The Annals of Applied Statistics
2016, Vol. 10, No. 4, 2377–2404
DOI: 10.1214/16-AOAS986

A PHYLOGENETIC LATENT FEATURE MODEL FOR CLONAL DECONVOLUTION[1]

BY FRANCESCO MARASS*, FLORENT MOULIERE*, KE YUAN†, NITZAN ROSENFELD* AND FLORIAN MARKOWETZ*

University of Cambridge and University of Glasgow†*

Tumours develop in an evolutionary process, in which the accumulation of mutations produces subpopulations of cells with distinct mutational profiles, called clones. This process leads to the genetic heterogeneity widely observed in tumour sequencing data, but identifying the genotypes and frequencies of the different clones is still a major challenge. Here, we present Cloe, a phylogenetic latent feature model to deconvolute tumour sequencing data into a set of related genotypes. Our approach extends latent feature models by placing the features as nodes in a latent tree. The resulting model can capture both the acquisition and the loss of mutations, as well as episodes of convergent evolution. We establish the validity of Cloe on synthetic data and assess its performance on controlled biological data, comparing our reconstructions to those of several published state-of-the-art methods. We show that our method provides highly accurate reconstructions and identifies the number of clones, their genotypes and frequencies even at a modest sequencing depth. As a proof of concept, we apply our model to clinical data from three cases with chronic lymphocytic leukaemia and one case with acute myeloid leukaemia.

1. Introduction. Cancers evolve through waves of mutation and clonal expansion [Nowell (1976)]. Darwinian selection operates on the increased variation within the tumour, favouring clones with increased fitness, according to microenvironmental and therapeutic pressures [Fearon and Vogelstein (1990); Stratton, Campbell and Futreal (2009); Aparicio and Caldas (2013); Beerenwinkel et al. (2015)]. As a consequence of this evolutionary process, tumours are generally genetically heterogeneous [Gerlinger et al. (2012); Nik-Zainal et al. (2012)] and consist of related populations of cancer cells (*clones*) with distinct genotypes, which encode the evolutionary history of each cell population [Nik-Zainal et al. (2012)]. This genetic heterogeneity is important clinically because it can confound the molecular profiling of biopsies, and increased variation may equip tumours with more avenues to escape treatment, leading to worse prognosis [Schwarz et al. (2015)].

Received April 2016; revised August 2016.

[1]Supported in part by CRUK core grant C14303/A17197, in particular A20240 (Rosenfeld lab core grant) and A19274 (Markowetz lab core grant).

Key words and phrases. Clonal deconvolution, tumour heterogeneity, latent feature model, phylogeny, admixture.

The clonal deconvolution problem. Identifying clones and their proportions is a difficult task [Beerenwinkel et al. (2015)], aggravated by the fact that cancer genomics data generally come from bulk sequencing experiments, which profile a mixture of cells from different clones. Clones are related to each other and can be thought of as nodes in a phylogenetic tree that describes tumour development. The root of the tree corresponds to a normal, nonmutated cell; every other node is a cancer clone with a distinct complement of mutations (its *genotype*). Each clone inherits the mutations of its parent and adds more to them. This encodes a subset relationship between parent and child nodes.

However, none of this is directly observable. Instead, the data only consist of a set of mutations and their proportions (called *allele fractions*) in a collection of tumour samples (Figure 1). The *clonal deconvolution problem* thus asks to identify the clonal genotypes, phylogeny and clonal fractions that best explain the observed data [El-Kebir et al. (2015)].

Additional challenges. The clonal deconvolution problem is further complicated by factors such as the selection of alleles during tumour evolution and the specifics of the data obtained from sequencing experiments. In particular, convergent evolution and mutational loss contradict the common assumption that mutations arise only once in the phylogeny (the *infinite sites assumption*) and never disappear. Tumours are subjected to internal selective pressures in their microenvironment and external pressures from therapeutic interventions. In such cases, multiple tumour clones may acquire the same mutation in convergent evolution, especially if it is a hotspot mutation or it confers resistance to the treatment. At the same time, mutations can be removed by several mechanisms, including loss of heterozygosity, the deletion of the chromosome fragment carrying the mutation. Another challenge is that, for cost-effective sequencing options like targeted amplicon sequencing, which we will use for the validation and the case studies, the depth of sequencing is not informative of the chromosomal copy number of the tumour. This contradicts assumptions often made by previous methods.

Previous approaches. Various methods have been proposed in the literature [Beerenwinkel et al. (2015)] to improve on manual analyses [Gerlinger et al. (2012); Nik-Zainal et al. (2012)]. To put our approach in context, it is useful to distinguish *direct* reconstructions that directly infer clonal genotypes [e.g., CloneHD [Fischer et al. (2014)], Clomial [Zare et al. (2014)] and BayClone [Sengupta et al. (2015)]] from *indirect* reconstructions that obtain clusters of mutations rather than full genotypes and require additional phylogenetic analysis to obtain genotypes [e.g., PyClone [Roth et al. (2014)], SciClone [Miller et al. (2014)], PhyloWGS [Deshwar et al. (2015)] and BitPhylogeny [Yuan et al. (2015)]].

Direct reconstructions generally aim to infer two quantities, a matrix of mutation assignments and a matrix of clonal fractions, which come together in an admixture in the sampling model. The mutation assignments matrix associates each mutation with zero or more classes, which can be intuitively interpreted as clonal

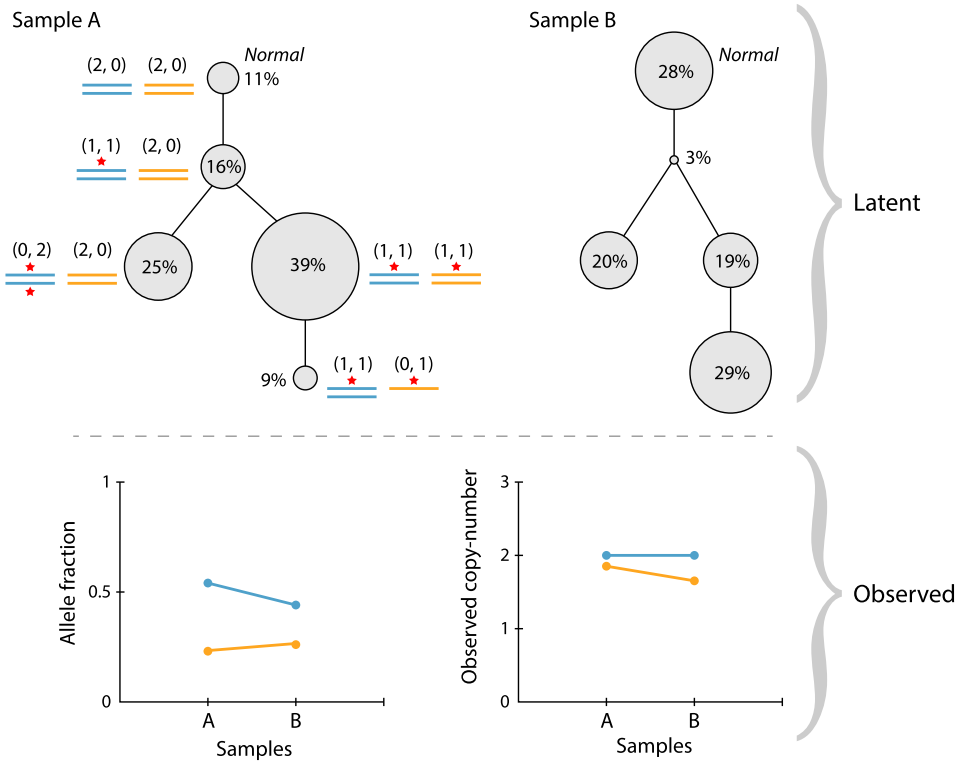

FIG. 1. *Overview of the general clonal deconvolution problem. Sample A shows how from a normal root node four clones evolved according to the displayed phylogeny. In this example, genotypes consist of two loci (blue and orange), which may be mutated (red star), gained or lost. The normal genotype consists of two nonmutated alleles for each locus. Clonal fractions are represented by the diameter of the node and reported as a percentage. A second sample B from the same patient would also consist of the same tree and genotypes; clonal fractions, however, may change. These latent parameters give rise to the observed mutation and copy-number data, shown at the bottom. The allele fraction of a mutation is the proportion of that allele in the sample. The observed copy number is the total copy number of each clone weighted by the clonal fractions. The increase of the orange mutation's allele fraction and the decrease of its observed copy number are due to the growth of the clone with a single and mutated copy.*

genotypes. For models that lack a phylogeny, inference may yield biologically implausible genotypes, as shown later in the benchmarking studies (Section 3.3).

