ_t, we let $b = 3$ and specify a by setting the median of the observed N_{st} to be the prior mean. For each value of C, we initialized $\mathbf{Z}$ by using the observed sample proportions and $\mathbf{L}$ by using the initial $\mathbf{Z}$. We generated initial values for $\boldsymbol{\theta}_{tc}$ and p_0 by prior draws. We generated $b_{st} \sim^{\mathrm{IID}} \mathrm{Be}(25, 975)$ to construct the training set and ran the MCMC simulation over 16000 iterations, discarding the first 6000 iterations as initial burn-in.

Fig. 3(a) shows $p(C|\mathbf{n}, \mathbf{N})$. The broken vertical line marks the simulation truth $C^{\mathrm{TRUE}} = 2$. The posterior mode $C^* = 2$ recovers the truth. Figs 3(d)–3(f) show the posterior point estimates $\mathbf{L}^*$, $\mathbf{Z}^*$ and $\mathbf{w}^*$. Compared with the simulation truth in Fig. 2, the posterior estimate recovers subclone 1 with high accuracy, but $\mathbf{l}_c^*$ for subclone $c = 2$ shows some discrepancies with the simulation truth. This is due to small w_{tc}^{TRUE}, $c = 2$, across all four samples (the last column in Fig. 2(c)). The discrepancy between $\mathbf{l}_2^*$ and $\mathbf{l}_2^{\mathrm{TRUE}}$ is related to the misspecification of $\mathbf{z}_c^*$ under $c = 2$. Conditionally on C^*, we computed $\hat{M}_{st}$ and $\hat{p}_{st}$ and compared them with the true values. Figs 3(b) and 3(c) show a good fit under the model for a majority of loci and samples although the histograms include a small pocket of differences between the true values and their estimates on the right-hand tail, also possibly due to the misspecification of $\mathbf{l}_2$ and $\mathbf{z}_2$. This simulation study illustrates that the model proposed reasonably recovers the simulation truth even with a small number of samples when the underlying structure is not complex.

3.3. Comparison with PyClone

For comparison, we implemented PyClone (Roth *et al.*, 2014) with the same simulated data. We select PyClone for the comparison, since it considers copy number changes and point mutations similarly to the method proposed. We let the normal copy number, the minor parental copy number and the major parental copy number be 2, 0 and 3 respectively at each locus. Although accounting for copy number changes, PyClone estimates the variant allelic prevalence (the fraction of clonal population having a mutation) at a locus in a sample. The interpretation of variant allelic prevalences, which is referred to as 'cellular prevalences' in PyClone, is similar to that of p_{st} in the model proposed. PyClone uses a Dirichlet process model to identify a (non-overlapping) clustering of the loci on the basis of their cellular prevalences. Cellular prevalences over loci and samples may vary but the clustering of loci is shared by samples. Fig. 4(a) shows posterior estimates of the cellular prevalences (by colour) and mutational clustering (by separations with white horizontal lines) under PyClone. Fig. 4(b) shows a heat map of p_{st}^{TRUE}. The loci (rows) of the two heat maps are rearranged in the same order for easy comparison. By comparing the two heat maps, the cellular prevalence estimates under PyClone are close to p_{st}^{TRUE} and lead to a reasonable estimate of a clustering of the loci. However, PyClone does not attempt to construct a description of subclones with genomic variants.

