The Annals of Applied Statistics
2020, Vol. 14, No. 4, 1856–1877
https://doi.org/10.1214/20-AOAS1368

RNDCLONE: TUMOR SUBCLONE RECONSTRUCTION BASED ON INTEGRATING DNA AND RNA SEQUENCE DATA

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

[1]*Department of Public Health Sciences, University of Chicago,* [*]*tjzhou95@gmail.com;* [†]*yji@health.bsd.uchicago.edu*
[2]*Research Institute, NorthShore University HealthSystem, subhajit06@gmail.com*
[3]*Department of Mathematics, University of Texas at Austin, pmueller@math.utexas.edu*

Tumor cell population consists of genetically heterogeneous subpopulations, known as subclones. Bulk sequencing data using high-throughput sequencing technology provide total and variant DNA and RNA read counts for many nucleotide loci as a mixture of signals from different subclones. We present RNDClone as a tool to deconvolute the mixture and reconstruct the subclones with distinct DNA genotypes and RNA expression profiles. In particular, we infer the number and population frequencies of subclones as well as subclonal copy numbers, variant allele numbers and gene expression levels by jointly modeling DNA and RNA read counts from the same tumor samples based on generalized latent factor models. Incorporating data at the RNA level provides new insights into intra-tumor heterogeneity in addition to the existing DNA-based inference. Performance of RNDClone is assessed using simulated and real-world datasets, including an analysis of three samples from a lung cancer patient in The Cancer Genome Atlas (TCGA). A potential fatal subclone is identified from the primary tumor which could explain the rapid prognosis and sudden death of the patient despite a promising diagnosis by conventional standards. The R package `RNDClone` is available in the Supplementary Material (Zhou et al. (2020)) and online at https://github.com/tianjianzhou/RNDClone.

1. Introduction. We develop a novel framework for statistical inference to understand intra-tumor heterogeneity. Biologically, the proposed approach is the first one to coherently combine information from both DNA and RNA level data using total and variant DNA and RNA read counts. Methodologically, the proposed inference is the first approach to formally cast the question as a generalized factor analysis problem. Adequate inference for tumor heterogeneity is a key ingredient for precision oncology.

1.1. *Background.* During tumorigenesis, tumor cells acquire and accumulate somatic mutations that give rise to genetically different cell subpopulations (Nowell (1976), Heppner (1984), Shackleton et al. (2009)). This phenomenon is known as intra-tumor heterogeneity. Each cell subpopulation, referred to as a *subclone*, consists of cells that have the same genetic architecture, such as point mutations and copy number aberrations (CNAs). These modifications must exert their effects through downstream cascades of molecular events, such as transcription or translation. Therefore, it is of great interest and importance to understand the downstream functional effects of the genetic modifications. For example, do subclonal DNA mutations affect mRNA expression? Is the effect subclonal as well? With DNA and RNA sequencing data on the same set of tumor samples, it is possible to simultaneously infer the subclones, their genetic architecture and the impact on mRNA expression.

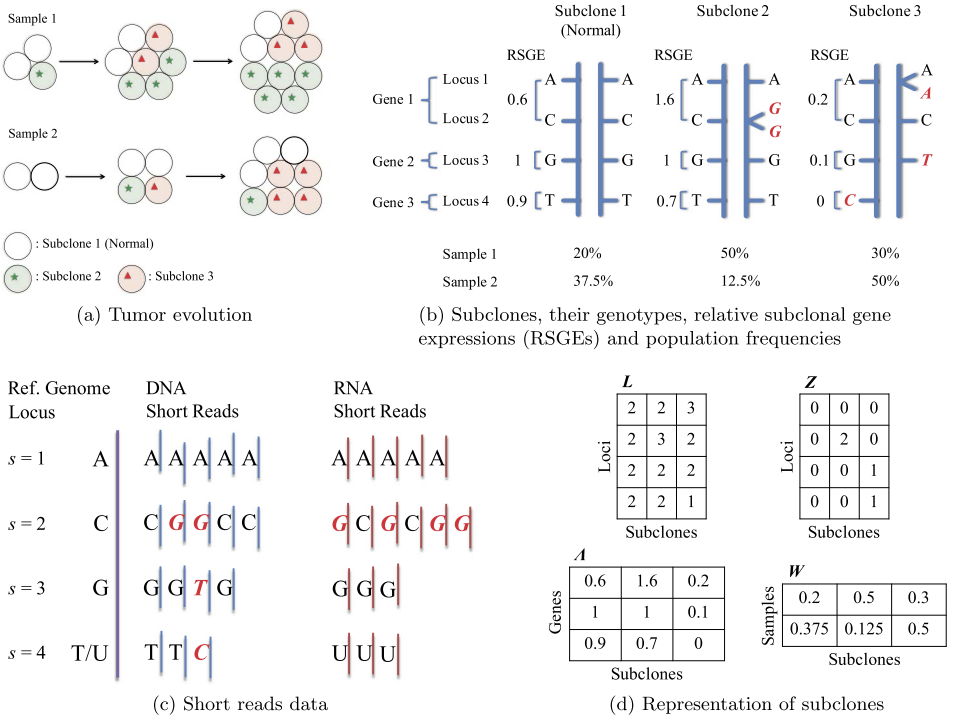
In this paper the problem of *subclone reconstruction* is about the identification of the number, population frequencies, genotypes and gene expression profiles of the subclones

Received March 2020; revised June 2020.
Key words and phrases. Copy number, gene expression, high-throughput sequencing, intra-tumor heterogeneity, latent factor model, somatic mutation.

from the matched DNA and RNA sequencing data. We take an important step beyond the previous work in subclone reconstruction, based on DNA modifications only (Section 1.2), and ask the questions whether and how genetic changes affect transcriptomic landscape in a subclone-specific fashion. Knowledge of tumor subclones is clinically important because it provides information about tumor progression and further suggests personalized treatment strategy (Misale et al. (2012), Landau et al. (2013), Schmitt, Loeb and Salk (2016)). Through a reanalysis of DNA and RNA data from a deceased stage 1 lung cancer patient, we show that subclone reconstruction might have altered the treatment strategy for the patient and might have potentially avoided the unexpected rapid disease progression that led to the fatality event.

The advent of next-generation sequencing (NGS) technology (Mardis (2008)) has enabled researchers to study the genomic landscape of tumor subclones with greater details. In NGS experiments for DNA or RNA molecules, fragmented DNA or cDNA (complementary DNA for RNA sequencing) molecules are extracted from the tumor cells and are sequenced using short or long reads that are mapped to the corresponding loci in the reference genome. We consider experiments in which DNA and RNA from the same tumor samples are sequenced. In particular, for DNA we consider whole-exome sequencing (WES), and for RNA we consider RNA sequencing (RNA-seq) which covers the entire transcriptome. Comprehension of Figure 1 is essential in order to follow the upcoming discussion and requires domain knowledge in cancer biology, bioinformatics and statistics. We display four plots in the figure to simplify the discussion which lay out the biological and statistical problems to be addressed. Figure 1(a) illustrates the biology of tumor subclonal evolution, Figure 1(b) shows the heterogeneous tumor subclones with different somatic point mutations, copy number changes and gene expression levels; Figure 1(c) demonstrates the sequencing data from a tumor sample, and Figure 1(d) shows the statistical quantities (in matrix form) to be inferred that describe the subclone structure. We provide a detailed description of the four plots to motivate our problem next. At the DNA level, point mutations and copy number changes usually drive the subclonal expansion. This is also seen in Figure 1(b), such as the mutation from "C" to "G" at locus 2, with a copy number gain. The *relative subclonal gene expression* (RSGE) refers to the relative expression level of a certain gene in a specific subclone which measures the relative abundance of the mRNA molecules produced by that gene. Here, "relative" means that the expression levels of different genes in different subclones are compared with each other. For example, if the RSGE of gene g_1 in subclone c_1 is two times larger than that of gene g_2 in subclone c_2, gene g_1 in subclone c_1 produces two times more mRNA molecules than gene g_2 in subclone c_2. Figure 1(b) shows hypothetical RSGEs for different genes in different subclones. Figure 1(c) shows that the DNA allele carrying the "G" nucleotide at locus 2 is captured by the short reads which are mapped to the reference genome. In addition, the RNA-seq data show that both "G" and "C" alleles are expressed at this locus. However, this is not always the case. Consider locus $s = 4$. Although there is a point mutation from "T" to "C" in subclone 3 (Figure 1(b)), the RNA short reads only show the "U" allele but not the "C" allele, potentially indicating that the DNA mutation is not transcribed to RNA. Of course, proper modeling is needed to account for the variability of the data which is the main goal of this paper. If done properly, using the short read data in Figure 1(c), the statistical model should provide the four matrices in Figure 1(d) which describes the subclonal architecture in both DNA and RNA level.

NGS data are usually overdispersed and are subject to noise and artifacts. Furthermore, the observed data may be explained by different subclone structures, that is, multiple solutions exist for the same subclone reconstruction problem. Therefore, proper statistical modeling and assumptions are necessary for valid subclone reconstruction, including a propagation of uncertainties that arise from this ambiguity. In this paper we propose a Bayesian approach,



(a) Tumor evolution

(b) Subclones, their genotypes, relative subclonal gene expressions (RSGEs) and population frequencies

(c) Short reads data

(d) Representation of subclones

FIG. 1. *All loci are assumed to reside in exons. Panel* (*a*) *illustrates tumor cell evolution and emergence of three subclones from the original normal cell population. A star or a triangle represents distinct mutations. Panel* (*b*) *shows the genotypes of the three subclones at four loci belonging to three genes as well as their population frequencies in the two samples. The bold italic letters represent somatic mutations including point mutations and CNAs. Panel* (*c*) *shows a number of DNA and RNA short reads mapped to the four loci in panel* (*b*). *The short line segments indicate the alignment of the read relative to the four loci. The bold italic letters indicate the short reads that bear variant sequences. Panel* (*d*) *demonstrates a mathematical representation of the three subclones in panel*(*b*) *with four matrices, where* $\boldsymbol{L}$ *represents copy numbers,* $\boldsymbol{Z}$ *represents variant allele numbers,* $\boldsymbol{\Lambda}$ *represents RSGEs and* $\boldsymbol{W}$ *represents population frequencies.*

named *RNDClone*, to reconstruct tumor subclones by integrating matched DNA and RNA sequencing data from the same tumor samples. We develop sampling models which take into account over-dispersion and noise in the data. We represent subclones as latent factors to capture the coexistence of multiple mutations in the same subclone. Importantly, we allow a simultaneous inference of the RSGEs so that the transcriptional impact of the DNA modifications can be learned. The R package `RNDClone` is provided in the Supplementary Material (Zhou et al. (2020)) and can also be accessed at https://github.com/tianjianzhou/RNDClone.

1.2. *Existing methods.* Numerous methods have been developed for the subclone reconstruction problem. A large portion of these methods are based on DNA sequence data only and deal with intra-tumor heterogeneity, such as THetA (Oesper, Mahmoody and Raphael (2013)), PyClone (Roth et al. (2014)), PyloWGS (Deshwar et al. (2015)), Clomial (Zare et al. (2014)), BayClone2 (Lee et al. (2016)), Cloe (Marass et al. (2016)), PairClone (Zhou et al. (2019a)), TreeClone (Zhou et al. (2019b)) and SIFA (Zeng, Warren and Zhao (2019)). Some methods are based on RNA sequence data only and handle intertumor heterogeneity, including csSAM (Shen-Orr et al. (2010)), CAM (Wang et al. (2016)) and BayCount (Xie, Zhou and Xu (2018)). Yet, inference on intra-tumor heterogeneity based on integrating DNA and RNA data remains an underexplored topic. A few methods (Wilkerson et al. (2014), Radenbaugh et al. (2014)) have proposed integrated analyses for mutation detection using both DNA and RNA data. We take a different perspective. We aim to reconstruct the subclones that have

distinct DNA mutations and RNA expression profiles, potentially better characterizing the impact of subclonal DNA mutations to the downstream RNA expression.

Statistically, we develop an approach based on latent-factor modeling which has been widely used already in many applications (e.g., West (2003), Carvalho et al. (2008), Bhattacharya and Dunson (2011) and Gao, Brown and Engelhardt (2013)). The idea of latent-factor modeling has been applied to infer intra-tumor heterogeneity based on DNA sequence data as seen in Zare et al. (2014), Lee et al. (2016), Marass et al. (2016), Zhou et al. (2019a, 2019b) and Zeng, Warren and Zhao (2019). Interestingly, similar ideas have also been used in RNA subclone inference (Shen-Orr et al. (2010), Wang et al. (2016), Xie, Zhou and Xu (2018)), although the goal there is to infer inter-tumor heterogeneity.