On the other hand, indirect methods cluster mutations based on their allele fractions across multiple samples. Joint phylogenetic modelling allows these clusters to become nodes of a tree, displaying at which node each mutation first appeared. Hence, the assignment of mutation clusters to nodes of a tree is generally inflexible to episodes of convergent evolution or mutational loss.

Latent feature models. Here we introduce Cloe, a phylogenetic latent feature model for clonal deconvolution that belongs to the category of direct reconstruc-

tion methods. Latent feature models discover independent features with which to describe a set of observed objects. The set of features possessed by an object determines the parameters of its distribution [Ghahramani and Griffiths (2005)]. In our context, observed objects are mutations, and latent features are clonal genotypes representing clones.

Latent feature models have been previously applied to clonal deconvolutions, but maintained the assumption that features are conditionally independent [Zare et al. (2014); Sengupta et al. (2015)]. In parallel, extensions to these models have been developed to relate features hierarchically, but placed features as the leaves of the tree [Heaukulani, Knowles and Ghahramani (2014)]. Moreover, these tree structures only correlated the feature assignments, making such a model unsuitable for clonal deconvolutions.

The model we propose lifts the independence assumption and relates features with a latent hierarchy, where each feature is independent of the others given its parent. In our framework, features correspond to the nodes of the tree, which encodes a noisy subset relationship in the mutation assignments. Our model differs from the phylogenetic Indian Buffet Process, as the latter relates observed objects with a latent phylogeny rather than the features [Miller, Griffiths and Jordan (2012)]. Our approach is more general than previously published methods because it relies on fewer assumptions on clones and the evolutionary model: we can readily model multiple independent primary tumours, account for loss of mutations and penalise, though still allow, convergent evolution.

We validate Cloe on simulated data, on a controlled biological dataset, and apply it to two published clinical datasets: longitudinal samples from three chronic lymphocytic leukaemia patients [Schuh et al. (2012)] and from an acute myeloid leukaemia case [Griffith et al. (2015)]. Cloe is available as an R package at https://bitbucket.org/fm361/cloe.

2. The Cloe model. Our model follows the overview of Figure 1. A latent phylogenetic tree influences the clonal genotypes; these, together with clonal fractions and additional nuisance parameters, describe the distribution of the data.

We observe data for J mutations in T samples, and the data are collected in two $J \times T$ matrices: $\mathbf{X}$ for mutant read counts and $\mathbf{D}$ for read depths, the number of times a particular locus of the genome is covered by sequencing reads.

The phylogenetic tree is defined by a vector $\mathcal{T}$ with $K > 1$ elements, one for each clone. We consider the normal contamination as a fixed clone. Our analysis is restricted to mutations in copy-number neutral regions: each clone, including the normal, contributes exactly two copies of each allele, of which at most one can be mutated. Clonal genotypes are defined in a binary $J \times K$ matrix $\mathbf{Z}$, where each column $z_{\cdot k}$ represents the genotype of clone k. The proportions of each clone in each sample are summarised in a $K \times T$ matrix $\mathbf{F}$ formed by T stochastic vectors.

Our goal is to infer the phylogeny $\mathcal{T}$, clonal genotypes $\mathbf{Z}$ and clonal fractions $\mathbf{F}$ from the posterior distribution $P(\mathcal{T}, \mathbf{Z}, \mathbf{F} \mid \mathbf{X}, \mathbf{D})$. To do this, in the following

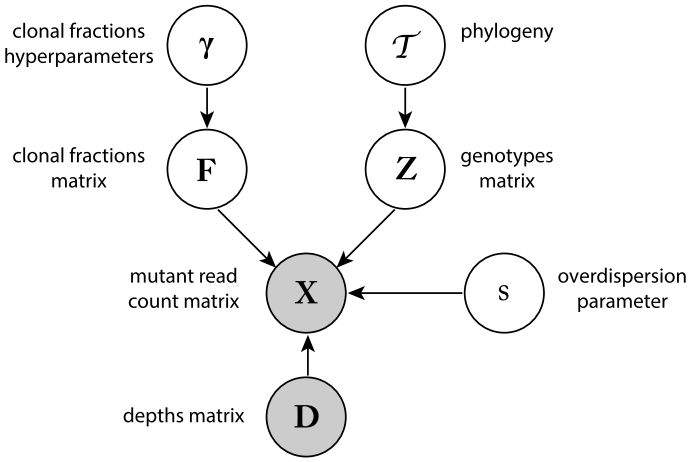

FIG. 2. *Simplified graphical model corresponding to Cloe, omitting overlapping plates and convergent evolution relations.*

sections we develop a probabilistic model that links observed and unobserved variables, and an inference algorithm to explore the posterior distribution.

2.1. *Model definition.* For guidance, a simplified version of our model is outlined in Figure 2, while at the end of this section Figure 3 presents the complete model.

Phylogeny. For $K > 1$ populations, we model the phylogenetic tree as a vector $\mathcal{T}$ of length K, where $\mathcal{T}_k = l$ means that the parent of k is l. The normal clone is fixed as the first entry, the root of the tree. To ensure that the graph encoded by $\mathcal{T}$ is a tree, we let each entry only take values on the previous entries. This definition is flexible, as the tree can assume any shape, even allowing phylogenies with multiple primary tumours. $\mathcal{T}$ is defined by

$$\begin{aligned} \mathcal{T}_1 &= 0, \\ \mathcal{T}_2 &= 1, \\ \mathcal{T}_k &\sim \mathcal{U}(\delta, k-1) \qquad \text{for } k \in \{3, \dots, K\}, \end{aligned} \tag{2.1}$$

where $\mathcal{U}(\delta, a)$ is a one-deflated discrete uniform distribution with values in $\{1, \dots, a\}$. The probability of drawing a 1 is δ, and the probability of drawing an integer between 2 and a is uniform:

$$\mathcal{U}(x; \delta, a) = \begin{cases} \delta & \text{if } x = 1, \\ \dfrac{1-\delta}{a-1} & \text{if } x \in \{2, 3, \dots, a\}. \end{cases} \tag{2.2}$$

We penalise multiple independent primary tumours (multiple children of the normal clone) by setting $\delta = (2k)^{-1}$.