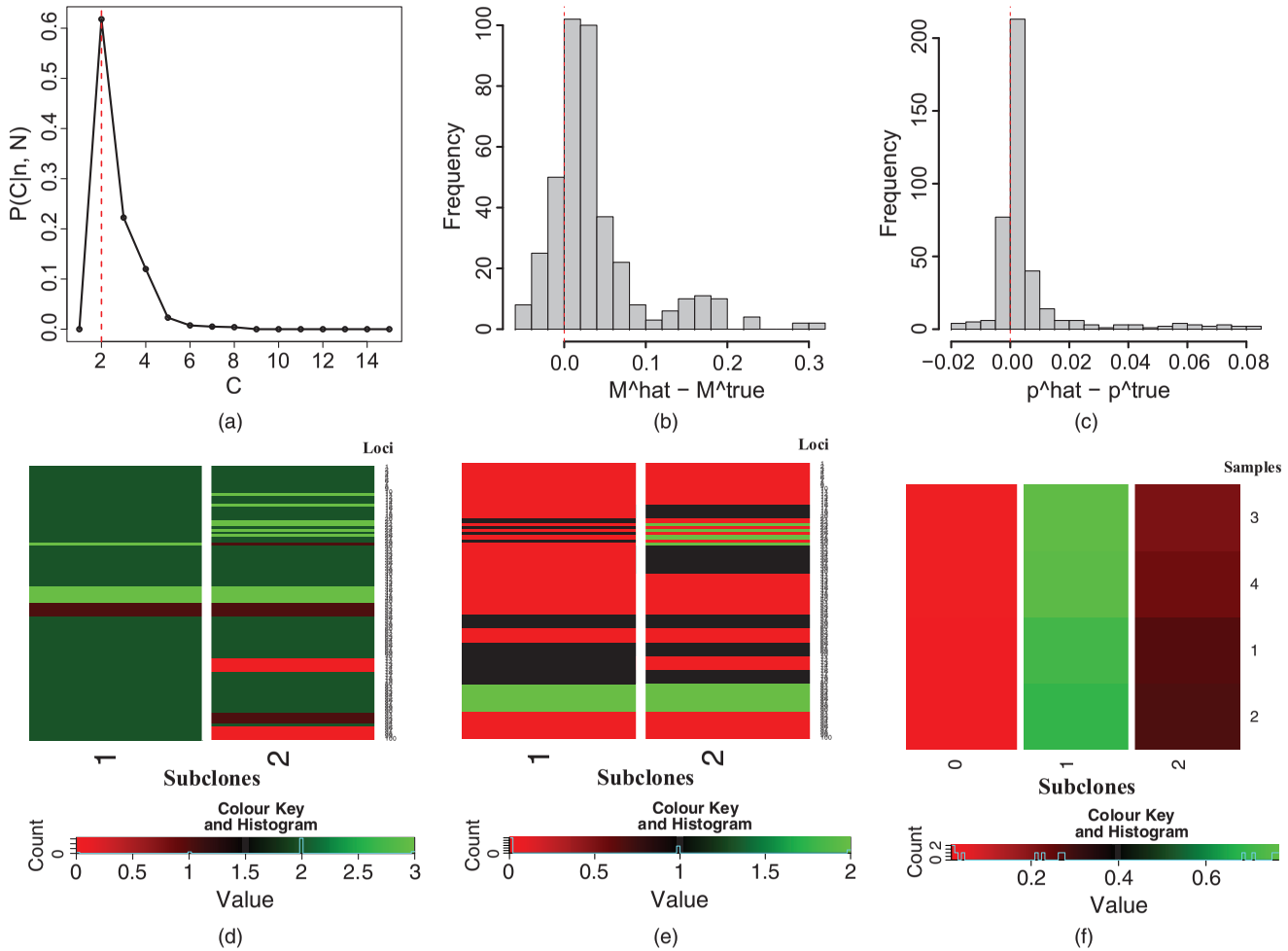

Fig. 3. Posterior inference for simulation: (a) $p(C|\mathbf{n},\mathbf{N})$; (b) $\hat{M}_{st} - M_{st}^{\text{TRUE}}$; (c) $\hat{p}_{st} - p_{st}^{\text{TRUE}}$; (d) $\mathbf{L}^*$; (e) $\mathbf{Z}^*$; (f) $\mathbf{w}^*$