The rest of the paper is structured as follows. In Section 2 we develop a statistical framework for RNDClone, including a sampling model and a prior model. In Section 3 we propose a scheme for posterior inference. In Section 4 we evaluate operating characteristics of RNDClone with two simulation studies. In Section 5 we apply RNDClone to the analysis of three samples from one patient in a lung cancer dataset from The Cancer Genome Atlas (TCGA). Finally, in Section 6 we conclude with a discussion.

2. Statistical model.

2.1. *Notation.* We first introduce some notation to represent the observed data. Suppose T tissue samples are dissected from the same patient, obtained either at different time points, at different spatial locations within the same tumor or at different metastatic sites. Let $s = 1, \ldots, S$ index the loci of the nucleotides (base pairs) that are covered by short reads produced by NGS experiments. The observed data are collected into four $S \times T$ matrices, $\boldsymbol{N} = [N_{st}]$, $\boldsymbol{n} = [n_{st}]$, $\boldsymbol{M} = [M_{st}]$ and $\boldsymbol{m} = [m_{st}]$. We denote the data as $\mathcal{D} = (\boldsymbol{N}, \boldsymbol{n}, \boldsymbol{M}, \boldsymbol{m})$ in short. The values of N_{st}, n_{st}, M_{st} and m_{st} represent the total number of DNA reads, number of DNA reads that bear a variant sequence, total number of RNA reads and number of RNA reads that bear a variant sequence at locus s for sample t, respectively. The total read count is also referred to as read depth, and we refer to the reads that bear a variant sequence as *variant reads* in short. In principle, each read can bear any of the four possible nucleotides, A, C, G and T at any locus. However, it is unlikely to observe more than two sequences across short reads at a single locus, as this would require repeated mutations at the same locus. Therefore, we only distinguish the reads according to whether a read possesses a reference or a variant sequence (compared to the reference genome). For example, in Figure 1(c) a total of $N_{2t} = 5$ DNA reads and $M_{2t} = 6$ RNA reads are mapped to the locus $s = 2$. Among all the reads, there are $n_{2t} = 2$ DNA variant reads and $m_{2t} = 4$ RNA variant reads.

2.2. *Representation of subclones.* Next, we introduce a mathematical representation of subclones. Since we only consider intra-tumor heterogeneity where the T samples are from the same patient, we assume the samples share the same subclones. However, the population frequencies of the same subclone are allowed to vary across different samples. The same assumption is made by most existing methods and is thought to be realistic. Denote by C the number of subclones, where C is unknown and needs to be inferred. Tumor samples are in general not pure, in the sense that they contain some proportions of normal cells. Therefore, among the C subclones we always include a first subclone of normal cells. The normal subclone does not possess any somatic mutation, as in Figure 1(b).

To represent the gene-level RSGE, suppose the S nucleotide loci reside in G genes. We index the genes by $g = 1, \ldots, G$, where $g = g(s) : \{1, \ldots, S\} \mapsto \{1, \ldots, G\}$ represents the gene in which nucleotide s resides. In WES and RNA-seq data, most of the genes possess up to one variant, while a few hypermutated genes possess multiple variants.

We encode the underlying subclone structure in the following four matrices, the construction of which is implicitly conditional on C:

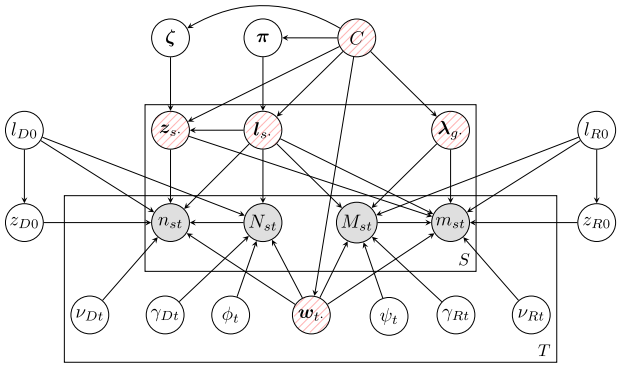


FIG. 2. *Graphical representation of the RNDClone model. Nodes represent random variables, arrows indicate dependencies and plates (rectangles) represent replicates. The shaded nodes are observations, and the striped nodes represent subclone structure and are the quantities of major interest. A "$\cdot$" in the subindex indicates a vector over all subclones $c = 1, \ldots, C$.*

1. An $S \times C$ matrix $\boldsymbol{L} = [l_{sc}]$, with l_{sc} representing the subclonal copy number of locus s in subclone c, $s = 1, \ldots, S$, $c = 1, \ldots, C$. The first subclone (normal cells) does not have any CNA, that is, $l_{s1} = 2$ for all s.
2. An $S \times C$ matrix $\boldsymbol{Z} = [z_{sc}]$, with z_{sc} recording the number of variant alleles for locus s in subclone c, $z_{sc} \le l_{sc}$, $s = 1, \ldots, S$, $c = 1, \ldots, C$. The first subclone (normal cells) does not have any mutation, that is, $z_{s1} = 0$ for all s.
3. A $G \times C$ matrix $\boldsymbol{\Lambda} = [\lambda_{gc}]$, with λ_{gc} representing the RSGE of gene g in subclone c, $g = 1, \ldots, G$, $c = 1, \ldots, C$.
4. A $T \times C$ matrix $\boldsymbol{W} = [w_{tc}]$, where w_{tc} represents the population frequency of subclone c in sample t, $t = 1, \ldots, T$, $c = 1, \ldots, C$. The proportion of tumor cells in a tumor sample, $(1 - w_{t1})$, is called tumor purity.

Using the terminology of latent factor models, $\boldsymbol{L}$, $\boldsymbol{Z}$, $\boldsymbol{\Lambda}$ are essentially factor matrices, and $\boldsymbol{W}$ is the loading matrix. Figure 1(d) demonstrates the mathematical representation of the three subclones in Figure 1(b) using the four matrices. For example, $l_{22} = 3$ represents three copies of the base pair at locus 2 in subclone 2; $z_{22} = 2$ indicates that two out of the three alleles have a variant sequence.

Our goal is to infer the latent quantities of interest, $\boldsymbol{L}$, $\boldsymbol{Z}$, $\boldsymbol{\Lambda}$, $\boldsymbol{W}$ and C, from the observed data, $\mathcal{D} = (\boldsymbol{N}, \boldsymbol{n}, \boldsymbol{M}, \boldsymbol{m})$. Taking a Bayesian approach, the desired inference is achieved by sampling from the posterior distribution

$$p(\boldsymbol{L}, \boldsymbol{Z}, \boldsymbol{\Lambda}, \boldsymbol{W}, C \mid \mathcal{D}) \propto p(\mathcal{D} \mid \boldsymbol{L}, \boldsymbol{Z}, \boldsymbol{\Lambda}, \boldsymbol{W}, C) p(\boldsymbol{L}, \boldsymbol{Z}, \boldsymbol{\Lambda}, \boldsymbol{W}, C),$$

where the sampling model $p(\mathcal{D} \mid \boldsymbol{L}, \boldsymbol{Z}, \boldsymbol{\Lambda}, \boldsymbol{W}, C)$ and the prior model $p(\boldsymbol{L}, \boldsymbol{Z}, \boldsymbol{\Lambda}, \boldsymbol{W}, C)$ are to be specified next.

The structure of the full inference model is summarized in Figure 2, which shows how the sampling distribution for the data $(\boldsymbol{N}, \boldsymbol{n}, \boldsymbol{M}, \boldsymbol{m})$ is indexed by the number of subclones C, the latent structure $(\boldsymbol{L}, \boldsymbol{Z}, \boldsymbol{\Lambda}, \boldsymbol{W})$ that describes the tumor heterogeneity and additional hyperparameters. In particular, in Figure 2 $\boldsymbol{l}_{s\cdot} = (l_{s1}, \ldots, l_{sC})$, $\boldsymbol{z}_{s\cdot} = (z_{s1}, \ldots, z_{sC})$, $\boldsymbol{\lambda}_{g\cdot} = (\lambda_{g1}, \ldots, \lambda_{gC})$, $\boldsymbol{w}_{t\cdot} = (w_{t1}, \ldots, w_{tC})$, and ζ, π, l_{D0}, z_{D0}, ν_{Dt}, γ_{Dt}, ϕ_t, l_{R0}, z_{R0}, ν_{Rt}, γ_{Rt} and ψ_t represent additional hyperparameters, the meaning of which will be more clear later.

2.3. *Sampling model.* Throughout the paper we write $\mathrm{E}(\cdot)$ for expectation and $\mathrm{Var}(\cdot)$ for variance. We denote by $\mathrm{Bin}(\cdot, \cdot)$ a binomial distribution, $\mathrm{Neg\text{-}Bin}(\cdot, \cdot)$ a negative binomial distribution, $\mathrm{Be\text{-}Bin}(\cdot, \cdot)$ a beta-binomial distribution, $\mathrm{Be}(\cdot, \cdot)$ a beta distribution, $\mathrm{Ga}(\cdot, \cdot)$ a

gamma distribution, $\mathrm{Dir}(\cdot,\ldots,\cdot)$ a Dirichlet distribution and $\mathrm{Unif}(\cdot,\cdot)$ a uniform distribution. Lastly, we use subscripts D and R to represent parameters related to DNA and RNA, respectively. We specify the sampling model $p(\mathcal{D} \mid \boldsymbol{L}, \boldsymbol{Z}, \boldsymbol{\Lambda}, \boldsymbol{W}, C)$ as follows.

Sampling model for the total number of DNA reads N_{st}. We start with the sampling model for N_{st}. It is generally assumed that the total number of DNA short reads mapped to a genomic region (i.e., read depth) scales linearly with the number of times the region appears in the DNA sample (i.e., copy number, Magi et al. (2011)). Based on this assumption, previous methods (Klambauer et al. (2012), Lee et al. (2016)) used a Poisson distribution to model N_{st}. The Poisson model relies on the assumption that the reads are randomly and independently sampled from any location of the test genome with equal probability, while, in reality, the distribution of read counts is typically slightly overdispersed (Magi et al. (2011)). To account for this overdispersion of read counts, we model N_{st} using a negative-binomial distribution. Let ϕ_t denote the expected number of DNA reads in sample t if there were no CNAs. We assume

$$N_{st} \mid \phi_t, A_{st}, \gamma_{Dt} \sim \text{Neg-Bin}(\phi_t A_{st}/2, \gamma_{Dt}), \tag{2.1}$$

with

$$\begin{aligned}&\Pr(N_{st} \mid \phi_t, A_{st}, \gamma_{Dt})\\ &\quad= \frac{\Gamma(N_{st}+\gamma_{Dt}^{-1})}{N_{st}!\Gamma(\gamma_{Dt}^{-1})}\left(\frac{1}{1+\gamma_{Dt}\phi_t A_{st}/2}\right)^{1/\gamma_{Dt}}\left(\frac{\gamma_{Dt}\phi_t A_{st}/2}{1+\gamma_{Dt}\phi_t A_{st}/2}\right)^{N_{st}}.\end{aligned}$$

We have $\mathrm{E}(N_{st}) = \phi_t A_{st}/2$ and $\mathrm{Var}(N_{st}) = \mathrm{E}(N_{st}) + \gamma_{Dt}\mathrm{E}(N_{st})^2$, with γ_{Dt} being a dispersion parameter. Here, A_{st} is the average copy number for locus s in sample t and is modeled as

$$A_{st} = w_{t0} l_{D0} + \sum_{c=1}^{C} w_{tc} l_{sc}.$$

Specifically, w_{t0} represents the proportion of a "background" subclone in sample t with no biological meaning. The background subclone is only used as a mathematical tool to account for tiny subclones that are not detectable or cannot be inferred with sufficient statistical power and also for noise and artifacts in the NGS data (sequencing errors, mapping errors, etc.). The term $w_{t0} l_{D0}$ models random noise in the total DNA read counts, where l_{D0} can be viewed as the copy number in the background subclone. We assume the random noise does not differ across different loci, thus l_{D0} does not have an index s.

