A BAYESIAN NONPARAMETRIC MODEL FOR RECONSTRUCTING TUMOR SUBCLONES BASED ON MUTATION PAIRS

SUBHAJIT SENGUPTA[1*], TIANJIAN ZHOU[2*], PETER MÜLLER[3], YUAN JI[1,4†]

[1]*Program for Computational Genomics and Medicine, NorthShore University HealthSystem;* [2]*Department of Statistics and Data Sciences, The University of Texas at Austin;* [3]*Department of Mathematics, The University of Texas at Austin;* [4]*Department of Public Health Sciences, The University of Chicago*

We present a feature allocation model to reconstruct tumor subclones based on mutation pairs. The key innovation lies in the use of a pair of proximal single nucleotide variants (SNVs) for the subclone reconstruction as opposed to a single SNV. Using the categorical extension of the Indian buffet process (cIBP) we define the subclones as a vector of categorical matrices corresponding to a set of mutation pairs. Through Bayesian inference we report posterior probabilities of the number, genotypes and population frequencies of subclones in one or more tumor sample. We demonstrate the proposed methods using simulated and real-world data. A free software package is available at http://www.compgenome.org/pairclone.

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

1. Introduction

1.1. *Background*

With the recent development of next-generation sequencing (NGS) technology, whole-genome or whole-exome sequencing has been used to interrogate genetic landscape of tumors within and across different patients. Using single nucleotide variants (SNVs), NGS data can reveal whether a tumor sample is composed of cell subpopulations, i.e., subclones that contain somatic mutations.[1–6] In essence, the main problem of subclone reconstruction is to identify more than two haploid genomes in a tumor sample. Since humans are diploid, a homogeneous cell population can only harbor two distinct haploid genomes, or else the cell population must be heterogeneous and contain at least two different subclones with different genomes. In NGS data, short reads are mapped to each SNV locus. Compared to the reference nucleotide base on the locus, some short reads may harbor the same reference base while others may bear a variant base. The latter are called variant reads and the proportion of variant reads among all the reads mapped to the SNV is called the observed variant allele fraction (VAF). If all the cells in a tumor sample share the same genome, i.e., they are genetically homogeneous, the VAFs must be close to 0, 0.5, or 1, reflecting the three possible genotypes at a single locus – AA, AB, or BB. For example, when all the cells in the tumor bear the heterozygous AB genotype, roughly half of the reads will harbor A and the other half B. Therefore, the observed VAF should be close to 0.5. Homozygous alleles should give rise to observed VAFs close to 0 or 1. When the VAF at the SNV is neither of 0, 0.5, or 1, the cellular genomes might

*have equal contributions.

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

be heterogeneous containing distinct gentoypes at the SNV. For example, a sample of 50% of cells bearing genotype AB and 50% of cells bearing AA results in 75% of A alleles and 25% B alleles. If the A allele is the reference genome, the VAF is expected to be around 25%, or 0.25, which is not close to 0, 0.5, or 1. Based on this basic logic, many methods[7–12] have been developed to infer subclones using NGS data.

1.2. *Main idea*

Inference of subclones that hinges on "unusual" VAFs is vulnerable to the noise and artifacts in the NGS data. In particular, due to the complexity and limitation of the NGS experiment, the observed VAF at an SNV can deviate from ideal values 0, 0.5, or 1 even when the cell population is homogeneous. When the population is indeed heterogeneous, noise in the NGS data can still affect the accuracy of subclone reconstruction. Currently the noise and artifacts in NGS data cannot be properly modeled and accounted for due to its complexity,[13] and therefore SNV-based subclone callers often require lengthy and ad-hoc noise filters. The effects of these noise filters on the subclone reconstruction is usually unknown.

To mitigate this problem, we consider a different approach. We assume that paired-end short reads are used in the NGS experiment. Instead of modeling reads mapped to individual SNVs, we consider a pair of them, i.e., mutation pairs. We consider proximal mutation pairs that are close enough to be phased by some of the same short reads. Such mutation pairs can be retrieved by existing tool[14] with high confidence. Since there are two loci in each mutation pair, the observed data are haplotypes (of two phased SNVs). With four possible nucleotides at each SNV, there could be up to 16 different haplotypes at each mutation pair (details in Section 2.2). Observing more than two haplotypes is evidence of tumor heterogeneity, again, due to diploidy. See Fig. 1 for an example.

We assume a total of T ($T \geq 1$) samples are obtained from a single patient, and consider intra-tumor heterogeneity as the main inference goal. Consider a finite number of K mutation pairs that are shared across the T samples, and assume that an unknown number of C subclones are present. We denote a subclone by a set of matrices $\boldsymbol{z}_{kc}$ for mutation pairs $k = 1, 2, \ldots, K$. Each $\boldsymbol{z}_{kc}$ is a 2×2 matrix that codes the two diploid genotypes of mutation pair k for subclone c. Detail of $\boldsymbol{z}_{kc}$ is given in the upcoming discussion. We also assume that the C subclones are shared by the T samples, with different population frequencies for each sample, denoted by $\boldsymbol{w}_t = (w_{t0}, w_{t1}, \ldots, w_{tC})$ for sample t, where $0 < w_{tc} < 1$ for all c and $\sum_{c=0}^{C} w_{tc} = 1$. Using the NGS data we infer $\boldsymbol{Z}$ and $\boldsymbol{w}$ based on a simple idea that that variant reads can only arise from subclones with variant genotypes.

Among exiting methods, SciClone, TrAp, Clomial, PhyloSub and PhyloWGS [8–12] are of relevance to this work. The main difference of our method from all the other existing methods is that we use mutation pairs as experimental units instead of unpaired SNVs. Also our model is based on latent feature allocation methods that allow overlapping mutations across subclones. This is different from cluster-based methods in the literature.

The paper is structured as follows: Sec. 2 and Sec. 3 describes the Bayesian feature allocation model and posterior inference, respectively. Sec. 4 presents two simulation studies. Sec. 5 reports analysis results for a real-world dataset. Sec. 6 concludes with a final discussion.