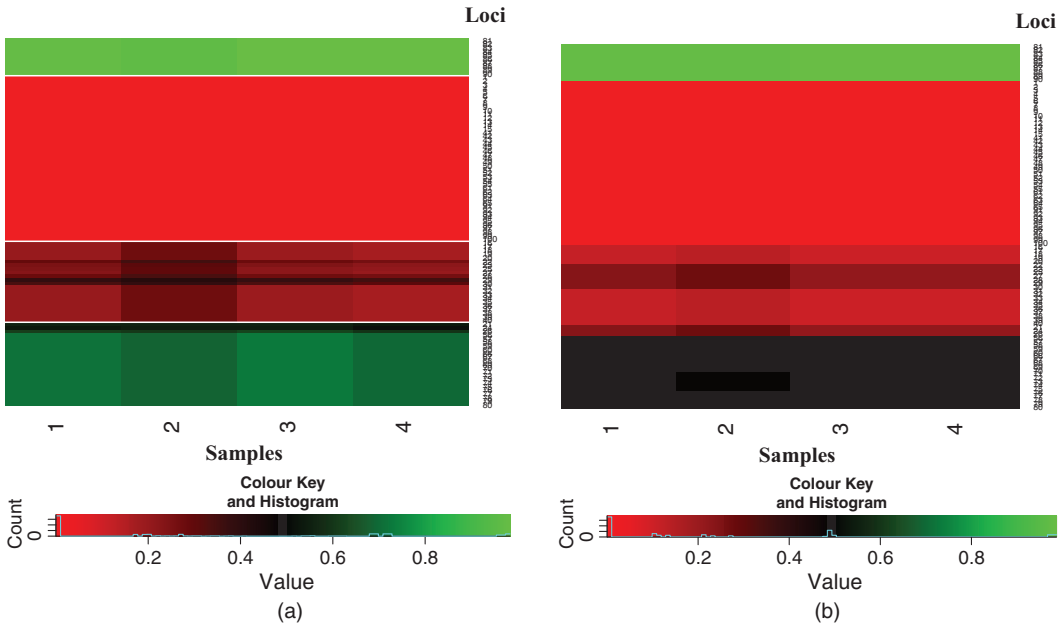

Fig. 4. Heat maps of (a) estimated cellular prevalences from PyClone and (b) p_{st}^{TRUE} for simulation

3.4. *Sensitivity analysis*

We studied how posterior inference changes over different specification of Q. We use $Q = 3, 4, 5$ to analyse the same data simulated with $Q = 3$. A negligible change in the marginal posterior distribution of C across the values of Q is observed. Under all Q, the maximum *a posteriori* estimate for C correctly recovers C^{TRUE}. The comparison of $\mathbf{L}^*$, $\mathbf{Z}^*$ and $\mathbf{w}^*$ with the simulation truth shows that any $Q \geqslant 3$ recovers the true subclonal structure with great accuracy. In particular, subclone 1 that has a large w^{TRUE} for all the samples is well reconstructed compared with subclone 2. See the on-line supplementary materials for figures that summarize the comparison.

We also studied sensitivity with respect to the choice of α in the beta–Dirichlet prior. We considered $\alpha = 1, 5, 10, 20$. Although the marginal posterior distribution of C slightly changes, the posterior mode recovers C^{TRUE} across all choices. See the on-line supplementary materials for detailed discussion on posterior inferences with the various values of α.

The model proposed is also examined with an additional simulation study assuming a more complicated subclonal structure and heterogeneous samples. The results are reported in the on-line supplementary material.

4. Lung cancer data

In an in-house experiment at NorthShore University HealthSystem, we record whole-exome sequencing for $T = 4$ surgically dissected tumour samples taken from the same lung cancer patient. The four samples were spatially close to each other. As an exploratory experiment, our clinical collaborators want to find out whether spatially proximal tumour samples are genetically homogeneous. For this, we extracted genomic DNA from each tissue and constructed an exome library from these DNA by using Agilent SureSelect capture probes. The exome library was then sequenced in paired end fashion on an Illumina HiSeq 2000 platform. About 60 million reads—