Sampling model for the number of variant DNA reads n_{st}. Conditional on N_{st}, we model n_{st} as the number of successful trials from a beta-binomial distribution; see, for example, Marass et al. (2016). For each read, let $\tilde{p}_{st}$ denote the probability that it bears a variant sequence. We assume

$$n_{st} \mid N_{st}, \tilde{p}_{st} \sim \mathrm{Bin}(N_{st}; \tilde{p}_{st}),$$

and the success probability $\tilde{p}_{st}$ follows a beta distribution, given by

$$\tilde{p}_{st} \mid p_{st}, \nu_{Dt} \sim \mathrm{Be}\big(\nu_{Dt}^{-1} p_{st}, \nu_{Dt}^{-1}(1-p_{st})\big).$$

Here, $\tilde{p}_{st}$ is centered at the variant allele fraction (VAF) in the cell population, denoted by p_{st} and ν_{Dt} controls the variance of the beta distribution. We define $p_{st} = \tilde{A}_{st}/A_{st}$, where $\tilde{A}_{st}$ is the average number of variant alleles for locus s in sample t, defined as

$$\tilde{A}_{st} = w_{t0} z_{D0} + \sum_{c=1}^{C} w_{tc} z_{sc}.$$

To see this, recall that $\tilde{A}_{st}$ is the average number of variant alleles, and A_{st} is the average number of alleles. Therefore, the ratio $\tilde{A}_{st}/A_{st}$ is the fraction of variant alleles, that is, the VAF. The term $w_{t0}z_{D0}$ is used to account for random noise in variant DNA read counts, and z_{D0} can be viewed as the number of variant alleles in the background subclone. Again, we assume the random noise does not differ across different loci, thus z_{D0} does not have an index s. Integrating out $\tilde{p}_{st}$, n_{st} follows a beta-binomial distribution given by

$$n_{st} \mid N_{st}, p_{st}, \nu_{Dt} \sim \text{Be-Bin}(N_{st}; p_{st}, \nu_{Dt}), \tag{2.2}$$

with

$$\Pr(n_{st} \mid N_{st}, p_{st}, \nu_{Dt}) = \frac{N_{st}!}{n_{st}!(N_{st}-n_{st})!} \times \frac{\Gamma(\nu_{Dt}^{-1})\Gamma(\nu_{Dt}^{-1}p_{st}+n_{st})\Gamma(\nu_{Dt}^{-1}(1-p_{st})+N_{st}-n_{st})}{\Gamma(\nu_{Dt}^{-1}p_{st})\Gamma(\nu_{Dt}^{-1}(1-p_{st}))\Gamma(\nu_{Dt}^{-1}+N_{st})},$$

where $\mathrm{E}(n_{st} \mid N_{st}, p_{st}, \nu_{Dt}) = N_{st}p_{st}$ and $\mathrm{Var}(n_{st} \mid N_{st}, p_{st}, \nu_{Dt}) = N_{st}p_{st}(1-p_{st})(1+N_{st}\nu_{Dt})/(1+\nu_{Dt})$.

Sampling model for the total number of RNA reads M_{st}. Next, we model M_{st}. It follows a similar construction as N_{st}. Suppose that the RSGE of gene g in subclone c is λ_{gc}. We are only concerned about the relative expression levels. For example, $\lambda_{g_1c_1} > \lambda_{g_2c_2}$ means that gene g_1 in subclone c_1 produces more RNA molecules than gene g_2 in subclone c_2. The average number of RNA copies in sample t that contain locus s is thus

$$B_{st} = w_{t0}\lambda_{g(s)0}l_{R0} + \sum_{c=1}^{C} w_{tc}\lambda_{g(s)c}l_{sc}. \tag{2.3}$$

As before, we include $w_{t0}\lambda_{g(s)0}l_{R0}$ to account for random noise in the total RNA read counts. We allow the random noise to be different for DNA and RNA reads by allowing l_{R0} and l_{D0} to be different.

Again, assuming that the total number of RNA reads mapped to a genomic region has a linear relationship with the number of RNA copies of that region, we model M_{st} using a negative-binomial distribution,

$$M_{st} \mid \psi_t, B_{st}, \gamma_{Rt} \sim \text{Neg-Bin}(\psi_t B_{st}/2, \gamma_{Rt}). \tag{2.4}$$

Here, ψ_t is the expected number of RNA reads at locus s in sample t if there were no CNAs at the locus and the RSGE for gene $g(s)$ was 1.

Sampling model for the number of variant RNA reads m_{st}. Finally, we model m_{st} conditional on M_{st}. The average RNA copies in sample t that bear a variant sequence at locus s is given by

$$\tilde{B}_{st} = w_{t0}\lambda_{g(s)0}z_{R0} + \sum_{c=1}^{C} w_{tc}\lambda_{g(s)c}z_{sc},$$

where $w_{t0}\lambda_{g(s)0}z_{R0}$ accounts for random noise in variant RNA read counts. We model m_{st} with a beta-binomial distribution,

$$m_{st} \mid M_{st}, q_{st}, \nu_{Rt} \sim \text{Be-Bin}(M_{st}; q_{st}, \nu_{Rt}), \tag{2.5}$$

where q_{st} is the fraction of RNA copies in sample t that bear a variant sequence at locus s, $q_{st} = \tilde{B}_{st}/B_{st}$.

Generalized factor analysis. Models (2.1), (2.2), (2.4) and (2.5) can be characterized as *generalized latent factor models*, where each observation $(N_{st}, n_{st}, M_{st}, m_{st})$ is modeled by C interpretable and biologically meaningful latent factors, $(l_{sc}, z_{sc}, \lambda_{g(s)c})$ for $c = 1, \ldots, C$, with the w_{tc} coefficients playing the role of factor loadings. In particular,

$$\mathrm{E}\begin{pmatrix} \boldsymbol{N} \\ \boldsymbol{n} \\ \boldsymbol{M} \\ \boldsymbol{m} \end{pmatrix} = f^{-1}\left[\begin{pmatrix} \boldsymbol{L} \\ \boldsymbol{Z} \\ \boldsymbol{\Lambda}^* \circ \boldsymbol{L} \\ \boldsymbol{\Lambda}^* \circ \boldsymbol{Z} \end{pmatrix} \boldsymbol{W}^\top\right],$$

where $\boldsymbol{\Lambda}^*$ is a $S \times C$ matrix with $\lambda^*_{sc} = \lambda_{g(s)c}$, symbol $\circ$ refers to the Hadamard product, that is, entrywise product and f^{-1} is a linear transformation as in equations (2.1), (2.2), (2.4) and (2.5).

Understanding the model as an instance of factor analysis, one might wonder why the factors $\boldsymbol{L}$ and $\boldsymbol{Z}$ should be constrained to integer scores. We strongly prefer the restriction to integer scores to maintain the interpretability of equations (2.1) through (2.5) as mapping the relevant biology. While a relaxation would likely improve mixing of posterior Markov chain Monte Carlo simulation, the loss in ease of communication cannot be justified by minor gains in computational efficiency, considering also that inference need not be carried out in real time.

2.4. *Prior model.* We build a hierarchical prior model for the unknown parameters, including the key quantities of interest, $\boldsymbol{L}$, $\boldsymbol{Z}$, $\boldsymbol{\Lambda}$, $\boldsymbol{W}$ and C.

Prior for C. We start with a truncated geometric prior for C,

$$\Pr(C) \propto (1-\alpha)^{C-1}\alpha, \quad 1 \le C_{\min} \le C \le C_{\max}.$$

The hyperparameter α $(0 < \alpha < 1)$ implies a prior preference for a parsimonious model; larger α represents more shrinkage. We assume that C is a priori restricted to $C_{\min} \le C \le C_{\max}$. The bounds are only used to simplify implementation. Since we only consider heterogeneous cell samples, we can set $C_{\min} = 2$. When computational resources allow, $C_{\max}$ can be set sufficiently large. Empirically, we find that the maximum number of subclones that can be inferred with sufficient statistical power is usually limited by the number of samples and sequencing depth. With the usual sequencing depths in WES experiments, setting $C_{\max} = T + 4$ suffices, where T is the number of tissue samples. This can also be understood from the latent factor model perspective. The number of factors (C) that can be reliably inferred is usually less than the rank of the data matrix which is bounded by the number of samples here.

Prior for $\boldsymbol{L}$. Next, we construct the prior model for the copy numbers $\boldsymbol{L}$. The first column of $\boldsymbol{L}$ represents the normal subclone and is fixed at $l_{s1} = 2$ for all s. For locus s of subclone c $(c \ne 1)$, we assume a truncated geometric-type prior for l_{sc},

$$\Pr(l_{sc} = k \mid \pi_c, C) \propto (1-\pi_c)^{|k-2|}\pi_c, \quad 0 \le K_{\min} \le k \le K_{\max}. \tag{2.6}$$

Here, π_c is the probability for subclone c having a copy number 2 at a locus, and $(1-\pi_c)$ is the probability of a copy number change. We assume a beta prior,

$$\pi_c \mid C \sim \mathrm{Be}(a_\pi, b_\pi).$$

The prior bounds of l_{sc}, $K_{\min}$ and $K_{\max}$ are only used to simplify implementation. Suitable values for $K_{\min}$ and $K_{\max}$ can be explored by running copy number callers on the WES data or comparing the minimum and maximum read depths with the mean read depth. When such information is not available, we recommend setting $K_{\min} = 0$ and choosing a sufficiently large $K_{\max}$.

Prior for $\boldsymbol{Z}$. Conditional on $\boldsymbol{L}$, we define the prior model for the variant allele numbers $\boldsymbol{Z}$, subject to $0 \le z_{sc} \le l_{sc}$. The first column of $\boldsymbol{Z}$ corresponds to the normal subclone and is fixed at $z_{s1} = 0$ for all s. For locus s of subclone c ($c \ne 1$), we assume a truncated geometric prior for z_{sc},

$$\Pr(z_{sc} = k \mid l_{sc}, \zeta_c, C) \propto (1-\zeta_c)^k \zeta_c, \quad 0 \le k \le l_{sc},$$

and $z_{sc} = 0$ if $l_{sc} = 0$. Here, $(1-\zeta_c)$ is the probability of observing a mutation in subclone c at a locus. We put a beta distribution prior on ζ_c,

$$\zeta_c \mid C \sim \mathrm{Be}(a_\zeta, b_\zeta).$$

The prior for $\boldsymbol{L}$ penalizes for deviations from copy number 2. Similarly, the prior for $\boldsymbol{Z}$ penalizes for large numbers of variant alleles. Such prior specifications impose some sparsity structure on the latent factors to improve the identifiability of the latent subclones, especially when the number of samples (T) is small. If desired, the prior models could be modified without changing anything in the rest of the discussion. For example, one may use discrete uniform priors for l_{sc} and z_{sc}.

Prior for $\boldsymbol{\Lambda}$. The RSGEs $\boldsymbol{\Lambda}$ should be nonnegative, $\lambda_{gc} \ge 0$. We assume a gamma prior,

$$\lambda_{gc} \mid C \overset{\text{i.i.d.}}{\sim} \mathrm{Ga}(a_\lambda, b_\lambda).$$

Prior for $\boldsymbol{W}$. The population frequencies of the subclones satisfy $\sum_{c=0}^{C} w_{tc} = 1$ for all t. Recall that w_{t1} stands for the proportion of the normal subclone (subclone 1), that is, (1 − tumor purity). In addition, w_{t0} is the proportion of the "background" subclone that is only used to model tiny subclones and noise. We assume a beta-Dirichlet prior on $\boldsymbol{w}_t$, such that

$$\begin{aligned} &w_{t1} \mid C \sim \mathrm{Be}(a_w, b_w) \quad \text{and} \\ &(w_{t0}, w_{t2}, \ldots, w_{tC})/(1-w_{t1}) \mid w_{t1}, \quad C \sim \mathrm{Dir}(d_0, d, \ldots, d). \end{aligned}$$

We set $d_0 \ll d$ to reflect the nature of the background subclone. Informative prior can be elicited for w_{t1} based on information from some tumor purity caller.