2. Probability Model

2.1. *Sampling Model*

We start the construction of a sampling model by considering one mutation pair k (see Fig. 1). Two loci, denoted by $r = 1, 2$ mark the mutation pair. A set of short reads are mapped to the genomic region that contain the two loci. Index short reads by d. When short reads are mapped to the region, we require that at least one of the two loci is covered, or else the short reads are excluded from our analysis since they do not provide any information on the mutation pair. Consider short read d mapped to mutation pair k in sample t. Define $\boldsymbol{s}_{tk}^{(d)} = \left\{s_{tkr}^{(d)}\right\}_{r=1,2} = \left(s_{tk1}^{(d)}, s_{tk2}^{(d)}\right)$, where $s_{tkr}^{(d)}$ takes three values of $\{0, 1, -\}$ representing that the base on read d mapped to locus r is reference, variant, or missing, respectively. For example, in Fig. 1 locus $r = 1$, $s_{tk1}^{(d)} = 0$ for read $d = 1$, $s_{tk1}^{(d)} = 1$ for read $d = 2$, and $s_{tk1}^{(d)} = -$ for read $d = 3$. Aggregating across two loci, each $\boldsymbol{s}_{tk}^{(d)}$ can take $G = 8$ possible genotypes, including the reference, variant, and missing genotypes, denoted by $\mathcal{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_G\} = \{(0,0), (0,1), (1,0), (1,1), (-,0), (-,1), (0,-), (1,-)\}$, where each $\boldsymbol{h}_g = \{h_{gr}\}_{r=1,2} = (h_{g1}, h_{g2})$ denotes the potential genotype at each locus r of a short read. Let $n_{tkg} = \sum_d I\left(\boldsymbol{s}_{tk}^{(d)} = \boldsymbol{h}_g\right)$ be the read count representing the number of short reads having genotype $\boldsymbol{h}_g$. Here $I()$ is the indicator function. The total number of reads that are mapped to the loci of the mutation pair k in sample t is then $N_{tk} = \sum_{g=1}^{G} n_{tkg}$. We assume a multinomial sampling model for n_{tkg} conditional on N_{tk}, given by

$$n_{tk1}, \ldots, n_{tkG} \mid N_{tk}, p_{tk1}, \ldots, p_{tkG} \overset{indep.}{\sim} \text{Multinomial}\left(N_{tk}; p_{tk1}, \ldots, p_{tkG}\right), \tag{1}$$

where $p_{tkg} = Pr(\boldsymbol{s}_{tk}^{(d)} = \boldsymbol{h}_g)$ is the probability that a read bears genotype $\boldsymbol{h}_g$ on mutation pair k in sample t.

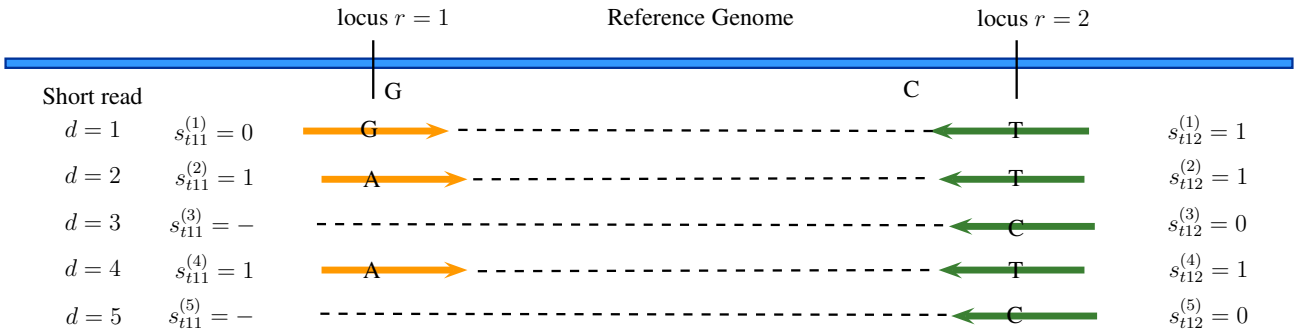


Fig. 1. Illustration of read count data for a mutation pair. There is a total of five reads mapped to the two loci that mark the mutation pair. The five reads exhibit genotypes $(0,1)$, $(1,1)$, $(-,0)$, $(1,1)$, $(-,0)$, which implies that there could be three haplotypes for the mutation pair in the sample.

2.2. *Subclone Representation using Z*

We collect all the $\boldsymbol{z}_{kc}$'s in a matrix format, denoted as a $K \times C$ matrix $\boldsymbol{Z} = [\boldsymbol{z}_{kc}]$. Technically, $\boldsymbol{Z}$ is a matrix of matrices, since each $\boldsymbol{z}_{kc}$ is itself a matrix. See Fig. 2. The total number of

subclones, denoted by C, is random. The c-th column of $\boldsymbol{Z}$, $\boldsymbol{z}_c = (\boldsymbol{z}_{1c}, \ldots, \boldsymbol{z}_{Kc})$ denotes one particular subclone. Each element $\boldsymbol{z}_{kc}$ records the two alleles of a particular mutation pair k for subclone c. Let $j = 1, 2$ index the two alleles in a subclone and $r = 1, 2$ represent the two loci in a mutation pair. We write $\boldsymbol{z}_{kc} = \{z_{kcjr}\} = ((z_{kc11}, z_{kc12}), (z_{kc21}, z_{kc22}))$. See Fig. 2 for an example. Note that $z_{kcjr} = 1$ indicates that r-th locus of j-th allele of $\boldsymbol{z}_{kc}$ bears a mutation compared to the reference genome. Clearly $\boldsymbol{z}_{kc}$ can take $Q = 16$ possible values i.e. $\boldsymbol{z}_{kc} \in \{\boldsymbol{z}^{(q)}\}_{q=1}^{16} = \{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(16)}\} = \{((0,0),(0,0)), ((0,0),(0,1)), \ldots, ((1,1),(1,1))\}$. For example, in Fig. 2 reference genome at the loci of mutation pair 1 is AT, and the corresponding genotype of subclone 3 is $((G,T),(G,C))$, which translates to $\boldsymbol{z}_{kc} = ((1,0),(1,1))$. However, we can collapse some $\boldsymbol{z}^{(q)}$ values since we do not distinguish the order of the two alleles for a mutation pair in a subclone. That is $\boldsymbol{z}_{kc} = ((z_{kc\underline{1}1}, z_{kc\underline{1}2}), (z_{kc\underline{2}1}, z_{kc\underline{2}2}))$ and $\boldsymbol{z}_{kc} = ((z_{kc\underline{2}1}, z_{kc\underline{2}2}), (z_{kc\underline{1}1}, z_{kc\underline{1}2}))$ lead to the same probability model. Therefore, the two alleles are coded invariant of their orders and we reduce the number of possible outcomes of $\boldsymbol{z}_{kc}$ to from 16 to $Q = 10$ and they are listed as: $\boldsymbol{z}^{(1)} = ((0,0),(0,0))$, $\boldsymbol{z}^{(2)} = ((0,0),(0,1))$, $\boldsymbol{z}^{(3)} = ((0,0),(1,0))$, $\boldsymbol{z}^{(4)} = ((0,0),(1,1))$, $\boldsymbol{z}^{(5)} = ((0,1),(0,1))$, $\boldsymbol{z}^{(6)} = ((0,1),(1,0))$, $\boldsymbol{z}^{(7)} = ((0,1),(1,1))$, $\boldsymbol{z}^{(8)} = ((1,0),(1,0))$, $\boldsymbol{z}^{(9)} = ((1,0),(1,1))$ and $\boldsymbol{z}^{(10)} = ((1,1),(1,1))$.