Genotypes. Genotypes are defined in a binary $J \times K$ matrix $\mathbf{Z} = (z_{jk})$, where 1 denotes a mutation and 0 the unmutated (*wild-type*) state, for each mutation j in each clone k. We fix the genotype of the normal clone to a zero vector of length J, implying that all mutations are somatic. More generally, the normal genotype could be modified to accommodate known germline variants. The genotype of a clone k for a mutation j is then defined as

$$z_{jk} \mid z_{j\mathcal{T}_k}, \mu, \rho \sim \text{Bernoulli}(p_{jk}), \tag{2.3}$$

where μ is the probability of mutating if the parent does not have a mutation, and ρ is the probability of reverting to wildtype if the parent is mutated:

$$p_{jk} = \begin{cases} \mu & \text{if } z_{j\mathcal{T}_k} = 0, \\ 1 - \rho & \text{if } z_{j\mathcal{T}_k} = 1. \end{cases} \tag{2.4}$$

Clonal fractions. Because the samples may be spatially or temporally separated, and collected at irregular intervals, we assume that clonal fractions are independent between samples. We represent clonal fractions with a $K \times T$ matrix $\mathbf{F} = (\mathbf{f}_{\cdot t})_{t=1,\dots,T}$ composed of stochastic column vectors $\mathbf{f}_{\cdot t}$ describing the proportions of each clone in a sample t. Clonal fractions for a sample t are modelled with a symmetric Dirichlet distribution with hyperparameter γ_t:

$$\mathbf{f}_{\cdot t} \mid \gamma_t \sim \text{Dirichlet}(\gamma_t). \tag{2.5}$$

Likelihood. Genotypes and clonal fractions come together in an admixture, their dot product representing the expected allele fractions for each mutation in each sample. We model mutant reads as successful trials from a beta-binomial distribution with overdispersion parameter s. The probability of success is a function of the expected allele fraction $p_{jt} = \frac{1}{2}(\mathbf{z}_{j\cdot} \cdot \mathbf{f}_{\cdot t})$. To capture sequencing noise at extreme values of p_{jt}, we replace it with a function $e(p_{jt})$ that depends on the sequencing error rate ε (e.g., 0.1%) such that

$$e(p_{jt}) = \begin{cases} \varepsilon & \text{if } p_{jt} = 0, \\ 1 - \varepsilon & \text{if } p_{jt} = 1, \\ p_{jt} & \text{otherwise.} \end{cases} \tag{2.6}$$

The likelihood is then specified by

$$x_{jt} \mid d_{jt}, \mathbf{z}_{j\cdot}, \mathbf{f}_{\cdot t}, s \sim \text{Beta-binomial}\big(d_{jt}, e(p_{jt}), s\big). \tag{2.7}$$

Nuisance parameters. We let the beta-binomial overdispersion parameter s and the Dirichlet hyperparameters $\boldsymbol{\gamma}$ have Gamma priors, whereas the mutation and reversion probabilities μ and ρ are fixed.

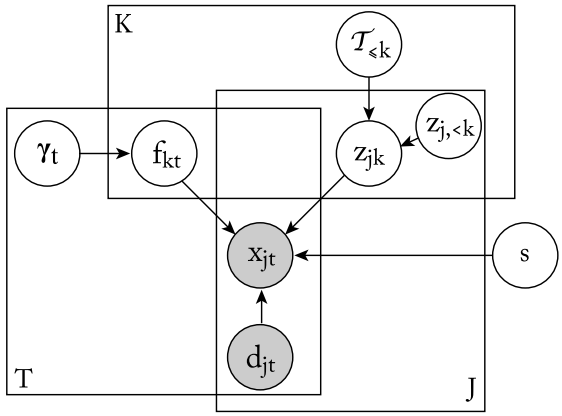

FIG. 3. *The full graphical model of Cloe.*

2.2. *Penalising convergent evolution.* One of the risks of assuming the independence of features in this biological application is that the inferred genotypes may largely display convergent evolution. We can penalise such occurrences by altering the definition of genotypes [cf. equations (2.3) and (2.4)].

Under the infinite sites assumption (ISA) every mutation occurs only once so that if multiple clones possess a mutation j, then the mutation must have appeared with their most recent common ancestor. In contrast, if the most recent common ancestor is not mutated, then the mutation must have appeared multiple times (convergent evolution). We thus say that a mutation assignment $z_{jk} = 1$ conflicts with ISA if the most recent common ancestor of $\{k' : z_{jk'} = 1, k' \leq k\}$ does not harbour mutation j.

We include ISA checks into our model by using an indicator function $I(j, k, a)$ that returns 1 if the most recent common ancestor of all clones $k' \leq k$ that harbour mutation j also possesses the mutation when $z_{jk} = a$; that is, $I(j, k, a) = 1$ if ISA is satisfied by setting $z_{jk} = a$, and 0 otherwise.

Thus, we redefine the distribution of genotypes making them conditional on all previous genotypes and weighting assignments by a user-defined parameter ν if they comply with ISA, or by $1 - \nu$ if they do not:

$$P(z_{jk} = 1 \mid \mathcal{T}, \mathbf{Z}_{j,<k}, z_{j\mathcal{T}_k} = 0, \mu, \rho, \nu) \propto (\mu\nu)^{I(j,k,1)}\big(\mu(1-\nu)\big)^{1-I(j,k,1)}. \tag{2.8}$$

In practice, only transitions that gain a mutation can clash with ISA, and the factor of ν immediately cancels out if the parental genotype is 1 at a given j. An ISA-check at j is thus only warranted if the parent is not mutated at j.

The graphical model corresponding to what has been described so far is shown in Figure 3.

2.3. *Inference.* We are interested in the posterior distribution of the latent variables given the observed variables $P(\mathcal{T}, \mathbf{Z}, \mathbf{F} \mid \mathbf{X}, \mathbf{D})$. We approximate the

Algorithm 1 MCMCMC sampling algorithm for Cloe

for $i = 1, \ldots,$ #iterations **do**
  for $m = 1, \ldots,$ #chains **do**
    for $j = 1, \ldots, J$ **do** ▷ $\mathbf{Z}$
      Propose new $\mathbf{z}_{j\cdot}^{*(m)}$
      Accept with probability 2.12
    end for
    for $k = 3, \ldots, K$ **do** ▷ $\mathcal{T}$
      Compute $P(\mathcal{T}_k^{*(m)} = l)$ for $l \in \{1, \ldots, k-1\}$ (eq. 2.9)
      Sample new $\mathcal{T}_k^{*(m)}$ from $P(\mathcal{T}_k^{*(m)})$
    end for
    Randomly swap two siblings
    With probability 1% propose a swap between a node and its parent
    Accept with probability 2.11
    for $t = 1, \ldots, T$ **do** ▷ $\mathbf{F}$
      Propose new $\mathbf{f}_{\cdot t}^{*(m)}$ from eq. 2.13
      Accept with probability 2.14
    end for ▷ Nuisance parameters
    Propose new $s^{*(m)} \sim \mathcal{N}(s^{(m)}, \sigma_s)$ and accept with probability 2.15
    for $t = 1, \ldots, T$ **do**
      Propose new $\gamma_t^{*(m)} \sim \mathcal{N}(\gamma_t^{(m)}, \sigma_\gamma)$ and accept with probability 2.16
    end for
  end for
  if i is a multiple of 100 **then** ▷ Chain swap
    Propose a chain $j \in \{1, \ldots, m-1\}$
    Accept the state swap between chains j and $j+1$ with probability 2.17
  end if
end for

posterior by Metropolis-coupled Markov chain Monte Carlo [MCMCMC; Geyer (1991)]. Within each MCMCMC chain, we use a generalised Gibbs sampler to draw samples of the tree vector and other parameters from their full conditionals.

Because the posterior landscape appears composed of high peaks separated by deep valleys (Supplementary Figure 1), we run five chains in parallel, with tempered posteriors. The sampling strategy described hereafter is applied to each chain and summarised in Algorithm 1.