Hyperpriors. We complete the model with priors for l_{D0}, z_{D0}, l_{R0}, z_{R0}, ϕ_t, ψ_t, γ_{Dt}, ν_{Dt}, γ_{Rt}, ν_{Rt} and λ_{g0}. We assume

$$\begin{aligned} &l_{D0} \sim \mathrm{Unif}(K_{\min}, K_{\max}), && z_{D0} \mid l_{D0} \sim \mathrm{Unif}(0, l_{D0}), \\ &l_{R0} \sim \mathrm{Unif}(K_{\min}, K_{\max}), && z_{R0} \mid l_{R0} \sim \mathrm{Unif}(0, l_{R0}), \\ &\phi_t \sim \mathrm{Ga}(a_{\phi_t}, b_{\phi_t}), && \psi_t \sim \mathrm{Ga}(a_{\psi_t}, b_{\psi_t}), \\ &\gamma_{Dt} \overset{\text{i.i.d.}}{\sim} \mathrm{Ga}(a_{\gamma_D}, b_{\gamma_D}), && \nu_{Dt} \overset{\text{i.i.d.}}{\sim} \mathrm{Ga}(a_{\nu_D}, b_{\nu_D}), \\ &\gamma_{Rt} \overset{\text{i.i.d.}}{\sim} \mathrm{Ga}(a_{\gamma_R}, b_{\gamma_R}), && \nu_{Rt} \overset{\text{i.i.d.}}{\sim} \mathrm{Ga}(a_{\nu_R}, b_{\nu_R}). \end{aligned}$$

Among these parameters, informative priors are necessary for ϕ_t and ψ_t. For example, the parameters (ψ_t, λ_{gc}) and $(\psi_t/a, a \cdot \lambda_{gc})$ lead to exactly the same data likelihood for any positive constant a. To help identify the parameters, we can match the prior means of ϕ_t and ψ_t with the mean DNA and RNA read depths in copy number neutral regions, respectively. This can be done by running a bioinformatics copy number caller on the DNA sequencing data. If such information is not available, we recommend setting the priors for ϕ_t and ψ_t such that $\mathrm{E}(\phi_t)$ and $\mathrm{E}(\psi_t)$ are centered at the mean DNA and RNA read depths, respectively. Informative priors can also be elicited for the overdispersion parameters, ϕ_t, ψ_t, γ_{Dt}, ν_{Dt}, γ_{Rt}, ν_{Rt}. For example, the variance of total DNA read count is $\mathrm{Var}(N_{st}) = \mathrm{E}(N_{st}) + \gamma_{Dt}\mathrm{E}(N_{st})^2$.

We can choose the prior for γ_{Dt}, according to the dispersion of N_{st}, in copy number neutral regions. Finally, we set $\lambda_{g0} = (\sum_{c=1}^{C} w_{tc}\lambda_{gc})/(\sum_{c=1}^{C} w_{tc})$ to let the magnitude of the random noise in the total RNA read counts (equation (2.3)) match the average expression level.

The RNDClone hierarchical model is summarized in Figure 2 using a graphical model representation.

2.5. *Special cases.* In special cases, RNDClone reduces to a model based on DNA or RNA data only:

DClone. When only DNA data ($\boldsymbol{N}$ and $\boldsymbol{n}$) are available, we can still infer $\boldsymbol{L}$, $\boldsymbol{Z}$, $\boldsymbol{W}$, C and DNA-related hyperparameters such as ϕ_t, γ_{Dt} and ν_{Dt}. We cannot estimate $\boldsymbol{\Lambda}$ and RNA-related hyperparameters. In this case we refer to our model as DClone. DClone is similar to some existing DNA-based methods such as Zeng, Warren and Zhao (2019) and Lee et al. (2016).

RClone. Typically, RNA-based methods only take as input the total RNA counts $\boldsymbol{M}$ and do not consider variant RNA counts. When only total RNA counts are available, we can still infer $\boldsymbol{\Lambda}$, $\boldsymbol{W}$, C and some RNA-related hyperparameters such as ψ_t and γ_{Rt}. We cannot estimate $\boldsymbol{L}$, $\boldsymbol{Z}$ and DNA-related hyperparameters. Without loss of generality, we can set $l_{sc} = 2$ and $z_{sc} = 0$ for all s and c. In this case we refer to our model as RClone. RClone is similar to some existing RNA-based methods such as Xie, Zhou and Xu (2018), although the goal there is to infer inter-tumor heterogeneity.

3. Posterior inference. Taking a Bayesian approach, inference on the quantities of interest is contained in their posterior distribution. We use Markov chain Monte Carlo (MCMC) simulations to draw J samples $\{\boldsymbol{L}^{(j)}, \boldsymbol{Z}^{(j)}, \boldsymbol{\Lambda}^{(j)}, \boldsymbol{W}^{(j)}, C^{(j)}, \ldots; j = 1, \ldots, J\}$ from the posterior distribution. Transdimensional MCMC and parallel tempering are needed to ensure proper convergence. The exact form of the posterior distribution is described in the Supplementary Material, Section S.1.1 (Zhou et al. (2020)).

Transdimensional MCMC. Let $\boldsymbol{x} = (\boldsymbol{L}, \boldsymbol{Z}, \boldsymbol{\Lambda}, \boldsymbol{W}, \boldsymbol{\pi}, \boldsymbol{\zeta}, \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\gamma}_D, \boldsymbol{\nu}_D, \boldsymbol{\gamma}_R, \boldsymbol{\nu}_R, l_{D0}, z_{D0}, l_{R0}, z_{R0})$ denote all unknown parameters, except the random number of subclones C. Sampling $(\boldsymbol{x}, C)$ involves transdimensional MCMC (Green (1995)), as the dimensions of $\boldsymbol{L}$, $\boldsymbol{Z}$, $\boldsymbol{\Lambda}$, $\boldsymbol{W}$, $\boldsymbol{\pi}$ and $\boldsymbol{\zeta}$ depend on C. Prior to each MCMC transition, denote the current state by $(\boldsymbol{x}, C)$. We propose a new $\tilde{C}$ from $q(\tilde{C} \mid C)$. We use a uniform proposal, $(\tilde{C} \mid C) \sim \text{Unif}(C_{\min}, \ldots, C_{\max})$. Next, we propose a new $\tilde{\boldsymbol{x}}$ whose dimension is consistent with $\tilde{C}$ from $q(\tilde{\boldsymbol{x}} \mid \tilde{C}) = p_\tau(\tilde{\boldsymbol{x}} \mid \tilde{C})$, where

$$p_\tau(\boldsymbol{x} \mid C) \propto p(\boldsymbol{x} \mid C)p(\mathcal{D} \mid \boldsymbol{x}, C)^\tau \quad \text{for } 0 \le \tau \le 1. \tag{3.1}$$

The proposal of $\tilde{\boldsymbol{x}}$ is motivated by the idea of power prior (Ibrahim and Chen (2000)), who used a fraction τ of the historical data likelihood to define an informative prior based on historical data. The acceptance probability of the proposal $(\tilde{\boldsymbol{x}}, \tilde{C})$ is calculated by

$$p_{\text{acc}}(\boldsymbol{x}, C, \tilde{\boldsymbol{x}}, \tilde{C}) = 1 \wedge \frac{p(\mathcal{D} \mid \tilde{\boldsymbol{x}}, \tilde{C})^{1-\tau} \cancel{p_\tau(\tilde{\boldsymbol{x}} \mid \tilde{C})} p(\tilde{C})}{p(\mathcal{D} \mid \boldsymbol{x}, C)^{1-\tau} \cancel{p_\tau(\boldsymbol{x} \mid C)} p(C)} \cdot \frac{\cancel{q(\boldsymbol{x} \mid C)} q(C \mid \tilde{C})}{\cancel{q(\tilde{\boldsymbol{x}} \mid \tilde{C})} q(\tilde{C} \mid C)}.$$

Here, $a \wedge b$ represents the minimum of a and b. In summary, the effect of using $q(\tilde{\boldsymbol{x}} \mid \tilde{C}) = p_\tau(\tilde{\boldsymbol{x}} \mid \tilde{C})$ is to replace the original posterior ratio in the acceptance probability with a fractional $(1 - \tau)$ power likelihood ratio.

The purpose of using the power prior and power likelihood is to achieve reasonable acceptance probabilities for the proposals and a well mixing Markov chain. Since the likelihood is highly informative, commonly used transdimensional proposals, such as the split-merge

proposals (Richardson and Green (1997)), have a very low chance of being accepted and result in very slowly mixing Markov chains. On the other hand, proposals from the power prior are more likely to be accepted. Importantly, the conditional posterior of $\boldsymbol{x}$ given C under this transdimensional MCMC is exactly the same as what under the original model. Details in the Supplementary Material, Section S.1.2 (Zhou et al. (2020)).

In contrast to many existing methods (such as Marass et al. (2016)) which use model selection criteria to choose the optimal C, here, we use transdimensional MCMC to quantify the uncertainty associated with the estimate of C. The marginal posterior of C implies a model comparison among different C's.

Parallel tempering and Gibbs sampler. The previously described transdimensional MCMC scheme requires sampling from $p_\tau(\boldsymbol{x} \mid C)$ in (3.1). We use (separate, up-front) MCMC simulation to generate from $p_\tau(\boldsymbol{x} \mid C)$. However, the posterior surface of $p_\tau(\boldsymbol{x} \mid C)$ is expected to be highly multimodal. We therefore use parallel tempering (Geyer (1991)) to further improve the mixing of the Markov chain. Consider I parallel Markov chains with decreasing temperatures $\{\Delta_1, \Delta_2, \ldots, \Delta_I\}$, where $\Delta_I = 1$. Let $\boldsymbol{x}_i$ denote the state of the ith chain. The target distribution for the ith chain is

$$p_{\tau,i}(\boldsymbol{x}_i \mid C) \propto p(\boldsymbol{x}_i \mid C)\, p(\mathcal{D} \mid \boldsymbol{x}_i, C)^{\tau/\Delta_i},$$

thus the target distribution of the Ith chain is the original target distribution $p_\tau(\boldsymbol{x} \mid C)$. At each MCMC iteration we first independently update all I chains. Gibbs sampling transition probabilities are used to update $\boldsymbol{x}_i$. Details of the full conditionals are in the Supplementary Material, Section S.1.3 (Zhou et al. (2020)). Then, for $i = 1, 2, \ldots, I-1$, we propose a swap between $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i+1}$ and accept the proposal with probability

$$p_{\text{swap}}(\boldsymbol{x}_i, \boldsymbol{x}_{i+1}) = 1 \wedge \left[\frac{p(\mathcal{D} \mid \boldsymbol{x}_{i+1}, C)}{p(\mathcal{D} \mid \boldsymbol{x}_i, C)}\right]^{\frac{1}{\Delta_i} - \frac{1}{\Delta_{i+1}}}.$$

The value from the Ith chain, $\boldsymbol{x}_I$, is kept.

To ensure that the Markov chain of $\boldsymbol{x}$ reaches the stationary distribution, for every possible $C \in \{C_{\min}, \ldots, C_{\max}\}$, we run a sufficiently large number J_0 of burn-in iterations. A draw of $\boldsymbol{x}$ after the burn-in period can be seen as a draw from $p_\tau(\boldsymbol{x} \mid C)$. The detailed MCMC scheme is summarized in the Supplementary Material, Algorithm S.1 (Zhou et al. (2020)).

Point estimate. Suppose we have obtained J posterior samples of $(\boldsymbol{x}, C)$, $\{(\boldsymbol{x}^{(j)}, C^{(j)}), j = 1, \ldots, J\}$. As a point estimate for the number of subclones, C, we report the estimated posterior mode, $\hat{C} = \text{Mode}(\{C^{(j)}, j = 1, \ldots, J\})$. Conditional on $\hat{C}$, we again use the posterior mode as a point estimate for $\boldsymbol{x}$, that is, $\hat{\boldsymbol{x}} = \boldsymbol{x}^{(\hat{j})}$, where

$$\hat{j} = \arg\max_j p\big(\mathcal{D} \mid \boldsymbol{x}^{(j)}, \hat{C}\big)\, p\big(\boldsymbol{x}^{(j)} \mid \hat{C}\big),$$

where the maximization is over all iterations j with $C^{(j)} = \hat{C}$. That is, we report $\hat{\boldsymbol{x}}$ based on the maximum a posteriori (MAP) estimator.