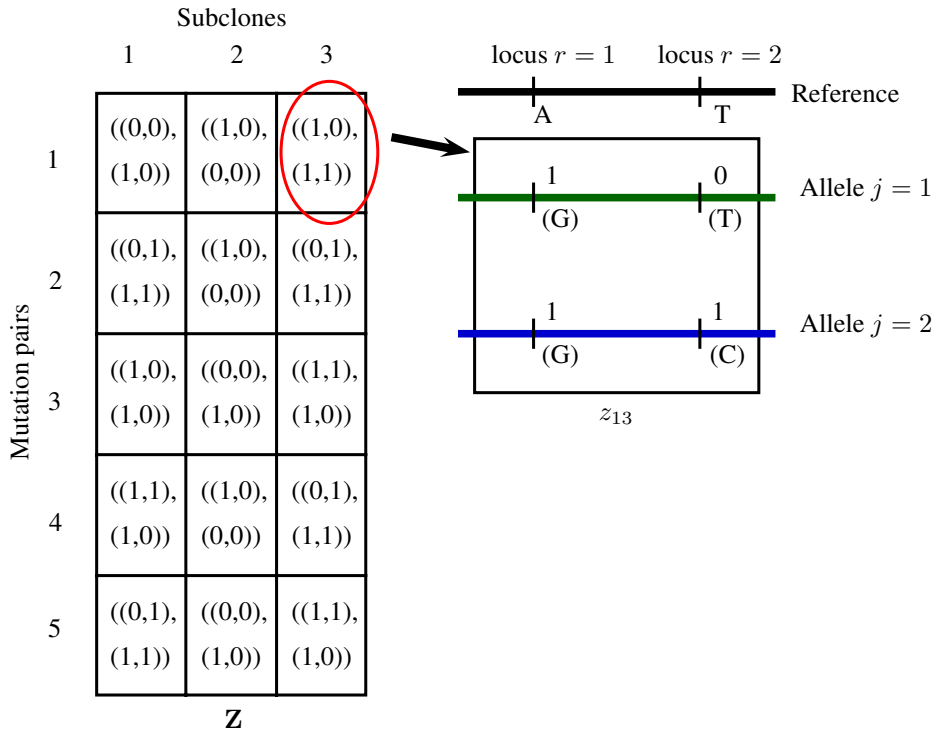


Fig. 2. Illustration of $\boldsymbol{Z}$ (left panel) for subclones in a sample and a particular subclonal genotypes for a mutation pair (right panel). Each column of Z represents a subclone, with each element representing the subclonal genotypes for a mutation pair. The genotypes for mutation 1 in subclone 3 is $((1,0),(1,1))$, which can be shown as a stylized example in the right panel.

Each sample is potentially an admixture of the subclones (columns of $\boldsymbol{Z}$), mixed in dif-

ferent proportions. Given $\boldsymbol{Z}$, we can denote the proportions of the C subclones by $\boldsymbol{w}_t = (w_{t0}, w_{t1}, \ldots, w_{tC})$ for sample t, where $0 < w_{tc} < 1$ for all c and $\sum_{c=0}^{C} w_{tc} = 1$. Notice that the subclones are common for all tissue samples, but the weights w_{tc} vary across samples. A background subclone, which has no biological meaning and is indexed by $c = 0$, is included to account for experimental noise (sequencing errors, mapping errors, etc.).

2.3. *Prior model*

Prior for p_{tkg}: The prior for the multinomial probabilities p_{tkg} in (1) is based on a simple idea: a short read harboring a particular haplotype $\boldsymbol{h}_g$ can only come from subclones that also harbor the same haplotype in their genomes. The probability of observing such a short read depends on the population frequencies $\boldsymbol{w}_t$ of such subclones harboring the haplotype. Therefore, we define

$$p_{tkg} \propto \sum_{c=1}^{C} w_{tc}\, A(\boldsymbol{h}_g, \boldsymbol{z}_{kc}) + w_{t0}\, \rho_g, \text{ for } g = 1, \ldots, 8, \tag{2}$$

where $A(\boldsymbol{h}_g, \boldsymbol{z}_{kc})$ is the expected proportion of alleles with genotype $\boldsymbol{h}_g$ at mutation pair k of subclone c. Accounting for the potential missing genotype at each of the two loci corresponding to the mutation pair, there are three ways a short read can cover the mutation pair: (i) the read maps to both loci; (ii) the read maps to the second locus but does not map to the first (left missing), and (iii) the read maps to the first locus but not the second locus (right missing). Therefore, we define

$$A(\boldsymbol{h}_g, \boldsymbol{z}_{kc}) = \begin{cases} \sum_{j=1}^{2} 0.5 \times I\left(h_{g1} = z_{kcj1}, h_{g2} = z_{kcj2}\right), & \text{for } g = 1, \ldots, 4; \\ \sum_{j=1}^{2} 0.5 \times I(h_{g2} = z_{kcj2}), & \text{for } g = 5, 6; \\ \sum_{j=1}^{2} 0.5 \times I(h_{g1} = z_{kcj1}), & \text{for } g = 7, 8. \end{cases} \tag{3}$$

In (3), the three equations correspond to the three coverage cases (i) – (iii) mentioned above. The factor 0.5 is used to reflect that any short read comes from one of the two alleles in the genome with equal probability. Quantifying the expected proportion of alleles in the genome, $A(\boldsymbol{h}_g, \boldsymbol{z}_{kc})$ can only take three values 0, 0.5 or 1. According to (3) and assuming no sequencing error, a read that covers both loci ($g = 1, 2, 3, 4$) and bears genotype $\boldsymbol{h}_g$ must be generated from a subclone having the same $\boldsymbol{h}_g$ genotype in at least one allele. When the read only covers one of the two loci, the requirement is to match the sequence on the covered locus only, and hence the equations in (3) for cases $g = 5, 6, 7, 8$.