Phylogeny. For each $\mathcal{T}_{k>2}$, we compute the conditional posterior of the parent assignment $\mathcal{T}_k = l$, for each $l < k$:

$$P(\mathcal{T}_k = l \mid \ldots) \propto P(\mathbf{Z} \mid \mathcal{T}_{-k}, \mathcal{T}_k = l) P(\mathcal{T}_k = l). \tag{2.9}$$

The likelihood term amounts to tallying the genotype transitions from parent l to child k, and reassessing how many transitions comply or clash with ISA for all

clones. The prior, according to equation (2.1), is equal to

$$P(\mathcal{T}_k = l) = \delta^{\boldsymbol{I}(l=1)}\left(\frac{1-\delta}{k-2}\right)^{1-\boldsymbol{I}(l=1)} = \left(\frac{1}{2k}\right)^{\boldsymbol{I}(l=1)}\left(\frac{2k-1}{2k(k-2)}\right)^{1-\boldsymbol{I}(l=1)}. \tag{2.10}$$

To facilitate the exploration of the space of tree and genotypes configurations, we uniformly propose a pair of siblings and swap their position in the tree and in the genotypes and clonal fractions matrices. Prior to this swap, the siblings had access to one linear topology. This move allows the other linear topology to be explored while leaving probabilities unaltered.

In addition, the swap between a node k and its parent l is proposed. A node k is chosen uniformly from $\{3, \ldots, K\}$. A tree $\mathcal{T}^*$ is created where k is the parent of l, while any children of k remain children of k; the same applies to l. As with the sibling swap, this move requires rearranging the clone order in the genotypes and clonal fractions matrices. The parent swap affects genotype transitions from $\mathcal{T}_l$, the original parent of l, to k, and from k to l. The proposal is accepted with probability

$$\min\left(1, \frac{P(\mathbf{Z}_{\{\mathcal{T}_l,k,l\}} \mid \mathcal{T}^*, \mu, \rho, \nu)P(\mathcal{T}^*_{\{\mathcal{T}_l,k,l\}})}{P(\mathbf{Z}_{\{\mathcal{T}_l,k,l\}} \mid \mathcal{T}, \mu, \rho, \nu)P(\mathcal{T}_{\{\mathcal{T}_l,k,l\}})}\right). \tag{2.11}$$

This move rescues the sampler from local maxima where the parental relationships are learnt in the wrong order. Because this situation is rare, and the move is computationally expensive, we perform this move with probability 0.01.

Genotypes. Because mutations are independent, we update $\mathbf{Z}$ by row, proposing a new row $\mathbf{z}^*_{j\cdot}$ by flipping each bit of $\mathbf{z}_{j\cdot}$ with probability θ. The proposal is symmetric and the move is accepted with probability

$$\min\left(1, \frac{P(\mathbf{x}_{j\cdot} \mid \mathbf{d}_{j\cdot}, \mathbf{z}^*_{j\cdot}, \mathbf{F}, s)P(\mathbf{z}^*_{j\cdot} \mid \mathcal{T}, \mu, \rho, \nu)}{P(\mathbf{x}_{j\cdot} \mid \mathbf{d}_{j\cdot}, \mathbf{z}_{j\cdot}, \mathbf{F}, s)P(\mathbf{z}_{j\cdot} \mid \mathcal{T}, \mu, \rho, \nu)}\right), \tag{2.12}$$

where the likelihood is only computed for mutation j, and the prior refers to the sequence of transitions from the root genotype at j to the leaves, with appropriate penalties for convergent evolution.

Clonal fractions. Because of the independence of the samples, the matrix $\mathbf{F}$ is updated by column. A new vector $\mathbf{f}^*_{\cdot t}$ is proposed from a Dirichlet distribution centred at the current value $\mathbf{f}_{\cdot t}$:

$$Q(\mathbf{f}^*_{\cdot t} \mid \mathbf{f}_{\cdot t}) = \text{Dirichlet}(\psi \mathbf{f}_{\cdot t} + \epsilon), \tag{2.13}$$

where ψ is a precision factor and ϵ a small bias to avoid sinks at 0. The proposal is accepted with probability

$$\min\left(1, \frac{P(\mathbf{x}_{\cdot t} \mid \mathbf{d}_{\cdot t}, \mathbf{Z}, \mathbf{f}^*_{\cdot t}, s)P(\mathbf{f}^*_{\cdot t} \mid \gamma_t)Q(\mathbf{f}_{\cdot t} \mid \mathbf{f}^*_{\cdot t})}{P(\mathbf{x}_{\cdot t} \mid \mathbf{d}_{\cdot t}, \mathbf{Z}, \mathbf{f}_{\cdot t}, s)P(\mathbf{f}_{\cdot t} \mid \gamma_t)Q(\mathbf{f}^*_{\cdot t} \mid \mathbf{f}_{\cdot t})}\right). \tag{2.14}$$

Nuisance parameters. The remaining parameters are updated with Metropolis moves using Gaussian proposals. The Metropolis–Hastings acceptance ratios are

$$\min\left(1, \frac{P(\mathbf{X} \mid \mathbf{D}, \mathbf{Z}, \mathbf{F}, s^*)P(s^*)}{P(\mathbf{X} \mid \mathbf{D}, \mathbf{Z}, \mathbf{F}, s)P(s)}\right) \qquad \text{for } s, \tag{2.15}$$

and

$$\min\left(1, \frac{P(\mathbf{f}_{\cdot t} \mid \gamma_t^*)P(\gamma_t^*)}{P(\mathbf{f}_{\cdot t} \mid \gamma_t)P(\gamma_t)}\right) \qquad \text{for } \gamma_t. \tag{2.16}$$

Temperatures and chain swaps. Regularly at user-defined intervals, a swap between two adjacent chains is proposed as a Metropolis–Hastings move. Let M denote the number of parallel chains, $P^{(m)}$ denote the tempered posterior of chain m, and ω_m denote the state of chain m. A chain m is selected among the first $M-1$ chains. The swap between chains m and $m+1$ is then accepted with probability

$$\min\left(1, \frac{P^{(m)}(\omega_{m+1})P^{(m+1)}(\omega_m)}{P^{(m)}(\omega_m)P^{(m+1)}(\omega_{m+1})}\right). \tag{2.17}$$

The temperature τ_m for each chain m is chosen according to the following scheme [Ronquist, Huelsenbeck and Teslenko (2005)]:

$$\tau_m = \big(1 + \Delta T(m-1)\big)^{-1}, \tag{2.18}$$

where $\Delta T > 0$ regulates the temperature differences between chains.

Parameter estimates. MCMCMC parameter estimates are derived solely from the first, untempered chain. After discarding a certain proportion of the initial samples as burn-in, and thinning the chain by a factor of i, thus considering every ith sample, we obtain a maximum *a posteriori* (MAP) estimate of the parameters by selecting the chain state of the sample with the highest posterior value.

Model selection. When Cloe is run on the same dataset with various values of K, a model selection criterion is needed to rank the solutions. To do so, we use the log-posterior probability of the MAP estimate, as it accounts for the fit to the data as well as the model complexity. Nevertheless, because this is a heuristic, manual review of the results is recommended. When multiple models attain similar log-posterior probabilities, we prefer solutions with higher log-likelihood values, denoting a better fit to the data.

3. Validation and benchmarks. We extensively validated and benchmarked Cloe by using simulated data and a controlled experimental setup based on mixtures of cell line DNA.