4. Simulation studies. We carry out simulation studies to explore the power of RNDClone to recover under realistic sample size and signal the true number of subclones C, copy numbers $\boldsymbol{L}$, variant allele numbers $\boldsymbol{Z}$, RSGEs $\boldsymbol{\Lambda}$ and cellular proportions $\boldsymbol{W}$. We define subclone reconstruction errors by comparing the point estimates $\hat{C}$, $\hat{\boldsymbol{L}}$, $\hat{\boldsymbol{Z}}$, $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{W}}$ with the simulation truth, similar to Marass et al. (2016). Let $C_{\text{err}} = |\hat{C} - C|$,

$$L_{\text{err}} = \frac{1}{S(C-1)}\left(\sum_{s,c} I\left(\hat{l}_{s\sigma(c)} \neq l_{sc}\right)\right),$$

$$Z_{\text{err}} = \frac{1}{S(C-1)}\left(\sum_{s,c} I\left(\hat{z}_{s\sigma(c)} \neq z_{sc}\right)\right),$$

$\Lambda_{\text{err}} = \sum_{g,c} |\hat{\lambda}^{\text{std}}_{g\sigma(c)} - \lambda^{\text{std}}_{gc}|/(GC)$ and $W_{\text{err}} = \sum_{t,c} |\hat{w}_{t\sigma(c)} - w_{tc}|/(TC)$. Here, σ is a permutation of subclones that minimizes Z_{err} to account for label-switching of subclones. The value $\lambda^{\text{std}}_{gc} = \lambda_{gc}/\text{sd}(\lambda_{gc})$ is the standardized RSGE. The reason for standardizing the RSGEs is to put them on the same scale to allow for comparison among different datasets. In some cases, RNDClone may fail to identify the correct number of subclones C. To avoid comparing two matrices with different dimensions and to ease the calculation of L_{err}, Z_{err}, Λ_{err} and W_{err}, we always report point estimates for $\hat{\boldsymbol{L}}$, $\hat{\boldsymbol{Z}}$, $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{W}}$ conditional on the correct C in the simulation studies.

For all simulation studies we fit the model with the following hyperparameters. We set $\alpha = 0.8$, $a_\pi = C - 1$, $b_\pi = 1$, $a_\zeta = C - 1$, $b_\zeta = 1$, $a_w = 1$, $b_w = 1$, $d = 1$, $d_0 = 0.03$, $a_\lambda = 1$, $b_\lambda = 1$, $b_{\phi_t} = 10$, $b_{\psi_t} = 10$, $a_{\gamma_D} = 1$, $a_{\nu_d} = 1$, $a_{\gamma_R} = 1$ and $a_{\nu_R} = 1$. Following the guidelines that we have introduced in Section 2.4, we set a_{ϕ_t} and a_{ψ_t} such that $\text{E}(\phi_t)$ and $\text{E}(\psi_t)$ match the mean DNA and RNA read depths in copy number neutral regions, respectively; we set b_{γ_D} and b_{ν_d} equal to 10 times the average DNA read depths and b_{γ_D} and b_{ν_d} equal to 10 times the average RNA read depths. We set $C_{\min} = 2$ and $C_{\max} = 7$ as the range of C. Based on empirical calibration, we set the power of the likelihood in the power prior $\tau = 0.99$ (equation (3.1)). We run MCMC simulation for 25,000 burn-in iterations and 5000 transdimensional transitions.

4.1. *Simulation* 1. In simulation 1 we assess RNDClone under multiple scenarios with a range of values for the number of subclones (C), number of samples (T), average read depth ($\text{E}(\phi_t)$ and $\text{E}(\psi_t)$) and maximum copy number ($\max(l_{sc})$).

Simulation 1(a). First, we validate RNDClone on simulated datasets with a range of values for the number of subclones (C) and samples (T). We consider nine scenarios, one for each combination of $C \in \{3, 4, 5\}$ and $T \in \{3, 4, 5\}$. Suppose there are $S = 100$ loci. The average DNA and RNA read depths for sample t in copy number neutral regions are generated from $\phi_t \sim \text{Ga}(200, 1)$ and $\psi_t \sim \text{Ga}(200, 1)$, with $\text{E}(\phi_t) = \text{E}(\psi_t) = 200$. For simplicity, we use, in the simulation truth copy, numbers ranging from 1 to 3, $l_{sc} \in \{1, 2, 3\}$. Accordingly, we set $K_{\min} = 1$ and $K_{\max} = 3$ as prior bounds for l_{sc} (equation (2.6)). Next, typically, data include for each gene only one locus that carries a mutation. Mimicking this, we assume that the 100 loci span $G = 91$ genes, with 84 genes contain one locus, five genes contain two loci and two genes contain three loci. The RSGEs λ_{gc} are randomly generated from a gamma distribution, $\lambda_{gc} \sim \text{Ga}(1, 1)$. Finally, for each scenario, 50 hypothetical datasets $\mathcal{D} = (\boldsymbol{N}, \boldsymbol{n}, \boldsymbol{M}, \boldsymbol{m})$ are generated from the assumed sampling models (2.1), (2.2), (2.4) and (2.5). More details on the simulation parameters are reported in the Supplementary Material, Section S.2.1 (Zhou et al. (2020)).

Table 1 reports the reconstruction errors under the nine scenarios, averaged over the repeatedly simulated datasets. In all nine cases, inference under RNDClone attains low reconstruction errors. The reconstruction errors on C, $\boldsymbol{Z}$ and $\boldsymbol{W}$ can be directly compared to those reported in Marass et al. (2016) (e.g., Figure 4). In general, the errors given by RNDClone are comparable with those in Marass et al. (2016), despite that RNDClone uses a more complex model which considers copy number aberrations and gene expression levels. We note that the reconstruction errors depend on the complexity of the subclone structure, for example, the number of subclones and the similarity among subclones. In general, the simulation truth is better recovered when the number of subclones C is smaller (relative to the number of samples T). It is also easier to distinguish multiple subclones when their genotypes and gene expression profiles are less similar. Table 1 also reports the computation times for RNDClone (using an Intel E5-2680 v4 2.40 GHz processor). Running time increases with the number of samples (T).

TABLE 1

Simulation 1(*a*). *Reconstruction errors and computation times* (*in minutes*) *of RNDClone under nine scenarios, one for each combination of* $C \in \{3, 4, 5\}$ *and* $T \in \{3, 4, 5\}$. *Values shown are averages over* 50 *repeat simulations with numerical Monte Carlo standard errors in subscripts*

C	T	C_{err}	L_{err}	Z_{err}	Λ_{err}	W_{err}	Time
3	3	$0.00_{0.00}$	$0.019_{0.001}$	$0.018_{0.001}$	$0.098_{0.001}$	$0.008_{0.000}$	141
3	4	$0.00_{0.00}$	$0.023_{0.001}$	$0.023_{0.001}$	$0.091_{0.001}$	$0.008_{0.000}$	171
3	5	$0.08_{0.04}$	$0.025_{0.001}$	$0.025_{0.001}$	$0.085_{0.001}$	$0.007_{0.000}$	203
4	3	$0.00_{0.00}$	$0.020_{0.001}$	$0.002_{0.000}$	$0.271_{0.003}$	$0.008_{0.000}$	142
4	4	$0.00_{0.00}$	$0.016_{0.001}$	$0.001_{0.000}$	$0.212_{0.002}$	$0.008_{0.000}$	174
4	5	$0.00_{0.00}$	$0.012_{0.001}$	$0.003_{0.001}$	$0.191_{0.002}$	$0.007_{0.000}$	203
5	3	$0.00_{0.00}$	$0.047_{0.002}$	$0.004_{0.000}$	$0.295_{0.002}$	$0.010_{0.000}$	141
5	4	$0.00_{0.00}$	$0.007_{0.001}$	$0.001_{0.000}$	$0.222_{0.002}$	$0.008_{0.000}$	171
5	5	$0.00_{0.00}$	$0.005_{0.000}$	$0.002_{0.000}$	$0.173_{0.002}$	$0.007_{0.000}$	204

Simulation 1(*b*). Next, we explore how read depth affects the performance of RNDClone. For each of the three simulation scenarios in Simulation 1(a), with $C = 4$ and $T \in \{3, 4, 5\}$, we generate two more hypothetical scenarios with $\mathrm{E}(\phi_t) = \mathrm{E}(\psi_t) \in \{100, 400\}$. Table 2 reports the reconstruction errors under the six scenarios and also includes the errors under the three scenarios in Simulation 1(a) (with $\mathrm{E}(\phi_t) = \mathrm{E}(\psi_t) = 200$) for comparison.

Again, in all scenarios RNDClone maintains low reconstruction errors. In general, the simulation truth is better recovered with higher read depths $\mathrm{E}(\phi_t)$ and $\mathrm{E}(\psi_t)$ and a larger number of samples T. This is similar to existing methods and is well understood in the literature.

Simulation 1(*c*). In Simulations 1(a), (b) we considered simulation truths with copy numbers ranging from one to three. However, some cancer cells can undergo more extensive CNAs. We therefore include an additional simulation scenario with copy numbers ranging from zero to 10. We consider $S = 100$ loci, $C = 4$ subclones, $T = 3$ samples and average read depths $E(\phi_t) = E(\psi_t) = 200$. We generate 50 datasets under this scenario and fit the datasets with RNDClone, setting $K_{\min} = 0$ and $K_{\max} = 10$, accordingly. The second row of Table 3 reports the reconstruction errors for this scenario. Again, the truth is recovered with small errors.

Since $K_{\min}$ and $K_{\max}$ are artificial prior bounds, it is possible that they are different from the true range of copy numbers. To explore how different choices of $K_{\min}$ and $K_{\max}$ affect

TABLE 2

Simulation 1(*b*). *Reconstruction errors and computation times* (*in minutes*) *of RNDClone under nine scenarios, one for each combination of* $T \in \{3, 4, 5\}$ *and* $\mathrm{E}(\phi_t) = \mathrm{E}(\psi_t) \in \{100, 200, 400\}$. *Values shown are averages over* 50 *repeat simulations with numerical Monte Carlo standard errors in subscripts*

T	$\mathrm{E}(\phi_t)$	C_{err}	L_{err}	Z_{err}	Λ_{err}	W_{err}	Time
3	100	$0.00_{0.00}$	$0.066_{0.002}$	$0.008_{0.001}$	$0.323_{0.003}$	$0.013_{0.001}$	141
3	200	$0.00_{0.00}$	$0.020_{0.001}$	$0.002_{0.000}$	$0.271_{0.003}$	$0.008_{0.000}$	142
3	400	$0.00_{0.00}$	$0.006_{0.001}$	$0.003_{0.000}$	$0.221_{0.003}$	$0.006_{0.000}$	142
4	100	$0.00_{0.00}$	$0.063_{0.002}$	$0.006_{0.001}$	$0.278_{0.002}$	$0.010_{0.000}$	171
4	200	$0.00_{0.00}$	$0.016_{0.001}$	$0.001_{0.000}$	$0.212_{0.002}$	$0.008_{0.000}$	174
4	400	$0.00_{0.00}$	$0.004_{0.000}$	$0.003_{0.000}$	$0.157_{0.002}$	$0.006_{0.000}$	174
5	100	$0.00_{0.00}$	$0.045_{0.002}$	$0.004_{0.001}$	$0.244_{0.002}$	$0.010_{0.000}$	202
5	200	$0.00_{0.00}$	$0.012_{0.001}$	$0.003_{0.001}$	$0.191_{0.002}$	$0.007_{0.000}$	203
5	400	$0.12_{0.05}$	$0.026_{0.009}$	$0.025_{0.009}$	$0.172_{0.014}$	$0.014_{0.004}$	206

TABLE 3

Simulation 1(*c*). *Reconstruction errors and computation times* (*in minutes*) *of RNDClone under two scenarios with different copy number ranges. For each scenario, RNDClone is run with* $(K_{\min}, K_{\max}) = (1, 3)$ *and* $(0, 10)$. *Values shown are averages over* 50 *repeat simulations with numerical Monte Carlo standard errors in subscripts*

min, max(l_{sc})	$K_{\min}, K_{\max}$	C_{err}	L_{err}	Z_{err}	Λ_{err}	W_{err}	Time
1, 3	1, 3	$0.00_{0.00}$	$0.020_{0.001}$	$0.002_{0.000}$	$0.271_{0.003}$	$0.008_{0.000}$	142
0, 10	0, 10	$0.00_{0.00}$	$0.105_{0.003}$	$0.038_{0.002}$	$0.303_{0.004}$	$0.012_{0.001}$	224
1, 3	0, 10	$0.00_{0.00}$	$0.025_{0.002}$	$0.003_{0.001}$	$0.272_{0.003}$	$0.009_{0.000}$	220
0, 10	1, 3	$1.14_{0.05}$	$0.552_{0.005}$	$0.231_{0.002}$	$0.559_{0.003}$	$0.061_{0.002}$	142

the performance of RNDClone, we conduct more simulations under two additional scenarios. First, we consider one scenario in Simulation 1(a) with $C = 4$, $T = 3$ and $E(\phi_t) = E(\psi_t) = 200$; we fit the 50 datasets generated under this scenario with $K_{\min} = 0$ and $K_{\max} = 10$, while the true copy numbers range from one to three. Next, we consider the previous scenario in Simulation 1(c) with copy numbers ranging from zero to 10; we fit the 50 datasets generated under this scenario with $K_{\min} = 1$ and $K_{\max} = 3$. The reconstruction errors under these two scenarios are summarized in Table 3 (rows three and four). When the true copy numbers are within the range of $K_{\min}$ and $K_{\max}$, the truth can still be recovered. On the other hand, when the true copy numbers exceed the range of $K_{\min}$ and $K_{\max}$, the reconstruction errors are high. Therefore, we recommend setting $K_{\min} = 0$ and choosing a sufficiently large $K_{\max}$. Increasing the range of $K_{\min}$ and $K_{\max}$ leads to longer computation time.