In (2) we also include a background subclone denoted by $c = 0$ with proportion of w_{t0} to account for experimental noise. The background subclone does not exist and is only used as a mathematical device to account for noise and artifacts in the NGS data. See Ref. [15] for details.

Prior for $\boldsymbol{Z}$: We develop a latent-feature-allocation prior for the latent matrix $\boldsymbol{Z}$, the elements of which take categorical values. The prior $p(\boldsymbol{Z} \mid C)$ is constructed under fixed C. Let $\boldsymbol{\pi}_c = (\pi_{c1}, \pi_{c2}, \ldots, \pi_{cQ})$ where $p(\boldsymbol{z}_{kc} = \boldsymbol{z}^{(q)}) = \pi_{cq}$ and $\sum_{q=1}^{Q} \pi_{cq} = 1$. We use the beta-Dirichlet distribution[16] as the prior for $\boldsymbol{\pi}_c$. Conditional on C, $p(\boldsymbol{z}_{kc} = \boldsymbol{z}^{(1)}) = \pi_{c1}$ follows a beta distribution with parameters 1 and α/C, and $(\tilde{\pi}_{c2}, \ldots, \tilde{\pi}_{cQ})$, where $\tilde{\pi}_{cq} = \pi_{cq}/(1 - \pi_{c1})$ with $q = 2, \ldots, Q$,

follows a Dirichlet distribution with parameters $(\gamma_2, \ldots, \gamma_Q)$. Here $\boldsymbol{z}^{(1)}$ is special because it refers to the reference genome. We write

$$\boldsymbol{\pi}_c \sim \text{Beta-Dirichlet}(\alpha/C, 1, \gamma_2, \ldots, \gamma_Q).$$

As shown in Ref. [17], the marginal limiting distribution of $\boldsymbol{Z}$ follows a categorical Indian buffet process (cIBP) as $C \to \infty$.

Prior for $\boldsymbol{w}$: Next, we introduce a prior distribution for $\boldsymbol{w}_t$ as

$$\boldsymbol{w}_t \mid C \overset{iid}{\sim} \text{Dirichlet}(d_0, d, \ldots, d),$$

for $t = 1, \ldots, T$. For all practical purpose, we set $d_0 < d$ to imply that the hypothetical background subclone has a small population frequency.

Prior for $\boldsymbol{\rho}$ and C: Then we construct the prior for $\boldsymbol{\rho}$, where ρ_g is the conditional probability of observing a read with a genotype $\boldsymbol{h}_g$ due to experimental noise. We assume Dirichlet priors on ρ_g's,

$$\rho_{g_1} \sim \text{Dirichlet}(d_1, \ldots, d_1);\ \rho_{g_2} \sim \text{Dirichlet}(2d_1, 2d_1);\ \rho_{g_3} \sim \text{Dirichlet}(2d_1, 2d_1) \tag{4}$$

where $g_1 = \{1, 2, 3, 4\}$, $g_2 = \{5, 6\}$ and $g_3 = \{7, 8\}$.
Finally, we put a geometric distribution prior on number of subclones i.e. $C \sim \text{Geom}(r)$, and hence $E(C) = 1/r$ a priori.

3. Posterior Inference

3.1. *Posterior computation:*

Markov chain Monte Carlo (MCMC) simulation[18] is used to draw samples from the posterior of the unknown parameters. Let $\boldsymbol{x} = (\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{w}, \boldsymbol{\rho})$ denote all the parameters except C. With fixed C, sampling $\boldsymbol{x}$ from the respective posterior distribution is straightforward. Gibbs sampling transition probabilities are used to update $\boldsymbol{Z}$ and $\boldsymbol{\pi}$, and Metropolis-Hastings transition probabilities are used to update $\boldsymbol{w}$ and $\boldsymbol{\rho}$.
Updating the value of C is more challenging, since it involves change of dimension of parameter space. We use an approach similar to Ref. [19], which is a reversible jump[20] style algorithm, with a model comparison approach using modified fractional Bayes factor.[21,22] The basic idea is to consider a finite number of possible C, denoted by $\{C_{\min}, \ldots, C_{\max}\}$, split the data into a training set $\boldsymbol{n}' = b\boldsymbol{n}$ and a test set $\boldsymbol{n}'' = (1-b)\boldsymbol{n}$ (where $0 < b < 1$), and do a model comparison among those possible C. Details are given in [19].

3.2. *Estimate of Z:*

The point estimates for the parameters are determined as follows. We use the posterior mode C^* as a point estimate of C. Conditional on C^*, we follow Ref. [19] to find a point estimate of $\boldsymbol{Z}$. For any two $K \times C^*$ matrices $\boldsymbol{Z}$ and $\boldsymbol{Z}'$, $1 \le c, c' \le C^*$, let $\mathcal{D}_{cc'}(\boldsymbol{Z}, \boldsymbol{Z}') = \sum_{k=1}^{K} \|\boldsymbol{z}_{kc} - \boldsymbol{z}'_{kc'}\|_1$. Here we take the vectorized form of $\boldsymbol{z}_{kc}$ and $\boldsymbol{z}'_{kc'}$ to compute L^1 distance between them. The distance between $\boldsymbol{Z}$ and $\boldsymbol{Z}'$ is then defined as $d(\boldsymbol{Z}, \boldsymbol{Z}') = \min_{\boldsymbol{\sigma}} \sum_{c=1}^{C^*} \mathcal{D}_{c,\boldsymbol{\sigma}_c}(\boldsymbol{Z}, \boldsymbol{Z}')$, where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_{C^*})$

is a permutation of $\{1,\ldots,C^*\}$ and the minimum is over all possible permutations. A posterior point estimate for $\boldsymbol{Z}$ is defined as

$$\boldsymbol{Z}^* = \underset{\boldsymbol{Z}'\in\{\boldsymbol{Z}^{(l)},l=1,\ldots,L\}}{\arg\min} \frac{1}{L}\sum_{l=1}^{L} d(\boldsymbol{Z}^{(l)},\boldsymbol{Z}'),$$

where $\{\boldsymbol{Z}^{(l)}, l=1,\ldots,L\}$ are posterior Monte Carlo samples of $\boldsymbol{Z}$. Finally, we report posterior point estimates $\boldsymbol{w}^*$ and $\boldsymbol{\rho}^*$ conditional on C^* and $\boldsymbol{Z}^*$ and calculate posterior point estimates $\boldsymbol{p}^*$ in order to check goodness of fit of the model.