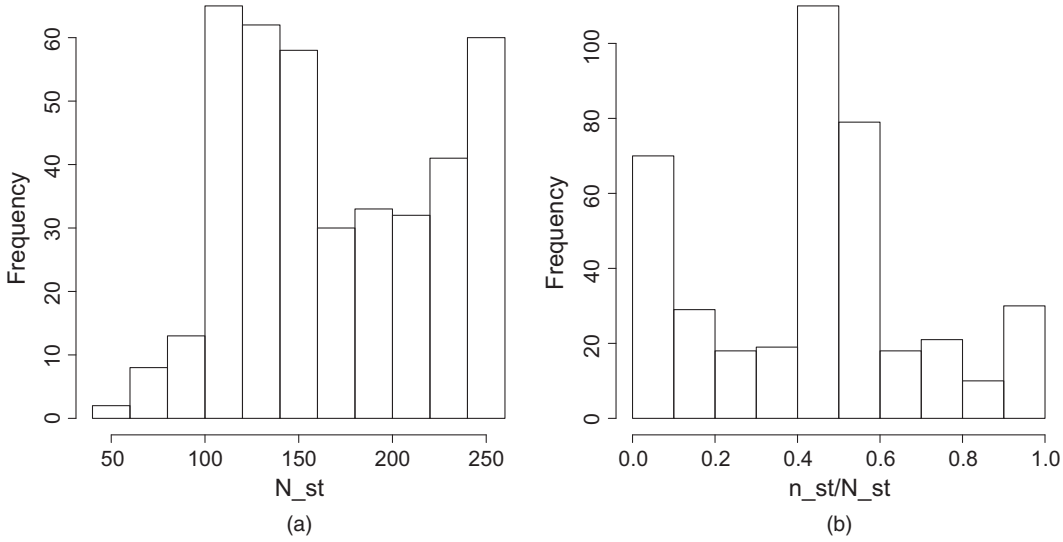

Fig. 5. Histograms of the lung cancer data set: (a) N_{st}; (b) n_{st}/N_{st}

each 100 bases long—were obtained. Since the SureSelect exome was about 50 megabases, raw (premapping) coverage was about 120-fold. We then mapped the reads to the human genome (version HG19) (Church *et al.*, 2011) by using BWA (Li and Durbin, 2009) and called variants by using GATK (McKenna *et al.*, 2010). Post mapping, the mean coverage of the samples was between 60- and 70-fold.

A total of nearly 115000 SNVs and small insertions and deletions were called within the exome co-ordinates. We restricted our attention to SNVs that

(a) make a difference to the protein translated from the gene (i.e. non-synonymous) and
(b) that exhibit significant coverage in all samples with n_{st}/N_{st} not being too close to 0 or 1; and
(c) we used expert judgement to exclude some more loci.

The filter rules described leave in the end $S = 101$ SNVs for the four intratumour samples. Fig. 5 shows the histograms of the total number of reads and the empirical read ratios, N_{st} and n_{st}/N_{st}.

We used hyperparameters that were similar to those in the simulation studies. Fig. 6 summarizes posterior inference under the model proposed. Fig. 6(a) shows $C^* = 2$, i.e. two estimated subclones. Using posterior samples with $C = C^*$, we computed $\hat{N}_{st}$ and $\hat{p}_{st}$ and compared them with the observed data. The differences are centred at 0, implying a good fit to the data. Conditionally on $C^* = 2$, we found $\mathbf{L}^*$, $\mathbf{Z}^*$ and $\mathbf{w}^*$. The loci in $\mathbf{L}^*$ and $\mathbf{Z}^*$ are rearranged in the same order for better illustration. From Fig. 5(a) we note that many positions have large numbers of reads: over 200 reads. This is reflected in $\mathbf{L}^*$ which estimates three copies at many positions. The estimated weights $\mathbf{w}^*$ in Fig. 6(f) show a great similarity across the four samples. This lack of heterogeneity across samples suggests that, for this tumour, spatial proximity is implicative of genetic homogeneity. Finally, we investigated sensitivity to hyperparameters by considering posterior inference under various specifications for Q, α or b_{st}. Only minor changes in the reported inference are observed.

Again, for comparison we implemented inference by using PyClone (Roth *et al.*, 2014) for