3.1. *Simulated data*. We first tested our model on 9 simulated datasets, one for each combination of number of clones (3, 4 or 5) and depth of sequencing (means: 50×, 200×, 1000×). The genotypes were created according to a random tree and using parameters $\mu = 0.5$, $\rho = 0.05$, $\nu = 0.9$; clonal fractions were *iid* draws from

TABLE 1
Running parameters for the simulated and validation datasets. For ε, the sequencing error parameter, the first value refers to the simulations, and the second to the validation dataset

Parameter	Value
MCMCMC	
Iterations	40,000
Chains	5
ΔT	0.4
Swap interval	50
Burn-in	50%
Thinning factor	4
Z	
μ	0.3
ρ	0.1
ν	0.75
θ (proposal)	0.20
ε (likelihood)	0.005, 0.002
F	
ψ (proposal)	200
ϵ (proposal)	4
Nuisance parameters	
γ (prior, shape)	2
γ (prior, rate)	1
σ_γ (proposal)	0.2
s (prior, shape)	11
s (prior, rate)	0.10
σ_s (proposal)	16

a symmetric Dirichlet distribution with parameter $\gamma = 2$. All datasets contained 100 mutations and 5 samples, with depths obtained from a Poisson distribution and mutant read counts obtained from a binomial distribution. Because of the fixed size of our model, we ran Cloe for 3, 4 and 5 clones on each of the 9 datasets (running parameters are reported in Table 1).

We measured Cloe's performance by its ability to identify the correct number of clones and the right parameters. In addition, we computed the frequentist coverage probability and assessed mixing performance from three consecutive runs of the algorithm. To compute the reconstruction error, we calculated two metrics, the

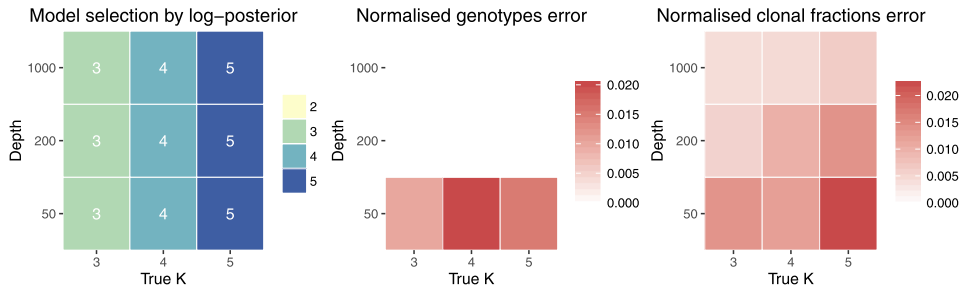

FIG. 4. *Results on the simulated datasets. Left: inferred model size for every combination of K and depths. Centre and right: reconstruction errors. All datasets consisted of five samples and* 100 *mutations.*

normalised genotypes error Z_{err} and the normalised clonal fractions error F_{err}, both defined as the sum of the absolute differences between inferred ($\mathbf{Z}^*$) and the true ($\mathbf{Z}$) matrices, normalised by the real matrix dimensions (ignoring the fixed genotype in $\mathbf{Z}$). To control for equivalent solutions with permuted clones, we find the permutation σ of the columns of $\mathbf{Z}^*$ that minimises Z_{err}:

$$Z_{\text{err}} = \frac{1}{J(K-1)} \min_{\sigma} \left(\sum_{j,k} |z^*_{j\sigma(k)} - z_{jk}| \right). \tag{3.1}$$

The same permutation is then used to rearrange the rows of $\mathbf{F}^*$, which is normalised by JK. If the true and inferred matrices have different sizes, we pad the smaller one with normal clones with zero clonal fractions. In this instance K refers to the correct number of clones.

Model selection. Using our automated model selection, we were able to recover the correct size K for every dataset (Figure 4, left) in two of the three runs. In one run, a higher log-posterior probability was given to the solution with 4 clones on the dataset with 50× and 5 clones (-14464.05 compared to the 5-clone solution's -14464.58; Supplementary Figure 3).

Reconstruction fidelity. To assess the reconstructions, we considered only the MAP solution for each dataset. The reconstruction error was low, with $Z_{\text{err}} \leq 0.027$ and $F_{\text{err}} \leq 0.033$. The largest errors were obtained at the lowest depth (Figure 4, centre and right), suggesting that on these random datasets Cloe can not only discover the correct number of clones, but also infer correct genotypes and clonal fractions with $>96.7\%$ accuracy (Supplementary Figures 4, 5 and 6).

MCMC performance. As measured by the reconstruction error metrics, Cloe can provide accurate inference of genotypes and clonal fractions on both simulated and biological data (Section 3.2). However, the shape of the posterior distribution may prevent a complete exploration of all peaks in all chains (Supplementary Figure 1). Convergence to high probability regions is quickly reached, yet it is possible that the sampler may climb the correct peak and move away from it with difficulty.

TABLE 2

Effective sample size per dataset, computed on 5000 *post-burn-in and post-thinning iterations for the first of the three replicate runs with the correct number of clones. The dataset is denoted by the average sequencing depth and the real number of clones. LP denotes the log-posterior, used as a proxy for the multidimensional parameters*

Dataset	γ_1	γ_2	γ_3	γ_4	γ_5	s	**LP**
50×, 3	99.87	140.05	283.02	75.51	53.55	791.69	999.84
200×, 3	102.72	184.97	262.84	87.50	74.97	446.42	620.97
1000×, 3	83.27	232.58	223.02	86.26	71.26	231.40	290.80
50×, 4	296.28	99.69	234.60	83.70	186.02	633.14	1369.27
200×, 4	461.14	87.64	330.48	77.61	219.94	555.83	894.29
1000×, 4	498.87	123.30	350.10	71.80	177.40	223.17	226.41
50×, 5	185.99	111.19	101.98	172.80	193.12	689.40	791.96
200×, 5	394.70	181.97	200.82	156.75	280.00	445.45	469.93
1000×, 5	507.30	202.85	202.03	160.05	340.71	269.50	293.98

To gain more insight into the performance of the sampler, we assessed convergence by the Gelman–Rubin statistic, coverage by computing frequentist coverage probabilities, and sampling by calculating the effective sample size (ESS). The Gelman–Rubin statistic was calculated from the log-posterior values of the untempered chains as a proxy for the multidimensional parameters. The potential scale reduction factor is within the accepted range, less than 1.1 (Supplementary Figure 2), for 24/27 cases. In three cases, at least one of the replicates did not converge to the same peak in the given number of iterations. Each run was started with a different random seed.

To calculate the coverage probability, we computed the 95% highest posterior density intervals from the MCMC traces, again using the log-posterior as a proxy. In each case, the true log-posterior probability was obtained by running the MCMC sampler with genotypes, clonal fractions and tree fixed to the correct values. Seven of the nine simulated datasets covered the true log-posterior in every replicate (Supplementary Table 1). In the remaining two cases, the true log-posterior lay close to but outside the upper bound of the interval. However, the accuracy of the reconstruction, as shown above, remained high, suggesting a discrepancy in the nuisance parameters.

Finally, we computed the ESS for the first run on each dataset, focussing on the cases where the sought number of clones matched the actual number of clones of the dataset. Table 2 reports the ESS of the nuisance parameters and the log-posterior, computed from 5000 post-burn-in and post-thinning iterations. Modulating the standard deviation of the Gaussian proposals for the nuisance parameters could decrease their autocorrelation. However, the shape of the posterior space (Supplementary Figure 1) may prevent efficient large moves.