4. Simulation

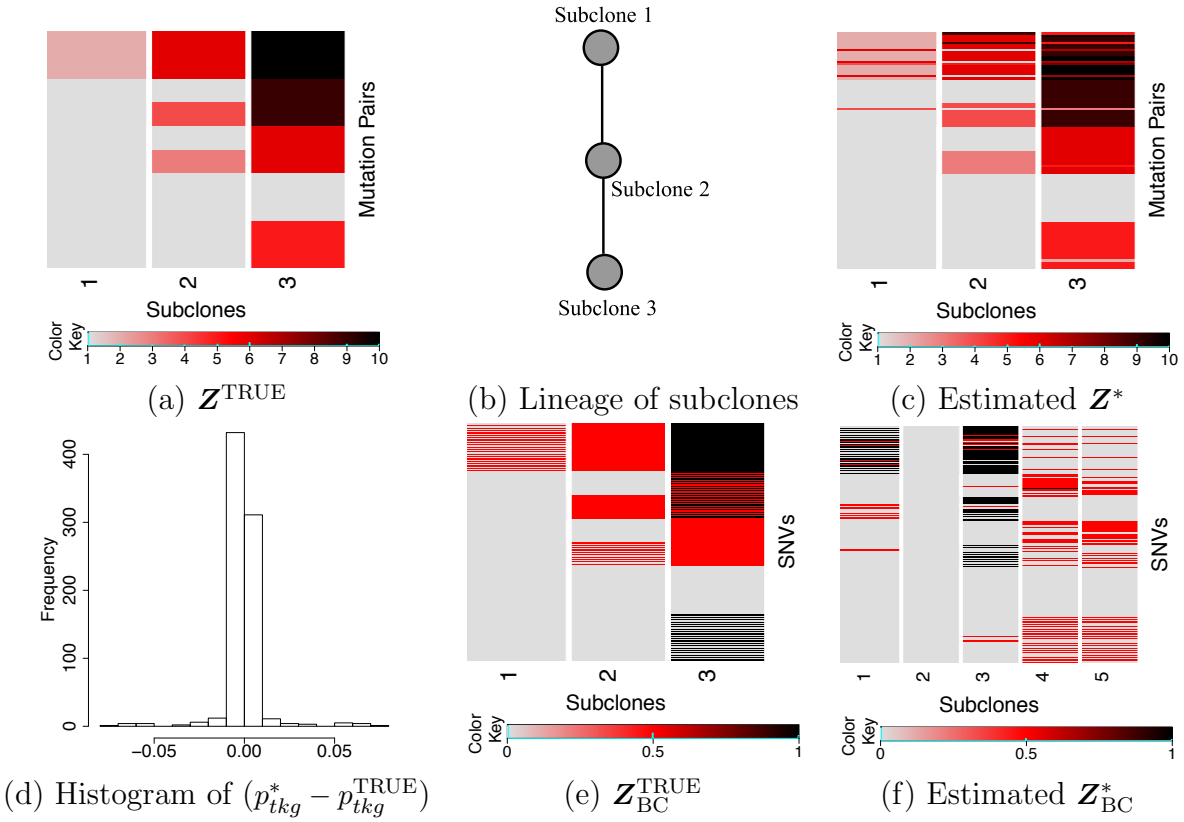
4.1. *Simulation 1*

We carry out two simulation studies to validate our proposed model. In the first simulation, we consider $K=100$ mutation pairs for $T=1$ sample. We assume the number of latent subclones is $C^{\text{TRUE}}=3$, and set the subclone proportions as $\boldsymbol{w}^{\text{TRUE}}=(1\times10^{-7},0.65,0.28,0.07)$ (note that 1×10^{-7} refers to the proportion of the background subclone). The latent $\boldsymbol{Z}^{\text{TRUE}}$ matrix is shown in Fig. 3(a) in the form of a heatmap. For example, subclone 3 has genotype $\boldsymbol{z}^{(q)}$ with different q values. Specifically, $q=10$ for mutation pairs 1-20, $q=9$ for mutation pairs 21-40, $q=6$ for mutation pairs 41-60, $q=1$ for mutation pairs 61-80, and $q=5$ for mutation pairs 81-100. Fig. 3(b) shows a possible lineage structure among subclones. We generate $\boldsymbol{\rho}^{\text{TRUE}}$ from its prior given in Eq. (4) with hyperparameter $d_1=1$. Next, we calculate multinomial probabilities p_{tkg}^{TRUE} shown in Eq. (2) and (3) from the simulated $\boldsymbol{Z}$, $\boldsymbol{w}$ and $\boldsymbol{\rho}$. We generate random numbers ranging from 400 to 600 as total read counts N_{tk}, and finally we generate read counts n_{tkg} from the multinomial distribution given N_{tk} as shown in Eq. (1).

We fit the model with hyperparameters as $\alpha=4$, $\gamma_2=\cdots=\gamma_Q=0.5$, $d=0.5$, $d_0=0.1$, $d_1=1$, and $r=0.4$. We set $C_{\min}=1$ and $C_{\max}=10$ as the range of C. The choice of b needs to be calibrated. We choose b such that the test sample size $(1-b)\sum_{t=1}^{T}\sum_{k=1}^{K}N_{tk}$ is approximately equal to $250/\sqrt{T}$. This choice leads to better posterior inference in our calibration process. We run MCMC simulation for $50,000$ iterations, discarding the first $20,000$ iterations as initial burn-in, and keep one sample every 10 iterations. The initial values are randomly generated from the prior.

The posterior mode $C^*=3$ recovers the truth. Fig. 3(c) shows the point estimate of $\boldsymbol{Z}^{\text{TRUE}}$, given by $\boldsymbol{Z}^*$, which is very close to the truth. Fig. 3(d) shows the difference between $(p_{tkg}^* - p_{tkg}^{\text{TRUE}})$, which can be considered as the residual of model fitting. The histogram is centered at zero with a small variance that indicates a good model fit. The estimated subclone weights are $\boldsymbol{w}^*=(1.20\times10^{-168},0.650,0.277,0.073)$, which is also close to the truth. Typically in a real scenario, the number of available samples are quite low. In fact, in most of the cases data for only one sample can be obtained. We perform this simulation example in order to show that our model performs quite well even with a single sample.