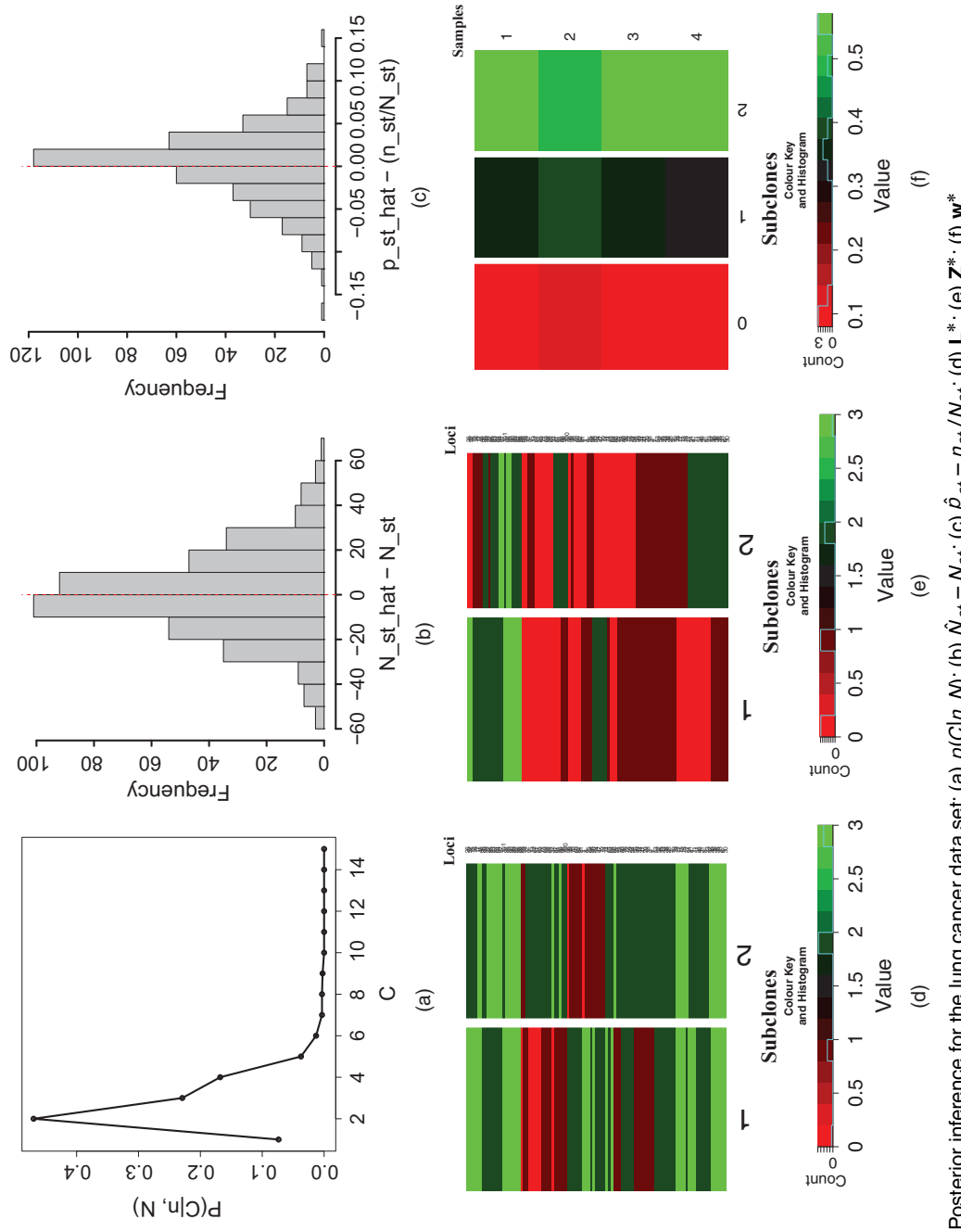

Fig. 6. Posterior inference for the lung cancer data set: (a) $p(C|n, N)$; (b) $\hat{N}_{st} - N_{st}$; (c) $\hat{\rho}_{st} - n_{st}/N_{st}$; (d) $\mathbf{L}^*$; (e) $\mathbf{Z}^*$; (f) $\mathbf{w}^*$