Lastly, we note that increasing the number of genomic loci or having more loci on each gene allows more borrowing of information and thus also improves the accuracy of subclone reconstruction. Computation time will increase as the number of loci (S) increases.

4.2. *Simulation* 1: *Comparison with alternatives*. There are no existing methods that simultaneously infer subclonal copy numbers, mutations and gene expressions. For comparison, we run SIFA (Zeng, Warren and Zhao (2019)), BayClone2 (Lee et al. (2016)) and BayCount (Xie, Zhou and Xu (2018)) on the same simulated datasets for the inference of DNA or RNA subclones but not both. SIFA and BayClone2 use only DNA sequencing data to infer subclonal copy numbers and mutations, while BayCount uses only total RNA counts to infer subclonal RNA expression profiles.

*SIFA and BayClone*2 (*DNA-based*). SIFA is one of the most recently published methods for DNA-based subclone reconstruction. Both SIFA and BayClone2 characterize subclonal copy numbers and variant allele numbers by latent feature matrices (same as $\boldsymbol{L}$ and $\boldsymbol{Z}$) and model total and variant DNA read counts ($\boldsymbol{N}$ and $\boldsymbol{n}$) using Poisson and binomial distributions, respectively. We run SIFA and BayClone2 under the default hyperparameter settings. The reconstruction errors of SIFA under Simulation 1(a) scenarios are reported in Table 4. Additional details, including the results of SIFA under Simulation 1(b), (c) scenarios, the results of BayClone2, and a discussion of the results, are presented in the Supplementary Material, Section S.2.2 (Zhou et al. (2020)). In all scenarios, SIFA and BayClone2 have higher reconstruction errors than RNDClone.

BayCount (*RNA-based*). BayCount is one of the most recently published methods for RNA-based subclone inference. BayCount was developed to infer intertumor heterogeneity when gene expression data from multiple patients are available. Yet, when multiple tissue samples from the same patient are available, BayCount can also be used to measure the heterogeneity in gene expression across these samples. BayCount characterizes subclonal expression profiles by a latent factor matrix (similar to $\boldsymbol{\Lambda}$) and models total RNA counts ($\boldsymbol{M}$) using a

TABLE 4

Reconstruction errors of SIFA and BayCount under the nine scenarios in Simulation 1(*a*). *Values shown are averages over* 50 *repeat simulations with numerical Monte Carlo standard errors in subscripts*

		SIFA				BayCount		
C	T	C_{err}	L_{err}	Z_{err}	W_{err}	C_{err}	Λ_{err}	W_{err}
3	3	$0.10_{0.04}$	$0.070_{0.004}$	$0.073_{0.009}$	$0.019_{0.002}$	$0.54_{0.08}$	$0.209_{0.007}$	$0.094_{0.003}$
3	4	$0.32_{0.07}$	$0.066_{0.003}$	$0.065_{0.003}$	$0.014_{0.001}$	$0.16_{0.07}$	$0.177_{0.004}$	$0.083_{0.002}$
3	5	$0.39_{0.07}$	$0.102_{0.009}$	$0.136_{0.020}$	$0.030_{0.005}$	$0.90_{0.05}$	$0.150_{0.002}$	$0.065_{0.001}$
4	3	$1.42_{0.10}$	$0.126_{0.004}$	$0.192_{0.008}$	$0.088_{0.004}$	$0.78_{0.11}$	$0.516_{0.004}$	$0.080_{0.001}$
4	4	$2.57_{0.12}$	$0.125_{0.005}$	$0.197_{0.010}$	$0.081_{0.004}$	$0.84_{0.08}$	$0.477_{0.012}$	$0.080_{0.003}$
4	5	$1.54_{0.08}$	$0.108_{0.005}$	$0.204_{0.011}$	$0.081_{0.005}$	$1.74_{0.08}$	$0.425_{0.009}$	$0.065_{0.002}$
5	3	$0.44_{0.08}$	$0.073_{0.004}$	$0.052_{0.010}$	$0.038_{0.005}$	$1.14_{0.08}$	$0.538_{0.004}$	$0.106_{0.001}$
5	4	$0.86_{0.09}$	$0.057_{0.006}$	$0.087_{0.012}$	$0.048_{0.005}$	$1.12_{0.12}$	$0.453_{0.009}$	$0.092_{0.003}$
5	5	$0.74_{0.07}$	$0.068_{0.005}$	$0.112_{0.010}$	$0.059_{0.005}$	$1.32_{0.17}$	$0.472_{0.007}$	$0.103_{0.003}$

negative binomial distribution. We run BayCount under its default hyperparameter setting. We rescale the outcome of BayCount and adjust for copy numbers so that it can be compared to $\boldsymbol{\Lambda}$. The reconstruction errors of BayCount under Simulation 1(a) scenarios are reported in Table 4. Additional results are presented in the Supplementary Material, Section S.2.2 (Zhou et al. (2020)). Overall, BayCount appears to have inflated errors compared with RNDClone.

Selfcomparison. We present a selfcomparison of our own modeling approach applied separately to DNA and RNA data. As described in Section 2.5, when only DNA data (or RNA data) are available, RNDClone reduces to DClone (or RClone) and can still infer DNA-related (or RNA-related) parameters, such as $\boldsymbol{L}$, $\boldsymbol{Z}$, $\boldsymbol{W}$ and C (or $\boldsymbol{\Lambda}$, $\boldsymbol{W}$ and C). The goal of this comparison is to better understand the advantage of jointly modeling DNA and RNA data.

The reconstruction errors of DClone and RClone under the nine scenarios in Simulation 1(a) are reported in Table 5. Additional results are presented in the Supplementary Material, Section S.2.2 (Zhou et al. (2020)). The reconstruction errors of RClone are generally higher than those of RNDClone. Although the reconstruction errors of DClone are comparable to those of RNDClone, using only DNA data, DClone does not provide estimations of the subclonal RNA expression profiles. One might think of combining the results from separate analyses post hoc to obtain a complete picture of the subclonal genotypes and gene expression profiles. However, such combination might not be straightforward, as DNA analysis and RNA analysis might yield different numbers of subclones and different subclone proportions, and the relationship between DNA and RNA subclones is also unknown. In contrast, inference under a joint model for DNA and RNA data provides coherent inference and easily interpretable results.

4.3. *Simulation* 2: *Frequentist coverage.* In simulation 2 we generate simulated data by mimicking the actual TCGA lung cancer dataset in Section 5. Same as the lung cancer dataset, we consider $S = 66$ loci and $T = 3$ samples. These loci reside on $G = 61$ genes. The number, genotypes and gene expression profiles of the subclones as well as the expected read depths and overdispersion parameters are generated by fitting the lung cancer dataset under RNDClone; see Section 5 for more detail of the lung cancer dataset. In particular, the number of subclones is $C = 3$, and the copy numbers are $l_{sc} \in \{0, 1, \ldots, 13\}$. Under this simulation truth, we repeatedly generate 50 datasets with the assumed sampling model to investigate the frequentist coverage properties of the Bayesian credible intervals given by RNDClone. This exercise gives us an assessment of the potential performance of RNDClone for the real data in Section 5.

TABLE 5

Reconstruction errors of DClone and RClone under the nine scenarios in Simulation 1(*a*). *Values shown are averages over* 50 *repeat simulations with numerical Monte Carlo standard errors in subscripts*

		DClone				RClone		
C	T	C_{err}	L_{err}	Z_{err}	W_{err}	C_{err}	Λ_{err}	W_{err}
3	3	$0.00_{0.00}$	$0.018_{0.001}$	$0.017_{0.001}$	$0.014_{0.000}$	$0.00_{0.00}$	$0.096_{0.001}$	$0.021_{0.001}$
3	4	$0.00_{0.00}$	$0.022_{0.001}$	$0.021_{0.001}$	$0.014_{0.000}$	$0.00_{0.00}$	$0.093_{0.001}$	$0.023_{0.001}$
3	5	$0.04_{0.03}$	$0.027_{0.001}$	$0.027_{0.001}$	$0.014_{0.000}$	$0.00_{0.00}$	$0.083_{0.001}$	$0.021_{0.001}$
4	3	$0.00_{0.00}$	$0.024_{0.002}$	$0.002_{0.000}$	$0.010_{0.000}$	$1.00_{0.00}$	$0.422_{0.004}$	$0.045_{0.002}$
4	4	$0.00_{0.00}$	$0.019_{0.001}$	$0.001_{0.000}$	$0.011_{0.000}$	$1.00_{0.00}$	$0.295_{0.004}$	$0.037_{0.002}$
4	5	$0.02_{0.02}$	$0.015_{0.001}$	$0.002_{0.001}$	$0.011_{0.000}$	$1.00_{0.00}$	$0.266_{0.002}$	$0.033_{0.002}$
5	3	$0.00_{0.00}$	$0.064_{0.003}$	$0.021_{0.004}$	$0.028_{0.002}$	$2.00_{0.00}$	$0.481_{0.003}$	$0.053_{0.002}$
5	4	$0.00_{0.00}$	$0.008_{0.001}$	$0.001_{0.000}$	$0.018_{0.000}$	$1.00_{0.00}$	$0.305_{0.003}$	$0.035_{0.002}$
5	5	$0.00_{0.00}$	$0.007_{0.001}$	$0.003_{0.000}$	$0.019_{0.000}$	$1.00_{0.00}$	$0.231_{0.002}$	$0.026_{0.001}$

We fit each simulated dataset with RNDClone using the same hyperparameter and MCMC setting described in the beginning of Section 4. Moreover, we set $K_{\min} = 0$ and $K_{\max} = 15$. The results are summarized in Table 6. The coverage probabilities are calculated as the proportion of the time that the 95% credible interval of the corresponding parameter contains the truth. For the multidimensional parameters $\boldsymbol{L}$, $\boldsymbol{Z}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{W}$, the reported coverage probabilities are averages over all of their entries. The frequentist coverage does not give rise to concerns. An exact match is not expected due to a relatively small sample size ($T = 3$) and a large parameter space. In addition, we report the reconstruction errors and root mean squared errors (RMSEs) averaged over the repeated datasets. Again, for $\boldsymbol{L}$, $\boldsymbol{Z}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{W}$, the RMSEs are averages over all of their entries. For example,

$$L_{\text{RMSE}} = \frac{1}{S(C-1)} \sum_{s,c} \sqrt{\frac{1}{J} \sum_{j} \bigl(l_{s\sigma(c)}^{(j)} - l_{sc}\bigr)^2}.$$

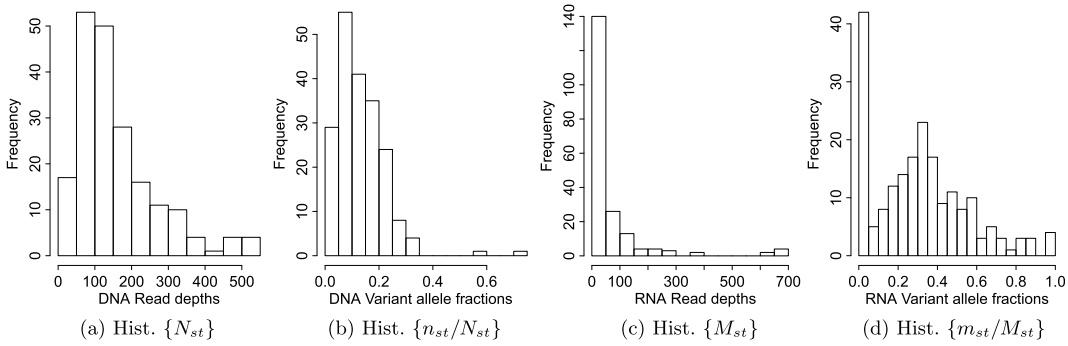
The reconstruction errors and RMSEs are small, indicating good recovery of the truth.