We compare the performance of our model against BayClone[15] which is an SNV-based subclone caller. It chooses the model based on log pseudo marginal likelihood (LPML). According to LPML, the estimated number of subclones under BayClone is $C^*=5$, which does not recover the truth. Fig. 3(e) shows the true subclone matrix in the form of Bay-



(a) $\boldsymbol{Z}^{\text{TRUE}}$ (b) Lineage of subclones (c) Estimated $\boldsymbol{Z}^*$

(d) Histogram of $(p^*_{tkg} - p^{\text{TRUE}}_{tkg})$ (e) $\boldsymbol{Z}^{\text{TRUE}}_{\text{BC}}$ (f) Estimated $\boldsymbol{Z}^*_{\text{BC}}$

Fig. 3. (a-d) Heatmap of the true subclone matrix $\boldsymbol{Z}^{\text{TRUE}}$, lineage structure and results from posterior inference. (e-f) Heatmap of the true and estimated subclone matrix using BayClone.

Clone's notation, denoted by $\boldsymbol{Z}^{\text{TRUE}}_{\text{BC}}$, and Fig. 3(f) shows the estimated matrix $\boldsymbol{Z}^*_{\text{BC}}$, where $z_{kc} = 0$, $z_{kc} = 0.5$ and $z_{kc} = 1$ refer to homozygous wild-type, heterozygous variant and homozygous variant at SNV locus k, respectively. The estimated subclone proportions are $\boldsymbol{w}^*_{\text{BC}} = (0.004, 0.364, 0.349, 0.171, 0.057, 0.054)$. From the BayClone's output, we can notice three problems. Firstly, BayClone could not recover the true number of subclones. Secondly, since BayClone infers the subclone structure by VAF of an SNV, the connection between adjacent SNVs is not modeled, and thus BayClone could not recover the $\boldsymbol{Z}$ matrix and cellular fractions accurately. For example, BayClone could not distinguish the difference between $\boldsymbol{z}_{kc} = \boldsymbol{z}^{(4)}$ and $\boldsymbol{z}_{kc} = \boldsymbol{z}^{(6)}$ in our model. Lastly, because of the noise in the data, BayClone includes a relatively larger proportion for the background subclone ($w_0 = 0.004$ in this example) which is significantly reduced for mutation pair data.

4.2. *Simulation 2*

In the second simulation study, we generate hypothetical reads data for $K = 100$ mutation pairs and $T = 5$ samples. We assume $C^{\text{TRUE}} = 4$. The subclone matrix $\boldsymbol{Z}^{\text{TRUE}}$ is shown in Fig. 4(a) and a possible lineage structure is given in Fig. 4(c). For each sample t, we generate the subclone proportions $\boldsymbol{w}^{\text{TRUE}}_t$ from $Dirichlet(0.01, \sigma(20, 14, 10, 4))$, where $\sigma(20, 14, 10, 4)$ is a random permutation of $(20, 14, 10, 4)$. The proportions $\boldsymbol{w}^{\text{TRUE}}$ which is now a matrix shown

by a heatmap in Fig. 4(b). In the heatmap for $\boldsymbol{w}$, darker color indicates high abundance of a subclone in a sample, and light grey color represents low abundance. The parameters $\boldsymbol{\rho}^{\text{TRUE}}$ and N_{tk} are generated using the same approach as before. Finally, we calculate p_{tkg}^{TRUE} and generate read counts n_{tkg} from Eq. (1) similar to previous simulation.

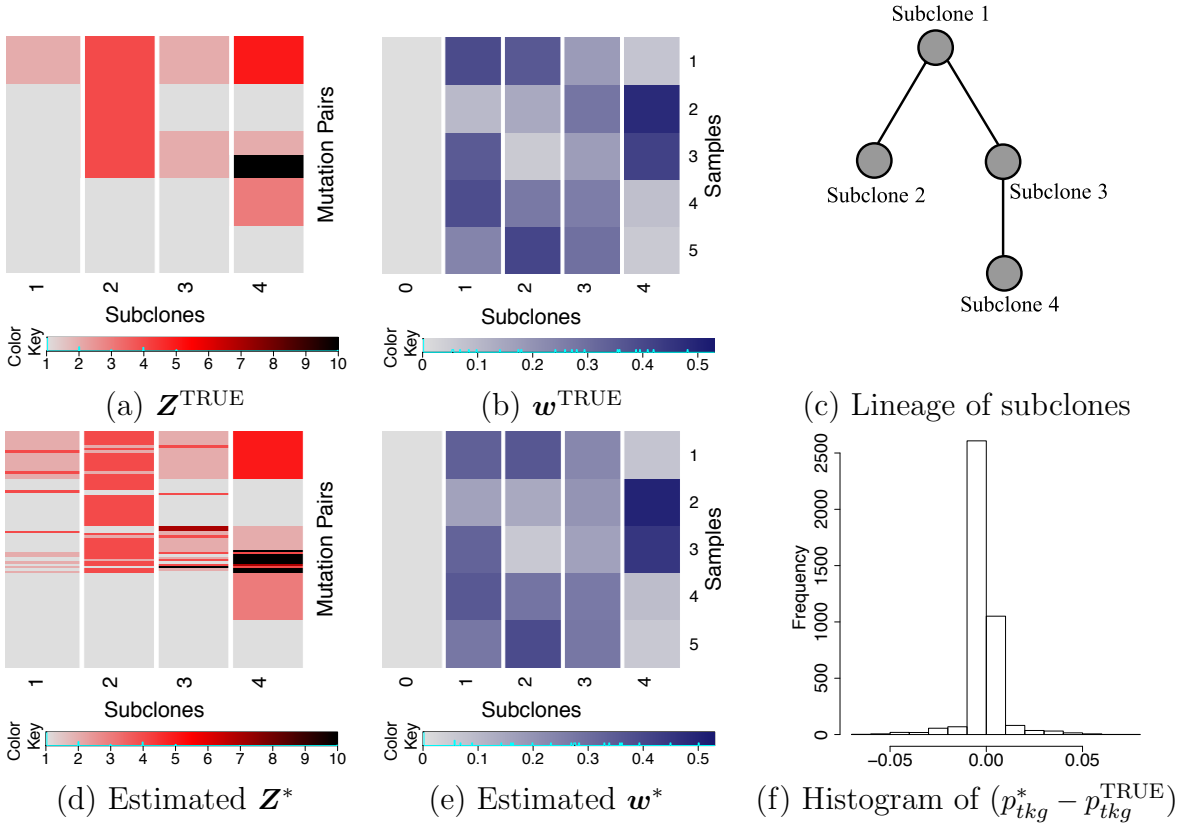


Fig. 4. Heatmap of the true subclone matrix, lineage structure and the results from posterior inference.