TABLE 3
Clonal fractions in the 14 *mixtures of the validation experiment*

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
N	0.26	0.34	0.22	0.13	0.37	0.38	0.72	0.26	0.13	0.45	0.65	0.38	0.53	0.00
C1	0.03	0.41	0.14	0.29	0.14	0.14	0.04	0.38	0.03	0.02	0.06	0.09	0.19	0.08
C2	0.18	0.00	0.19	0.25	0.06	0.04	0.11	0.05	0.32	0.18	0.03	0.28	0.00	0.30
C3	0.44	0.19	0.00	0.17	0.10	0.41	0.13	0.18	0.43	0.23	0.20	0.25	0.22	0.57
C4	0.10	0.05	0.44	0.15	0.33	0.05	0.00	0.14	0.10	0.12	0.07	0.00	0.06	0.04

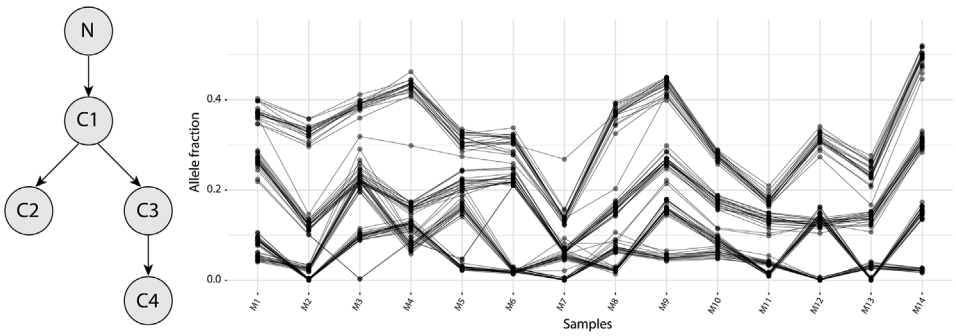

FIG. 5. *The observed data for the validation dataset. On the left is the artificial phylogeny, where N denotes the normal clone, and C*1 *to C*4 *are the genotypes derived from the cancer cell lines. On the right are the observed mutational dynamics* (*allele fractions over samples*) *at an average depth of* 17,260×.

3.2. *Controlled experimental data*. Because synthetic data may not capture the variability seen in real biological data, we tested our method on a bespoke experiment, described in detail in the Supplementary Material [Marass et al. (2016)]. Briefly, in order to mimic heterogeneous tumour samples, we genotyped DNA from four single-cell-diluted cancer cell lines and one normal cell line, and mixed it at known proportions (Table 3). Fourteen mixtures were created, with a median tumour content of 64%. In this experiment, the cancer cell lines represent tumour clones.

Because the cell lines were unrelated, we selected a subset of mutations (heterozygous single nucleotide variants and small indels) to embed the genotypes into an artificial phylogeny (Figure 5, left). The cell lines were genotyped by whole-exome sequencing and the allele fractions of the selected mutations were quantified by deep targeted sequencing [Forshew et al. (2012)]. The final dataset contained 82 mutations, quantified across 14 mixtures at a median depth of 17,260× (Figure 5, right).

The large number of samples and the high depth of sequencing that we obtained afforded a sensitivity analysis in which we varied the number of samples and the

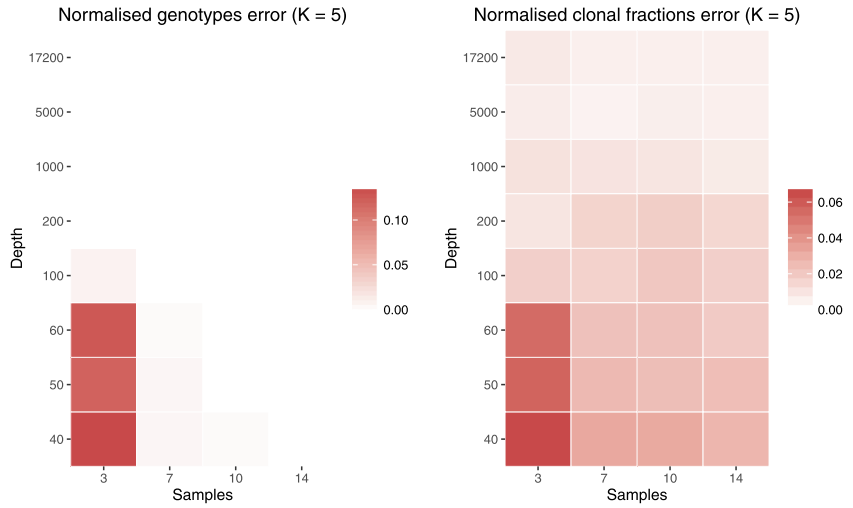

FIG. 6. *Reconstruction errors on validation data obtained running Cloe with $K = 5$ clones. The heatmaps show the genotypes error (left) and clonal fractions error (right) for various combinations of depth and samples.*

depth. Cloe was run on these datasets with the same parameters as for the synthetic data (Table 1), albeit with a smaller sequencing error parameter.

Model selection. We ran Cloe for $K \in \{3, 4, 5, 6\}$ and performed model selection based on the log-posterior values of the MAP estimates. In every case we were able to identify the correct number of clones (Supplementary Figure 7), suggesting that either a moderate depth of sequencing or multiple samples should suffice in obtaining good estimates of the number of clones.

Reconstruction fidelity. Overall, we obtained precise reconstructions for almost all depth-samples combinations. Considering for each combination only the first solution suggested by Cloe, on average 1% of mutation assignments were inaccurate (Z_{err} median 0, mean 0.013), and clonal fractions were inferred with an average error lower than 2% (F_{err} median 0.017, mean 0.019; Figure 6 and Supplementary Figure 8). As expected, we observed a pattern of decreasing errors as the data increase in the number of samples or in depth. Except for the low-depth cases described below, every solution reconstructed the correct tree topology (Supplementary Figure 9, left).

Specific low-depth cases. Poorer reconstructions were obtained at lower depths ($\leq 60\times$) for the datasets with three samples. In every case, the inferred genotypes showed a faulty separation between two expected genotypes (Supplementary Figure 10), which led to high error metrics: $Z_{\text{err}} \leq 0.134$ and $F_{\text{err}} \leq 0.065$. Inference of the phylogeny was also affected (Supplementary Figure 9, right). Despite the imprecise reconstruction, there is an overall good agreement with the observed data [Supplementary Figure 10(e)].

These results could be improved by tuning the running parameters of Cloe for these datasets. Because the height of the posterior peaks at these levels of depth is lower than at high depth, using less tempered chains may result in higher acceptance of chain swaps and, consequently, in a more complete exploration of the posterior space. Increasing the number of MCMCMC iterations could also prove to be beneficial.

It should be also noted that at low depths sampling noise may promote suboptimal parameter combinations to near-optimal. In this case, more mutations should be analysed in order to average sampling noise effects, though this may place a heavy burden on our implementation's runtime. Alternatively, one could model more data in terms of samples. If the clonal fractions are dynamic enough, meaning that most clones grow and shrink at some point in the samples, more opportunities are provided to separate clonal signals.

3.3. *Comparison to other approaches*. To further benchmark Cloe, we compared the results of four published methods compatible with targeted sequencing on our validation dataset: the latent feature models BayClone [Sengupta et al. (2015)] and Clomial [Zare et al. (2014)] as well as the nonparametric mixture models PhyloWGS [Deshwar et al. (2015)] and PyClone [Roth et al. (2014)]. Other methods, like CloneHD [Fischer et al. (2014)], are not applicable to targeted sequencing.

We ran two tests on the validation dataset described in the previous section, first using all samples and the entire depth, and then 3 samples and a depth of $100\times$. Our method's performance on the first dataset is a perfect reconstruction of the genotypes ($Z_{\text{err}} = 0$) and a near-perfect reconstruction of the clonal fractions ($F_{\text{err}} = 0.005$), with a correct identification of the number of clones. With less data, there are three misassignments ($Z_{\text{err}} = 0.009$) and the error of the clonal fractions is 0.017 (Figure 7); again, the number of clones is correctly inferred.