5. TCGA data analysis. We apply RNDClone to the analysis of a TCGA (The Cancer Genome Atlas) lung adenocarcinoma (LUAD) dataset. We have $T = 3$ tumor samples from a stage 1B T2 patient with lung adenocarcinoma (sample ID: TCGA-44-2668) who has passed away two years after diagnosis. In particular, all of the three samples are obtained from the primary tumor tissue and are sequenced using whole exome sequencing (WES) technology. We first retrieve the DNA sequence data by downloading the VCF (variant call format) files from the TCGA website. In particular, we choose the somatic variant callset generated by

TABLE 6

Simulation 2. *Frequentist coverage probabilities, reconstruction errors and root mean squared errors* (*RMSEs*) *for* C, $\boldsymbol{L}$, $\boldsymbol{Z}$, $\boldsymbol{\Lambda}$, $\boldsymbol{W}$ *and the log-likelihood. For the multidimensional parameters the values are averages over all entries*

	C	$\boldsymbol{L}$	$\boldsymbol{Z}$	$\boldsymbol{\Lambda}$	$\boldsymbol{W}$	log-likelihood
Coverage	100%	95.7%	98.0%	85.4%	76.2%	92%
Rec. error	0	0.292	0.065	0.261	0.023	–
RMSE	0.714	0.682	0.133	0.319	0.024	–



(a) Hist. $\{N_{st}\}$ (b) Hist. $\{n_{st}/N_{st}\}$ (c) Hist. $\{M_{st}\}$ (d) Hist. $\{m_{st}/M_{st}\}$

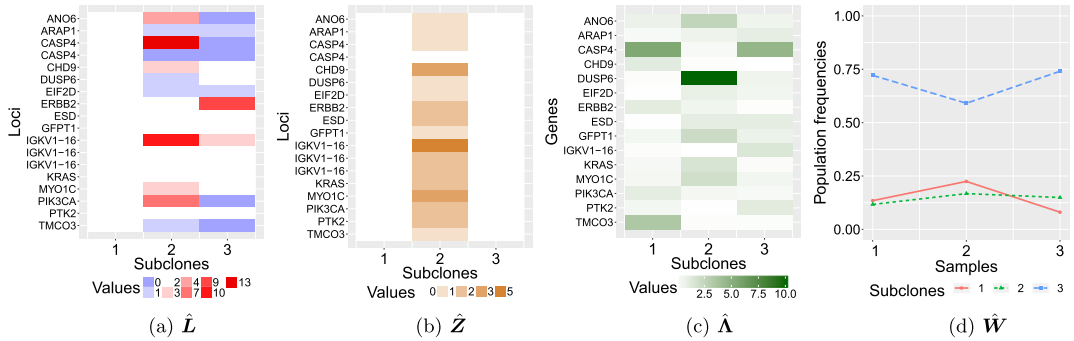
FIG. 3. *Some summary plots of the TCGA LUAD dataset. Histogram of DNA read depths* (*a*), *DNA variant allele fractions* (*b*), *RNA read depths* (*c*) *and RNA variant allele fractions* (*d*).

MuTect2 (Cibulskis et al. (2013)). The average sequencing depth is around 100×. Next, to obtain the RNA sequence data, we download the corresponding RNA BAM (binary alignment map) files which contain sorted and indexed RNA reads. Finally, we find the loci for which both DNA and RNA read data are available for the three samples. For each locus we record the total numbers of mapped DNA and RNA reads as well as numbers of mapped DNA and RNA reads that bear a variant sequence. We obtain a total of $S = 66$ loci across $G = 61$ genes. Figure 3 shows the histograms of DNA and RNA read depths and variant allele fractions. The average DNA read depths for the three samples are 141, 210 and 122, respectively. As shown in the simulation studies, RNDClone should provide useful inference with $T = 3$ samples and 100×–200× read depth.

We fit the dataset with the same hyperparameters described in the beginning of Section 4. To calibrate an informative prior for ϕ_t, for each sample we retrieve copy number estimates of specific DNA segments obtained from single nucleotide polymorphism (SNP) array analysis, using the TCGA SNP array data for the three samples. We then run a regression between the copy number estimates and the average read depths for the DNA segments. Lastly, we match the prior mean and variance of the average read depth in copy number neutral regions (i.e., copy number $= 2$) with the values estimated from the regression analysis. We run MCMC simulation with 25,000 burn-in iterations and 5000 transdimensional transitions. The computation takes around 150 minutes using a single core on an Intel Xeon (E5-2650 v4, 2.20 GHz) computing node.

Results. The posterior mode of C is $\hat{C} = 3$ (posterior distribution of C is shown in the Supplementary Material, Figure S.3, Zhou et al. (2020)). That is, we infer that the tumor samples have three subclones (including one normal subclone). The estimated population frequencies of the three subclones across the three samples are shown in Figure 4(d). For the upcoming biological interpretation we focus on 15 genes that have high expression levels in at least one subclone. Figures 4(a), (b) show the estimated copy numbers and variant allele numbers for the loci that reside on the 15 genes, and Figure 4(c) shows the estimated RSGEs for the 15 genes. The complete estimates are reported in the Supplementary Material, Figure S.4 (Zhou et al. (2020)). A convergence diagnostic for the MCMC simulation and a check of model fit are reported later.

From Figure 4, we can see that RNDClone identifies two tumor subclones (subclones 2 and 3), aside from the normal cells (subclone 1). Subclone 3 exhibits copy number losses (Figure 4(a)) in many genes but no point mutations (Figure 4(b)). It has the largest population frequencies (Figure 4(d)) in all three samples. However, subclone 2 shows mostly copy number gains (Figure 4(a)) and extensive mutations in almost all the genes (Figure 4(b)).



(a) $\hat{\boldsymbol{L}}$ (b) $\hat{\boldsymbol{Z}}$ (c) $\hat{\boldsymbol{\Lambda}}$ (d) $\hat{\boldsymbol{W}}$

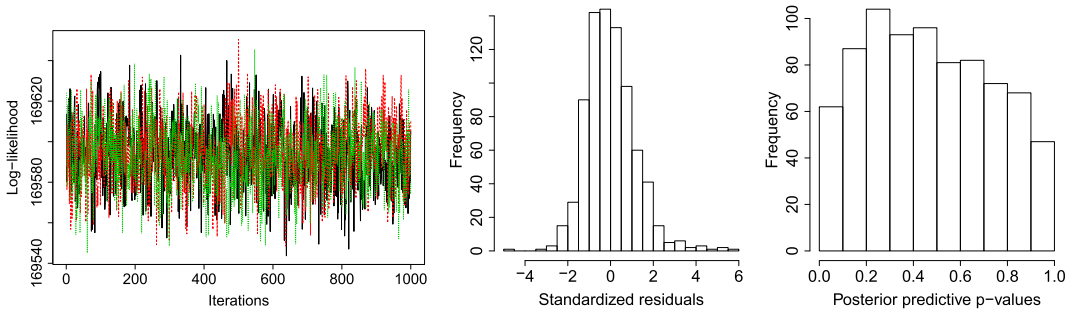
FIG. 4. *Posterior inference for the TCGA LUAD dataset using RNDClone. Estimated copy numbers*: (*a*) *and variant allele numbers*, (*b*) *for the loci that reside on the* 15 *selected genes and estimated RSGEs* (*c*) *for the* 15 *selected genes in the form of heatmaps. The color keys for the corresponding heatmaps are included in the bottom of each figure. Panel* (*d*) *shows the estimated population frequencies of the three subclones across the three samples.*

Its population frequencies are quite small as shown in Figure 4(d), indicating that this is a hypermutated small subclone.

The lung cancer patient was initially diagnosed with stage 1 disease and was declared a complete response, that is, no detectable tumor after a surgery. However, merely one year after the surgery, the patient relapsed with new tumor and died quickly after failing conventional treatment. This is unusual for a stage 1B patient. A possible explanation is that subclone 2 might have been harboring the fatal tumor cells that caused the disease relapse and death. The subclone-2 tumor could be highly malignant since most genes that we analyzed were mutated (Supplementary Material, Figure S.4, Zhou et al. (2020)), including the well-known oncogene KRAS that usually leads to poor prognosis. In addition, Figure 4(c) shows that the RSGEs of subclone 2 are distinct from those of subclones 1 and 3, indicating that the hypermutated subclone 2 also alters the transcription profile of the cells. We speculate that subclone analysis (such as using RNDClone) could have been helpful for this patient if the analysis had been performed and subclone 2 had been identified. Perhaps additional chemotherapy would have been ordered by the treating clinicians after the surgery, or the patient might have been followed more frequently for potential relapse. Sadly, without knowing the malignant subclone 2, the stage 1B tumor, considered as early stage, could still be lethal.

Convergence diagnostic and test of fit. To assess the convergence of the MCMC simulation, we run RNDClone with three different starting values (random variate seeds). As a convenient scalar summary to track for the convergence assessment, we use the log-likelihood values. Figure 5(a) shows the traceplot of the posterior samples of the log-likelihood from the three runs. To avoid complications related to the transdimensional nature of the MCMC implementation, we first select from each of the three runs 1000 iterations when the chain was imputing $C = 3$ subclones. We then evaluated diagnostics with these subchains. We evaluate the potential scale reduction factor (Gelman and Rubin (1992)) across the three chains. We find a factor of 1.01, with an upper confidence bound 1.05, indicating no evidence for lack of convergence.

Next, as an informal check of model fit, we inspect the histograms of standardized residuals, defined as $[a - \mathrm{E}(a)]/\sqrt{\mathrm{Var}(a)}$, where a is N_{st}, n_{st}, M_{st} or m_{st} and $\mathrm{E}(a)$ and $\mathrm{Var}(a)$ are functions of the parameters. The residuals are then calculated and averaged over all posterior draws of the parameters. Figure 5(b) shows the histogram of the standardized residuals. The residuals are centered around zero with little mass beyond ± 2, indicating a good model fit. We also use posterior predictive checking to examine the goodness-of-fit of the model; see, for example, Gelman et al. (2014) (Section 6.3) for a review. Using the J posterior



(a) Traceplot of the log-likelihood (b) Hist. of residuals (c) Hist. of p_{pp}

FIG. 5. *Panel* (*a*) *shows the traceplot of the log-likelihood from three independent Markov chains from fitting the TCGA LUAD dataset with RNDClone using different random seeds. Panel* (*b*) *is a histogram of standardized residuals. Panel* (*c*) *is a histogram of posterior predictive* p*-values for the multidimensional data.*

draws of the parameters, we generate replicated data $\{\mathcal{D}^{\text{rep}(j)} = (\boldsymbol{N}^{\text{rep}(j)}, \boldsymbol{n}^{\text{rep}(j)}, \boldsymbol{M}^{\text{rep}(j)}, \boldsymbol{m}^{\text{rep}(j)}), j = 1, \ldots, J\}$ from the RNDClone model. Since the data are multidimensional, for each entry a (a can be N_{st}, n_{st}, M_{st} or m_{st}) we calculate the posterior predictive p-value by $p_{\text{pp}} = \frac{1}{J}\sum_j I(a^{\text{rep}(j)} > a)$. Figure 5(c) shows the histogram of the posterior predictive p-values for all entries of the data. In particular, 3.66% of the p-values are less than 0.05, and 2.40% of the p-values are greater than 0.95 which does not give rise to concerns about model fit.

Comparison with alternatives. For comparison, we run SIFA and BayCount on the same dataset. The results under SIFA and BayCount as well as a discussion of the results are presented in the Supplementary Material, Section S.3.2 (Zhou et al. (2020)).