We fit the model with the same hyperparameters and same MCMC setting, except here we use $C_{\max} = 8$ in order to accelerate MCMC sampling. Also here due to the presence of multiple samples, we use a smaller (compared to simulation 1) test sample size. The posterior mode $C^* = 4$ recovers the truth. Fig. 4(d) shows the heatmap of $\boldsymbol{Z}^*$, and Fig. 4(e) shows the heatmap of $\boldsymbol{w}^*$. Comparing those two figures with Fig. 4(a) and 4(b), we can see that the truth is nicely recovered. Some mismatches are due to the relatively complex subclone structure. Fig. 4(f) shows the histogram of $(p_{tkg}^* - p_{tkg}^{\text{TRUE}})$ which indicates a good model fit.
We also compare our results with BayClone for this simulation. BayClone chooses the model with 5 subclones, which does not recover the truth.

5. Head and Neck Cancer Dataset

Whole exome data of 30 pairs of matched tumor (head and neck cancer) and normal samples are downloaded from the Sequence Read Archive (`http://www.ncbi.nlm.nih.gov/sra`).[23] We map the pair-end reads from the FASTQ format files to the human genome (version HG19)

using BWA to generate BAM files for each individual sample. GATK's UnifiedGenotyper is used to call variants and to generate a single VCF file for all of them. Next task is to find mutation pair positions, their genotypes and number of reads supporting them. It is done by a bioinformatics tool *LocHap*[14] which searches for multiple single nucleotide variants (SNVs) that are scaffolded by the same reads. The scaffolded SNVs are referred to as local haplotypes. When a local haplotype exhibits more than two genotypes, *LocHap* calls it a local haplotype variant (LHV). Using the individual BAM file and the combined VCF file, *LocHap* generates HCF format output file.[14] HCF files contain LHV with two or three SNV locations. This whole process runs very fast as *LocHap* is an ultra-fast tool that can process an WES sample with about 30X coverage under a minute.[14] On an average we find a few hundreds LHVs with high quality in a WES sample. We select LHVs with two SNV locations as we are interested in mutation pairs only. Among those LHVs, we first remove the LHVs where the loci of two SNVs are very close to each other (within, say 50 bps) or close to other types of structural variants such as indels. We remove the LHVs where most reads were aligned to any of the SNVs at a base near the end of the reads. Also we filter out those LHVs where any of the SNVs are mapped by most reads with strand bias. At first, we find the intersection of mutation pair loci between normal and tumor samples and then we select randomly around 100 loci for each sample and record the read data from HCF files. In order to compare the underlying subclonal structure of normal and tumor samples we run our model on both separately. We run MCMC for $50,000$ iterations and discard the first $20,000$ iterations as initial burn-in. We use thinning count equals to 10. Hyperparameter settings are exactly same as the simulation 1 (Section 4.1).

Fig. 5 shows the number of subclones of a tumor sample and its matched normal for all 30 samples. Note that in almost all the samples the number of subclones in tumor is higher than the matched normal. In Fig. 6, we put subclone matrix ($\boldsymbol{Z}$) from six tumor and matched

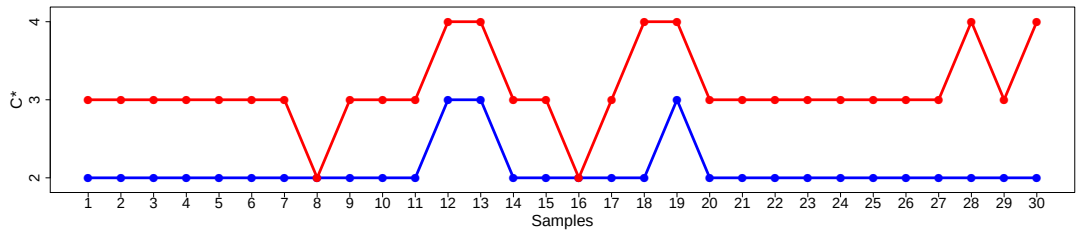


Fig. 5. Inferred number of subclones (C^*) for tumor (in red) and matched normal (in blue)

normal samples side by side. As one can notice, in tumor sample the corresponding subclonal structure is somewhat preserved with an addition of a new subclone. This indicates that tumor sample is more heterogeneous than the corresponding matching normal sample. We show the proportion of each subclone below each column of $\boldsymbol{Z}$ and columns of $\boldsymbol{Z}$ is reordered according to decreasing order of weights of the subclones.

We also run BayClone on those samples. The results look different. Due to space limitation, we omit the details since BayClone results are less reliable according to simulation studies.

Analysis of real data provides valuable clinical information. For example, one could seek potential biomarker mutation pairs for targeted therapy. These results could also be used as future diagnosis reference.