BayClone [Ji (2016)] was run with default parameters for 45,000 iterations, discarding the first 5000 and thinning the chain by a factor of 4. We tested the same model sizes as for our own method, namely, 3, 4, 5 and 6. Through the log-pseudo-marginal likelihood, BayClone was able to identify $K = 5$ as the best solution on the first dataset. However, its reconstruction of the genotypes was less precise (Figure 8), with $Z_{\text{err}} = 0.25$ and $F_{\text{err}} = 0.103$. Here Z_{err}, the normalised absolute difference between inferred and real genotypes matrices, ignores the ploidy of the mutation. The reconstruction was poorer on the second dataset because of the less precise data: six clones were inferred with $Z_{\text{err}} = 0.287$ and $F_{\text{err}} = 0.147$ (Supplementary Figure 11).

Clomial (version 1.6.0) implements an EM algorithm, and it was run with default parameters (1000 restarts and 100 maximum EM iterations) using model sizes of 3, 4, 5 and 6. On the first dataset, model selection with BIC (and AIC) indicated $K = 5$ as the best solution, with one misassignment ($Z_{\text{err}} = 0.003$) and an accurate

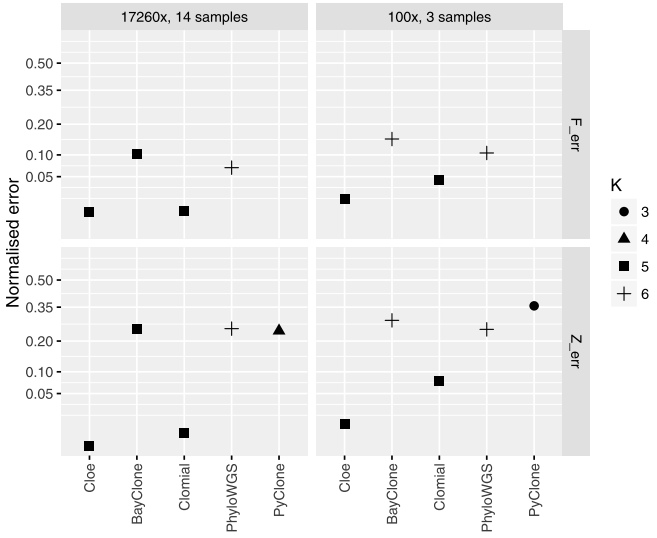

FIG. 7. *Comparison of Cloe against four published methods on our validation dataset. Z_{err} denotes the reconstruction error on genotypes, and F_{err} the error on clonal fractions. The column headers denote the data that were analysed in each instance. The legend refers to the number of inferred clones; the correct number is five.*

reconstruction of the clonal fractions ($F_{\text{err}} = 0.006$). With less data, model selection was unclear as AIC, BIC and the log-likelihoods were all discordant. Using the correct model size, Clomial obtained a $Z_{\text{err}} = 0.076$ and $F_{\text{err}} = 0.044$.

PyClone (version 0.12.9) was run for 30,000 iterations with a beta-binomial density and copy-number neutral states allowing a single mutant allele out of two (AB mode). PyClone was also provided with estimates of cellularity for each of the samples. We removed the first 3000 iterations as burn-in samples and thinned the chain by a factor of 4. The output of PyClone consists of a clustering of the observed mutations, where each cluster should correspond to one of the nonroot nodes of Figure 5 (left). Phylogenetic modelling can translate these clusters into genotypes. On the full dataset, PyClone produced three clusters, as shown in Figure 8. Because the cluster of stem mutations was merged with one of its two children, we were unable to interpret the results phylogenetically. Hence, we could not derive genotypes or clonal fractions. As such, the clusters were used as genotypes. The estimate of K is 4 with $Z_{\text{err}} = 0.241$. On less data, two clusters were produced, leading to an estimate of K of 3 with $Z_{\text{err}} = 0.357$.

PhyloWGS (commit `290645c`) was run with 1000 burn-in samples, 10,000 MCMC iterations, 5000 MH iterations and without copy-number information, so as to elicit copy-number neutral behaviour; only the top solution was considered. Like PyClone, PhyloWGS clusters mutations. However, joint phylogenetic analysis means that genotypes can be obtained by letting each clone have the mutations of the corresponding cluster plus the mutations of its ancestors. On both datasets,

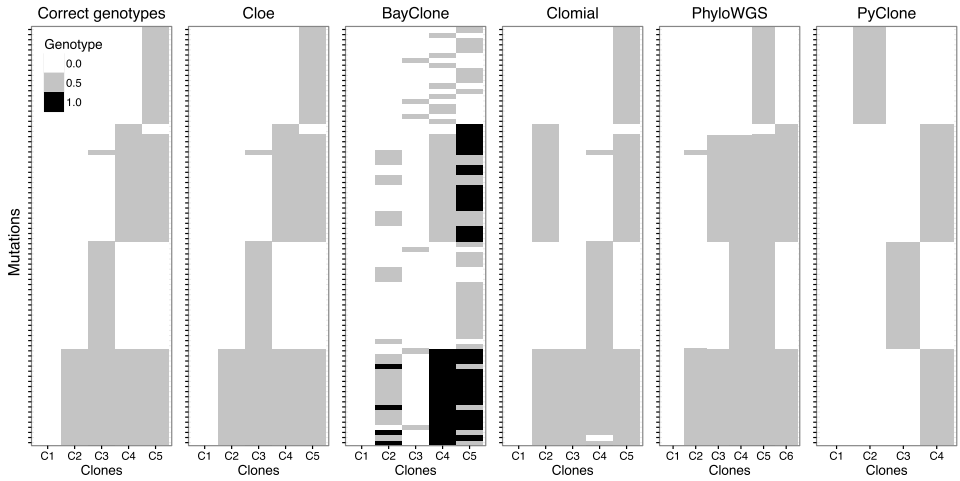

FIG. 8. *Comparison of the genotypes inferred by the five benchmarked methods using all the data in our validation dataset. PyClone's reconstruction is a clustering of the mutations. In this representation, solutions were padded with the normal clone* ($C1$) *for a more direct comparison with our method. The legend refers to the proportion of mutated alleles out of two.*

PhyloWGS inferred six clones, likely because it does not allow mutation losses. On the entire dataset the errors were $Z_{\text{err}} = 0.25$ and $F_{\text{err}} = 0.069$, while on the low-depth dataset the errors were $Z_{\text{err}} = 0.247$ and $F_{\text{err}} = 0.105$.

In summary, our benchmark shows that Cloe compares favourably against similar published methods (Figure 8). It is expected that the accuracy of the reconstruction would be affected by the quality of the data. Indeed, every model performed more poorly on less data, however, Cloe seemed to be affected to a lesser extent (Supplementary Figure 11).

4. Case studies. We show the applicability of Cloe to clinical data in two case studies.

4.1. *Chronic lymphocytic leukaemia.* This dataset consists of five time points for each of three chronic lymphocytic leukaemia patients [Schuh et al. (2012)]. The original study identified mutations by whole-genome sequencing (WGS; average depth across the mutation loci $39\times$) and quantified a subset of these with deep, targeted amplicon sequencing (TAR; average depth $101{,}600\times$).

The authors reported evolutionary trees and clonal fractions for each of the three cases, following k-means clustering of the mutations and a manual analysis. We used this information to run Cloe with known clonal fractions on all mutations, prioritising information from the higher depth datasets. We interpreted these results as ground truth mutation assignments for all three patients, and scored our reconstructions to these reference parameters.