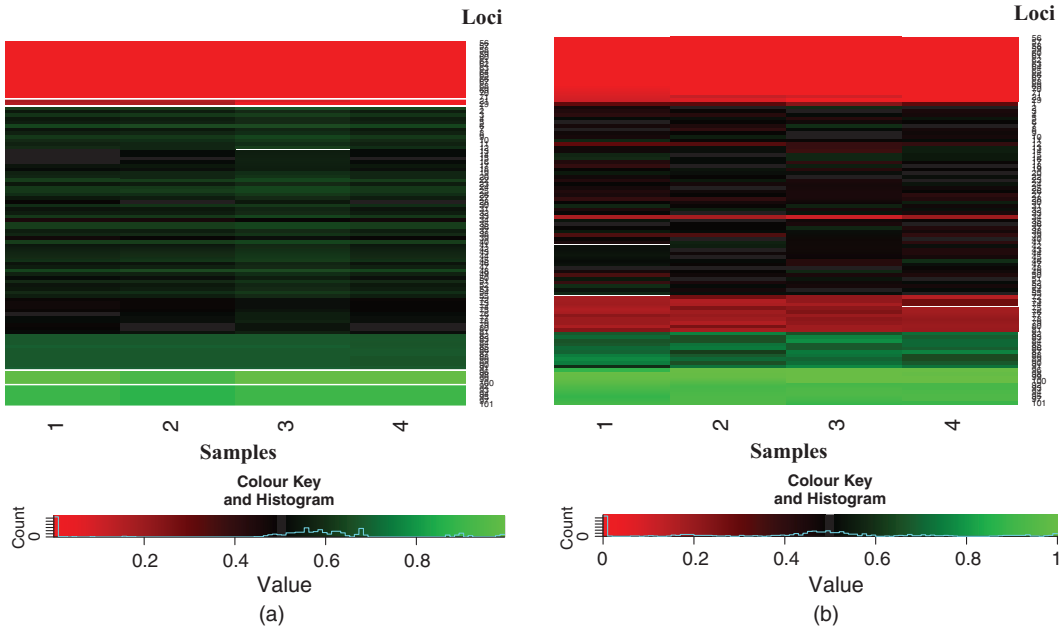

Fig. 7. Heat maps of (a) estimated cellular prevalences from PyClone and (b) n_{st}/N_{st} for the lung cancer data set

the same lung cancer data. Posterior estimates of prevalence and the estimated clustering of the loci are shown in Fig. 7(a). Inference identified five clusters of loci. The mean prevalences within a locus cluster are similar across samples, which is similar to $\mathbf{w}^*$ in Fig. 6(f). Fig. 7(b) is a heat map of fractions of reads bearing mutation for each locus and sample. Again, PyClone provides a reasonable estimate of a partition of loci based on the empirical fractions but does not provide (and is not meant to provide) inference on subclonal populations.

5. Conclusions

The approach proposed infers subclonal DNA copy numbers, subclonal variant allele counts and cellular fractions in a biological sample. By jointly modelling CNVs and SNVs, we provide the desired description of TH based on DNA variations in both, sequence and structure. Such inference will significantly impact downstream treatment of individual tumours, ultimately allowing personalized prognosis. For example, a tumour with large proportions of cells bearing somatic mutations on tumour suppressor genes should be treated differently from another tumour with a small proportion of such cells. In addition, metastatic or recurrent tumours may have very different compositions of cellular genomes and should be treated differently. Inference on TH can be exploited for improved treatment strategies for relapsed cancer patients and can spark significant improvement in cancer treatment in practice.

Various extensions are possible for the present model. For example, sometimes additional sources of information on CNVs such as a single-nucleotide polymorphism array may be available. We then extend the model to incorporate this information into the modelling of $\mathbf{L}$. Another meaningful extension is to cluster patients on the basis of the imputed TH, i.e. we link a random partition and a feature allocation model. This extension may help clinicians to assign different treatment strategies and may be a natural basis of adaptive clinical trial designs.

Inference for TH is a critical gap in the current literature. The ability to break down a tumour precisely into a set of subclones with distinct genetics would provide the opportunity for breakthroughs in cancer treatment by facilitating individualized treatment of the tumour that exploits TH. It would open the door for a cocktail type of combinational treatments, with each treatment targeting a specific tumour subclone on the basis of its genetic characteristics. We believe that the model proposed may provide an integrated view on subclones to explain TH that remains a mystery to scientists so far.

Acknowledgement

Yuan Ji and Peter Müller's research is partially supported by National Institutes of Health grant R01 CA132897.

References

Bedard, P. L., Hansen, A. R., Ratain, M. J. and Siu, L. L. (2013) Tumour heterogeneity in the clinic. *Nature*, **501**, 355–364.

Biesecker, L. G. and Spinner, N. B. (2013) A genomic view of mosaicism and human disease. *Nat. Rev. Genet.*, **14**, 307–320.

Broderick, T., Jordan, M. I. and Pitman, J. (2013) Cluster and feature modeling from combinatorial stochastic processes. *Statist. Sci.*, **28**, 289–312.

Brooks, S., Gelman, A., Jones, G. and Meng, X.-L. (2011) *Handbook of Markov Chain Monte Carlo*. Boca Raton: CRC Press.

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W. M., Ritchie, G. R., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., Matthews, L., Whitehead, S., Chow, W., Torrance, J., Dunn, M., Harden, G., Threadgold, G., Wood, J., Collins, J., Heath, P., Griffiths, G., Pelan, S., Grafham, D., Eichler, E. E., Weinstock, G., Mardis, E. R., Wilson, R. K., Howe, K., Flicek, P. and Hubbard, T. (2011) Modernizing reference genome assemblies. *PLOS Biol.*, **9**, no. 7, article e1001091.