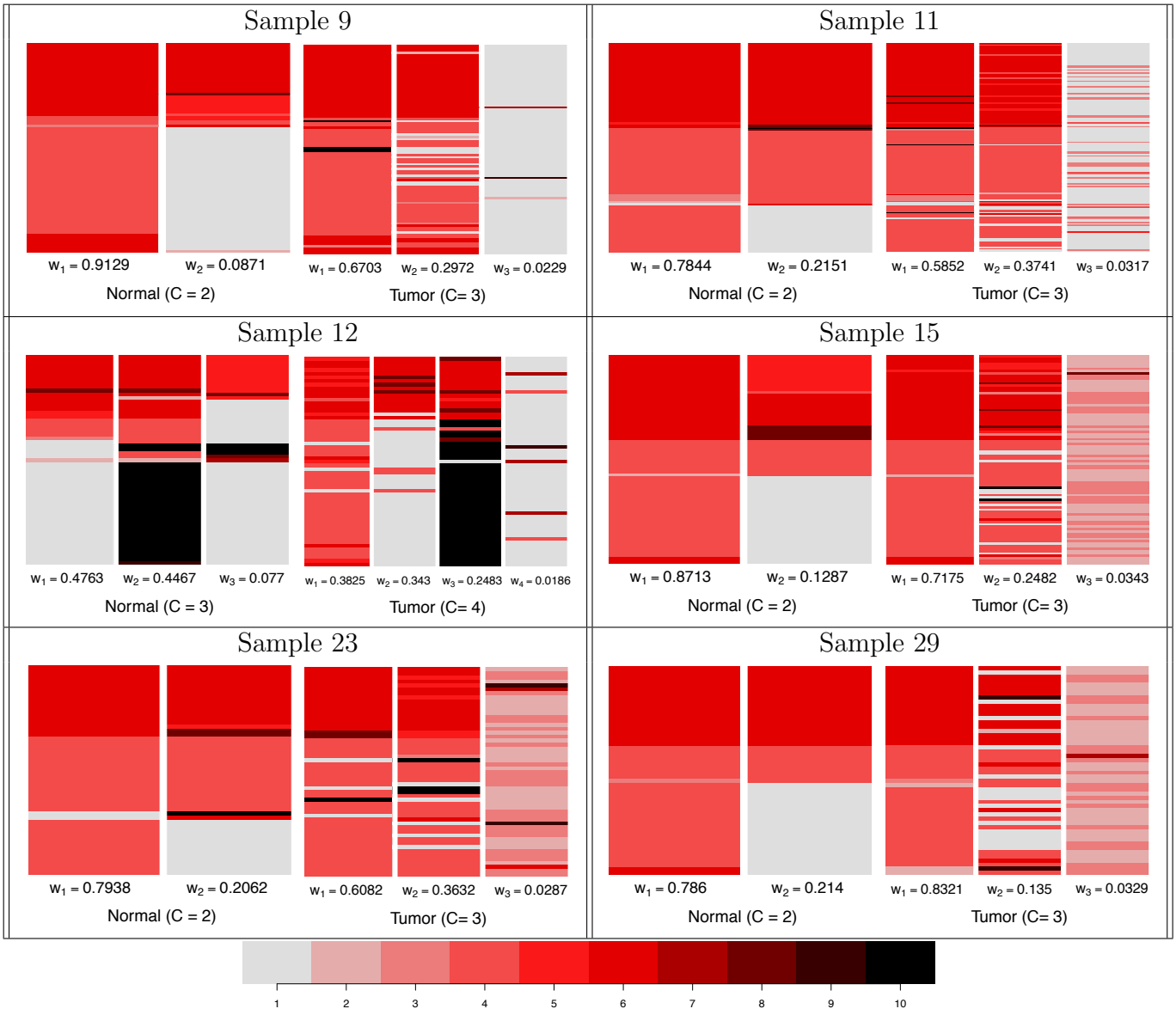


Fig. 6. Heatmap of subclone matrix $\boldsymbol{Z}$ from selected 6 samples (ordered according to age).

6. Discussion and future work

With the proposed model we infer subclonal structure and their proportions using mutation pairs data. The methods describe tumor heterogeneity in a principled manner based on a feature allocation model. It explicitly models overlapping mutation pairs across subclones. Through simulations, we show that mutation pair-based inference is more powerful

than SNV-based subclone calling. This is not surprising since mutation pairs naturally provide heterogeneity of tumor samples through poly-genotypic short reads. In other words, direct evidence of having more than two haplotypes from short reads can be used to infer subclones in a tumor sample rather than indirect modeling on unusual VAFs for SNVs.
Our approach can be extended to model more than two SNVs. In order to accommodate more number of SNVs we only need to increase the number of categorical values that the $\boldsymbol{Z}$ matrix can take. Also, we are working on extensions that explicitly take into account potential phylogenetic relationship of subclones, which requires modeling the dependence among columns of the $\boldsymbol{Z}$ matrix.

References

1. N. D. Marjanovic, R. A. Weinberg and C. L. Chaffer, *Clinical chemistry* **59**, 168 (2013).
2. V. Almendro, A. Marusyk and K. Polyak, *Annual Review of Pathology: Mechanisms of Disease* **8**, 277 (2013).
3. K. Polyak, *The Journal of clinical investigation* **121**, p. 3786 (2011).
4. J. Stingl and C. Caldas, *Nature Reviews Cancer* **7**, 791 (2007).
5. M. Shackleton, E. Quintana, E. R. Fearon and S. J. Morrison, *Cell* **138**, 822 (2009).
6. D. L. Dexter, H. M. Kowalski, B. A. Blazar, Z. Fligiel, R. Vogel and G. H. Heppner, *Cancer Research* **38**, 3174 (1978).
7. L. Oesper, A. Mahmoody and B. J. Raphael, *Genome Biol* **14**, p. R80 (2013).
8. C. A. Miller, B. S. White, N. D. Dees, M. Griffith, J. S. Welch, O. L. Griffith, R. Vij, M. H. Tomasson, T. A. Graubert, M. J. Walter *et al.*, *PLoS computational biology* **10**, p. e1003665 (2014).
9. F. Strino, F. Parisi, M. Micsinai and Y. Kluger, *Nucleic acids research* **41**, e165 (2013).
10. W. Jiao, S. Vembu, A. G. Deshwar, L. Stein and Q. Morris, *BMC bioinformatics* **15**, p. 35 (2014).
11. H. Zare, J. Wang, A. Hu, K. Weber, J. Smith, D. Nickerson, C. Song, D. Witten, C. A. Blau and W. S. Noble, *PLoS computational biology* **10**, p. e1003703 (2014).
12. A. G. Deshwar, S. Vembu, C. K. Yung, G. H. Jang, L. Stein and Q. Morris, *Genome biology* **16**, p. 35 (2015).
13. H. Li, *Bioinformatics* **30**, 2843 (2014).
14. S. Sengupta, K. Gulukota, Y. Zhu, C. Ober, K. Naughton, W. Wentworth-Sheilds and Y. Ji, *Nucleic Acids Research (to appear)* (2015).
15. S. Sengupta, J. Wang, J. Lee, P. Müller, K. Gulukota, A. Banerjee and Y. Ji, *Pacific Symposium of Biocomputing* , 467 (2015).
16. Y. Kim, L. James and R. Weissbach, *Biometrika* **99**, 127 (2012).
17. S. Sengupta, J. Ho and A. Banerjee, *Two Models Involving Bayesian Nonparametric Techniques*, tech. rep., University of Florida (2013).
18. S. Brooks, A. Gelman, G. Jones and X.-L. Meng, *Handbook of Markov Chain Monte Carlo* (CRC Press, 2011).
19. J. Lee, P. Müller, Y. Ji and K. Gulukota, *A Bayesian Feature Allocation Model for Tumor Heterogeneity*, tech. rep., UC Santa Cruz (2013).
20. P. J. Green, *Biometrika* **82**, 711 (1995).
21. A. O'Hagan, *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 99 (1995).
22. G. Casella and E. Moreno, *Journal of the American Statistical Association* **101**, 157 (2006).
23. N. Stransky, A. M. Egloff, A. D. Tward, A. D. Kostic, K. Cibulskis, A. Sivachenko, G. V. Kryukov, M. S. Lawrence, C. Sougnez, A. McKenna *et al.*, *Science* **333**, 1157 (2011).