6. Discussion. We have developed novel generalized latent factor models to reconstruct tumor subclones based on integrative analyses of both DNA and RNA sequence data. This is the first attempt to bridge existing DNA-based and RNA-based methods. We simultaneously infer subclonal genotype and gene expression profile. Such inference provides important clinical information about personalized treatment strategies.

Our modeling approach can be considered as a finite truncation (in C) of related Bayesian nonparametric models (Müeller, Quintana and Page (2018)). The truncation is used to facilitate posterior simulation, following, for example, the idea of truncated stick-breaking priors in Ishwaran and James (2001). Alternative split-merge proposals (Jain and Neal (2004), Griffiths and Ghahramani (2011)) in our application would be adding or deleting columns in $\boldsymbol{L}$, $\boldsymbol{Z}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{W}$. We found that such proposals lead to very low acceptance probabilities in our application due to the highly peaked and multimodal posterior landscape. We therefore developed the transdimensional and parallel-tempering MCMC algorithm to efficiently sample from the posterior space.

We have demonstrated the practical application of RNDClone on one real-world dataset and found some interesting results. Currently, publicly available synthetic or real-world datasets with both DNA and RNA sequencing counts are still rare. When more of such datasets become available, we plan to apply RNDClone to these datasets to further evaluate the validity of RNDClone.

It is generally assumed that tumor cells evolve along a phylogenetic tree. Each cell inherits the mutations of its parent and, possibly, also gains more mutations. An important extension of the current model is to explicitly model the potential phylogenetic relationship among the subclones, such as in Deshwar et al. (2015), Marass et al. (2016), Zeng, Warren and Zhao (2019) and Zhou et al. (2019b). However, modeling tumor phylogeny is very challenging

when CNA data are included. Existing methods either assume copy number neutrality, require copy number estimates from another software, or make strong assumptions about the occurrence of CNAs. Therefore, like many existing methods (Roth et al. (2014), Zare et al. (2014), Lee et al. (2016)), we chose not to explicitly model phylogeny. In simulation studies we find that RNDClone is able to recover an assumed true phylogenetic structure when it is included in the simulation scenario.

Currently, RNDClone characterizes tumor heterogeneity by point mutations, copy number aberrations and gene expression levels. A future direction is to incorporate data from other types of structural variations (Fan et al. (2014)) as well as DNA methylation data (Rhee et al. (2013)) in the analysis. It is also of interest to integrate further downstream gene expression data such as proteomics data into the current scheme. The focus of RNDClone is inference on intra-tumor heterogeneity. Nevertheless, the proposed methodology can be easily extended to infer intertumor heterogeneity across different patients. Lastly, RNDClone utilizes NGS data from bulk sequencing experiments. Alternatively, single-cell sequencing data (Schmidt and Efferth (2016), Kuipers, Jahn and Beerenwinkel (2017)) provide genomic and transcriptomic information at the cellular level. It is of interest to develop similar methodologies for single-cell data, and there is some very recent progress (Campbell et al. (2019)).

Acknowledgments. We thank the Editor, the Associate Editor and two anonymous reviewers for their invaluable comments which have greatly improved the quality of the paper.

SUPPLEMENTARY MATERIAL

Supplement to "RNDClone: Tumor subclone reconstruction based on integrating DNA and RNA sequence data" (DOI: 10.1214/20-AOAS1368SUPPA; .pdf). Supplementary details referenced in the main text, including details of the MCMC implementation, simulation, and TCGA data analysis.

Source code for "RNDClone: Tumor subclone reconstruction based on integrating DNA and RNA sequence data" (DOI: 10.1214/20-AOAS1368SUPPB; .zip). Source code for the R package `RNDClone`, and data files for the simulation studies and TCGA data analysis.

REFERENCES

BHATTACHARYA, A. and DUNSON, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. MR2806429 https://doi.org/10.1093/biomet/asr013

CAMPBELL, K. R., STEIF, A., LAKS, E., ZAHN, H., LAI, D., MCPHERSON, A., FARAHANI, H., KABEER, F., O'FLANAGAN, C. et al. (2019). Clonealign: Statistical integration of independent single-cell RNA and DNA sequencing data from human cancers. *Genome Biol*. **20** 54.

CARVALHO, C. M., CHANG, J., LUCAS, J. E., NEVINS, J. R., WANG, Q. and WEST, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc*. **103** 1438–1456. MR2655722 https://doi.org/10.1198/016214508000000869

CIBULSKIS, K., LAWRENCE, M. S., CARTER, S. L., SIVACHENKO, A., JAFFE, D., SOUGNEZ, C., GABRIEL, S., MEYERSON, M., LANDER, E. S. et al. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol*. **31** 213–219.

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 Biol*. **16** 35. https://doi.org/10.1186/s13059-015-0602-8

FAN, X., ZHOU, W., CHONG, Z., NAKHLEH, L. and CHEN, K. (2014). Towards accurate characterization of clonal heterogeneity based on structural variation. *BMC Bioinform*. **15** 299.

GAO, C., BROWN, C. D. and ENGELHARDT, B. E. (2013). A latent factor model with a mixture of sparse and dense factors to model gene expression data with confounding effects. Preprint. Available at arXiv:1310.4792.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci*. **7** 457–472.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2014). *Bayesian Data Analysis*, 3rd ed. *Texts in Statistical Science Series*. CRC Press, Boca Raton, FL. MR3235677

GEYER, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics*, *Proceedings of the* 23*rd Symposium on the Interface* 156–163. Interface Foundation of North America, Fairfax Station, VA.

GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810 https://doi.org/10.1093/biomet/82.4.711

GRIFFITHS, T. L. and GHAHRAMANI, Z. (2011). The Indian buffet process: An introduction and review. *J. Mach. Learn. Res.* **12** 1185–1224. MR2804598

HEPPNER, G. H. (1984). Tumor heterogeneity. *Cancer Res.* **44** 2259–2265.

IBRAHIM, J. G. and CHEN, M.-H. (2000). Power prior distributions for regression models. *Statist. Sci.* **15** 46–60. MR1842236 https://doi.org/10.1214/ss/1009212673

ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173. MR1952729 https://doi.org/10.1198/016214501750332758

JAIN, S. and NEAL, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *J. Comput. Graph. Statist.* **13** 158–182. MR2044876 https://doi.org/10.1198/1061860043001

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** e69. https://doi.org/10.1093/nar/gks003

KUIPERS, J., JAHN, K. and BEERENWINKEL, N. (2017). Advances in understanding tumour evolution through single-cell sequencing. *Biochimica et Biophysica Acta* (*BBA*)*-Reviews on Cancer* **1867** 127–138.

LANDAU, D. A., CARTER, S. L., STOJANOV, P., MCKENNA, A., STEVENSON, K., LAWRENCE, M. S., SOUGNEZ, C., STEWART, C., SIVACHENKO, A. et al. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152** 714–726.

LEE, J., MÜLLER, P., SENGUPTA, S., GULUKOTA, K. and JI, Y. (2016). Bayesian inference for intratumour heterogeneity in mutations and copy number variation. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **65** 547–563. MR3522952 https://doi.org/10.1111/rssc.12136

MAGI, A., TATTINI, L., PIPPUCCI, T., TORRICELLI, F. and BENELLI, M. (2011). Read count approach for DNA copy number variants detection. *Bioinformatics* **28** 470–478.

MARASS, F., MOULIERE, F., YUAN, K., ROSENFELD, N. and MARKOWETZ, F. (2016). A phylogenetic latent feature model for clonal deconvolution. *Ann. Appl. Stat.* **10** 2377–2404. MR3592061 https://doi.org/10.1214/16-AOAS986

MARDIS, E. R. (2008). Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9** 387–402. https://doi.org/10.1146/annurev.genom.9.081307.164359

MISALE, S., YAEGER, R., HOBOR, S., SCALA, E., JANAKIRAMAN, M., LISKA, D., VALTORTA, E., SCHIAVO, R., BUSCARINO, M. et al. (2012). Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature* **486** 532–536.

MÜELLER, P., QUINTANA, F. A. and PAGE, G. (2018). Nonparametric Bayesian inference in applications. *Stat. Methods Appl.* **27** 175–206. MR3807363 https://doi.org/10.1007/s10260-017-0405-z

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. https://doi.org/10.1186/gb-2013-14-7-r80

RADENBAUGH, A. J., MA, S., EWING, A., STUART, J. M., COLLISSON, E. A., ZHU, J. and HAUSSLER, D. (2014). RADIA: RNA and DNA integrated analysis for somatic mutation detection. *PLoS ONE* **9** e111516.

RHEE, J.-K., KIM, K., CHAE, H., EVANS, J., YAN, P., ZHANG, B.-T., GRAY, J., SPELLMAN, P., HUANG, T. H.-M. et al. (2013). Integrated analysis of genome-wide DNA methylation and gene expression profiles in molecular subtypes of breast cancer. *Nucleic Acids Res.* **41** 8464–8474.

RICHARDSON, S. and GREEN, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59** 731–792. MR1483213 https://doi.org/10.1111/1467-9868.00095

ROTH, A., KHATTRA, J., YAP, D., WAN, A., LAKS, E., BIELE, J., HA, G., APARICIO, S., BOUCHARD-CÔTÉ, A. et al. (2014). PyClone: Statistical inference of clonal population structure in cancer. *Nat. Methods* **11** 396–398.

SCHMIDT, F. and EFFERTH, T. (2016). Tumor heterogeneity, single-cell sequencing, and drug resistance. *Pharmaceuticals* (*Basel*) **9** 33.

SCHMITT, M. W., LOEB, L. A. and SALK, J. J. (2016). The influence of subclonal resistance mutations on targeted cancer therapy. *Nature Reviews Clinical Oncology* **13** 335–347.

SHACKLETON, M., QUINTANA, E., FEARON, E. R. and MORRISON, S. J. (2009). Heterogeneity in cancer: Cancer stem cells versus clonal evolution. *Cell* **138** 822–829.

SHEN-ORR, S. S., TIBSHIRANI, R., KHATRI, P., BODIAN, D. L., STAEDTLER, F., PERRY, N. M., HASTIE, T., SARWAL, M. M., DAVIS, M. M. et al. (2010). Cell type–specific gene expression differences in complex tissues. *Nat. Methods* **7** 287–289.

WANG, N., HOFFMAN, E. P., CHEN, L., CHEN, L., ZHANG, Z., LIU, C., YU, G., HERRINGTON, D. M., CLARKE, R. et al. (2016). Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues. *Sci. Rep.* **6** 18909.

WEST, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics*, 7 (*Tenerife*, 2002) 733–742. Oxford Univ. Press, New York. MR2003537

WILKERSON, M. D., CABANSKI, C. R., SUN, W., HOADLEY, K. A., WALTER, V., MOSE, L. E., TROESTER, M. A., HAMMERMAN, P. S., PARKER, J. S. et al. (2014). Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic Acids Res.* **42** e107.

XIE, F., ZHOU, M. and XU, Y. (2018). BayCount: A Bayesian decomposition method for inferring tumor heterogeneity using RNA-Seq counts. *Ann. Appl. Stat.* **12** 1605–1627. MR3852690 https://doi.org/10.1214/17-AOAS1123

ZARE, H., WANG, J., HU, A., WEBER, K., SMITH, J., NICKERSON, D., SONG, C., WITTEN, D., BLAU, C. A. et al. (2014). Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput. Biol.* **10** e1003703.

ZENG, L., WARREN, J. L. and ZHAO, H. (2019). Phylogeny-based tumor subclone identification using a Bayesian feature allocation model. *Ann. Appl. Stat.* **13** 1212–1241. MR3963569 https://doi.org/10.1214/18-AOAS1223

ZHOU, T., MÜLLER, P., SENGUPTA, S. and JI, Y. (2019a). PairClone: A Bayesian subclone caller based on mutation pairs. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 705–725. MR3937470 https://doi.org/10.1111/rssc.12328

ZHOU, T., SENGUPTA, S., MÜLLER, P. and JI, Y. (2019b). TreeClone: Reconstruction of tumor subclone phylogeny based on mutation pairs using next generation sequencing data. *Ann. Appl. Stat.* **13** 874–899. MR3963556 https://doi.org/10.1214/18-AOAS1224

ZHOU, T., SENGUPTA, S., MÜLLER, P. and JI, Y. (2020). Supplement to "RNDClone: Tumor subclone reconstruction based on integrating DNA and RNA sequence data." https://doi.org/10.1214/20-AOAS1368SUPPA, https://doi.org/10.1214/20-AOAS1368SUPPB.