De, S. (2011) Somatic mosaicism in healthy human tissues. *Trends Genet.*, **27**, 217–223.

Deshwar, A., Vembu, S., Yung, C. K., Jang, G. H., Stein, L. and Morris, Q. (2015) PhyloWGS: Reconstructing subclonal composition and evolution from whole genome sequencing of tumors. *Genome Biol.*, **16**, article 35.

Ding, L., Ley, T. J., Larson, D. E., Miller, C. A., Koboldt, D. C., Welch, J. S., Ritchey, J. K., Young, M. A., Lamprecht, T., McLellan, M. D., McMichael, J. F., Wallis, J. W., Lu, C., Shen, D., Harris, C. C., Dooling, D. J., Fulton, R. S., Fulton, L. L., Chen, K., Schmidt, H., Kalicki-Veizer, J., Magrini, V. J., Cook, L., McGrath, S. D., Vickery, T. L., Wendl, M. C., Heath, S., Watson, M. A., Link, D. C., Tomasson, M. H., Shannon, W. D., Payton, J. E., Kulkarni, S., Westervelt, P., Walter, M. J., Graubert, T. A., Mardis, E. R., Wilson, R. K. and DiPersio, J. F. (2012) Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, **481**, 506–510.

Fischer, A., Vázquez-García, I., Illingworth, C. J. and Mustonen, V. (2014) High-definition reconstruction of clonal composition in cancer. *Cell Rep.*, **7**, 1740–1752.

Frank, S. A. and Nowak, M. A. (2003) Cell biology: developmental predisposition to cancer. *Nature*, **422**, 494.

Frank, S. A. and Nowak, M. A. (2004) Problems of somatic mutation and cancer. *Bioessays*, **26**, 291–299.

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

Griffiths, T. and Ghahramani, Z. (2006) Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, pp. 475–482.

Jiao, W., Vembu, S., Deshwar, A., Stein, L. and Morris, Q. (2014) Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinform.*, **15**, article 35.

Kim, Y., James, L. and Weissbach, R. (2012) Bayesian analysis of multistate event history data: beta-dirichlet process prior. *Biometrika*, **99**, 127–140.

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.-A., Mitterecker, A., Bodenhofer, U. and Hochreiter, S. (2012) cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.*, **40**, no. 9, article e69.

Lee, J., Müller, P., Gulukota, K. and Ji, Y. (2015) A bayesian feature allocation model for tumor heterogeneity. *Ann. Appl. Statist.*, **9**, 621–639.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, B. and Li, J. Z. (2014) A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biol.*, **15**, article 473.
McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. and DePristo, M. A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
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., Ellis, M. J., Schierding, W., DiPersio, J. F., Ley, T. J., Mardis, E. R., Wilson, R. K. and Ding, L. (2014) Sciclone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLOS Computnl Biol.*, **10**, no. 8, article e1003665.
Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W. R., Hicks, J. and Wigler, M. (2011) Tumour evolution inferred by single-cell sequencing. *Nature*, **472**, 90–94.
Oesper, L., Mahmoody, A. and Raphael, B. J. (2013) THetA: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome Biol.*, **14**, article R80.
Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A. and Shah, S. P. (2014) PyClone: statistical inference of clonal population structure in cancer. *Nat. Meth.*, **11**, 396–398.
Russnes, H. G., Navin, N., Hicks, J. and Borresen-Dale, A.-L. (2011) Insight into the heterogeneity of breast cancer through next-generation sequencing. *J. Clin. Investgn*, **121**, 3810–3818.
Sengupta, S. (2013) Two models involving bayesian nonparametric techniques. *Phd Thesis*. University of Florida, Gainesville.
Sengupta, S., Gulukota, K., Lee, J., Müller, P. and Ji, Y. (2015) Bayclone: Bayesian nonparametric inference of tumor subclones using ngs data. *Proc. Pacif. Symp. Biocomput.*, **20**, 467–478.
Strino, F., Parisi, F., Micsinai, M. and Kluger, Y. (2013) Trap: a tree approach for fingerprinting subclonal tumor composition. *Nucleic Acids Res.*, **41**, no. 17, article e165.
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 Computnl Biol.*, **10**, no. 7, article e1003703.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary materials: Bayesian inference for intra-tumor heterogeneity in mutations and copy number variation'.