We ran Cloe on the reported mutations with $K \in \{3, \ldots, 7\}$ for each case and each experiment (WGS, low-depth, and TAR, high-depth), comparing our results with the original study, with PhyloSub's results [Jiao et al. (2014)], and with CloneHD's reconstruction of case CLL003 [Fischer et al. (2014)]. We also included another dataset, which consisted of the WGS dataset with data from the higher depth TAR dataset for mutations in common.

Case CLL003 displays a radical clonal shift [Supplementary Figures 12(a) and (b)]: the main clone in the early time points is replaced by a distinct new clone that appears only at the second time point and expands to become the predominant clone. Using targeted sequencing data, Cloe obtained a very accurate reconstruction, identifying the correct number of clones, with a single misassignment and average errors on clonal fractions of 1% (Figure 9, Supplementary Figure 13). On less data, our method opted for a solution with 4 clones that ignored the founding clone, only present in the first of five samples at a clonal fraction of 3%. Choosing the top solution with 5 clones recovered the correct clonal structure. On WGS data there were five incorrect mutation assignments (Supplementary Figure 14), whereas with the combined dataset only one (Supplementary Figure 15). Barring the rare founding clone, the 4-clone reconstructions are correct with one (combined data) and two (WGS data) misassignments.

The remaining cases showed more stable dynamics [Supplementary Figures 12(c)–(f)]. For CLL006, Cloe assigned the nine mutations of the TAR dataset to six clones without errors; three errors were observed with 18 mutations in the WGS dataset (Figure 9). Analysis of the combination of the two data types yielded an additional clone, though similar log-posterior probabilities and a higher log-likelihood were obtained by a six-clone solution. Removing clone C5 from the seven-clone solution yields a correct reconstruction (Supplementary Figure 16).

Finally, for CLL077, Cloe's analysis resulted in a perfect reconstruction of the genotypes with targeted sequencing data. Two misassignments were obtained for the combined dataset, whereas four of the five clones were identified in the WGS data: the founding clone, with only four of the 20 mutations, was merged with one of its children. After the four-clone solutions, solutions with six clones had high log-posterior probabilities. Indeed, the first of these solutions is an accurate reconstruction with two misassignments and one clone repeated twice almost identically. In the middle, solutions with the expected number of clones, five, had six errors (Supplementary Figure 17).

Overall, Cloe produced accurate reconstructions of the latent parameters. Higher errors were observed when an incorrect number of clones was inferred. However, even in these cases, our phylogenetic model allowed us to obtain close approximations of the ground truth.

Assuming that our reconstruction of the ground truth is correct, Cloe's inference results in a better reconstruction than reported by CloneHD, both using high-depth and low-depth data: first, because of our phylogenetic modelling, we were able to identify the founding clone; second, we could confidently identify the rising

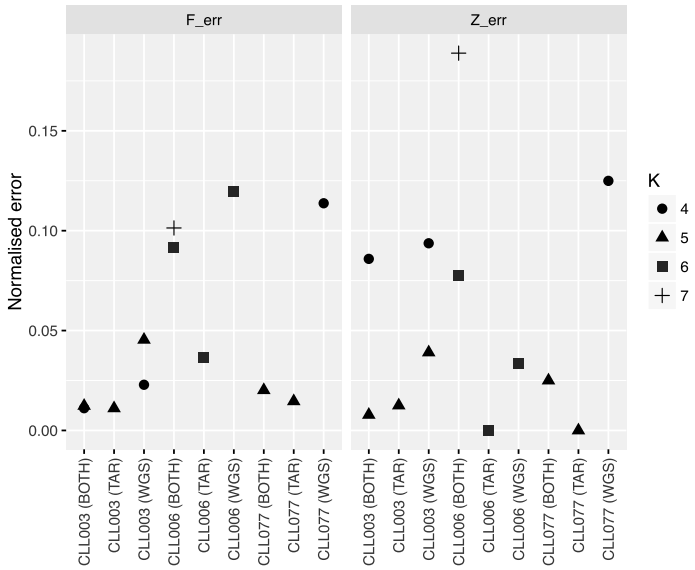

FIG. 9. *Performance metrics of Cloe on the CLL datasets. The correct number of clones for cases CLL*003 *and CLL*077 *is* 5, *whereas for CLL*006 *it is* 6. *TAR stands for targeted sequencing* (*average depth* 101,600×); *WGS stands for whole-genome sequencing* (*average depth* 39×); *BOTH is the WGS dataset with TAR data for shared mutations. The legend refers to the number of inferred clones. When the first solution inferred the wrong number of clones, the top solution for the correct number of clones is also shown.*

clone's parent [the ambiguous green clone in Fischer et al. (2014)]. With low-depth data, Cloe's automatic model selection preferred a model with four clones, but could also provide a more accurate five-clone solution.

Our results on targeted sequencing data largely agree with those obtained by PhyloSub, with two small exceptions. For CLL003, Cloe predicts that clone 4 [clone *c* of Figure 7 (right) in Jiao et al. (2014)] does not harbour the *IL11RA* mutation. This episode appears to be supported by the data (Supplementary Figure 18), as Cloe's reconstruction leads to a closer fit to the data (sum of absolute errors on the allele fractions of this mutation is 0.06 for Cloe, 0.13 for PhyloSub). Rather than a loss of mutation, this could be due to convergent evolution at the leaf nodes, leading to a sum of absolute errors of 0.07. For case CLL006, our reconstruction agrees with that of Schuh et al. (2012): five tumour clones are detected, and the *EGFR* mutation is predicted to stem from the founding clone. PhyloSub instead preferred to place the EGFR mutation in an additional clone after the founder, leading to a closer fit: the sum of absolute errors was 0.02 compared to Cloe's 0.07 for this mutation.

4.2. *Acute myeloid leukaemia.* AML31 refers to a patient with acute myeloid leukaemia, whose case was studied in great depth with several sequencing exper-

iments targeting bulk DNA at various scales, RNA and also single cells [Griffith et al. (2015)]. As each layer of data refined the authors' understanding of the evolution of this tumour, seven clusters and driver mutations were identified. Integration of all sequencing data revealed over 1300 mutations curated in a 'platinum list'. The tumour genomes appeared to be devoid of copy-number aberrations.

We considered a subset of platinum-list mutations for three datasets: ALLDNA (a pool of all the DNA sequencing data, median depth of 1841× for the primary tumour sample, 388× for the relapse), TORRENT (custom capture panel on an Ion Torrent platform, median depths 41× and 46.5×) and WGS (whole-genome sequencing on an Illumina platform, median depths 323× and 41×). For each dataset, we uniformly selected a random subset of 250 mutations, halving the number of mutant reads for hemizygous chromosome X mutations, and adding reported driver mutations.

Cloe was run with $K \in \{3, \ldots, 7\}$ on the datasets. Model selection on ALLDNA indicated $K = 5$ as the preferred solution, followed closely by $K = 6$, which provided a closer fit to the data (Supplementary Figure 19). The inferred mutation dynamics for both models are shown in Figure 10. Whereas both model sizes could capture the trends in the data, the solution for $K = 6$ correctly identified two groups of mutations that rise in allele fraction in the relapse sample.

Our reconstruction shows a decrease in tumour burden at relapse, a single origin for all clones and branched evolution after the founding clone (Figure 11). Clone 5 and its child, clone 6, become the main clones in the relapse sample, supplanting clones 3 and 4. The founding clone appears present only at very low clonal fractions.

Matching our clones to the original clusters, we found a close correspondence (Table 4), corroborating Cloe's inference. The only misassignment is *TP53* to clone 6, which in the original study required single-cell sequencing and additional time points to identify as belonging to a separate clone. Beyond the genotypes, there was also a close match between inferred and expected clonal fractions, with a maximum absolute difference of 4%.

Cluster 6 was not identified by our model. According to the original analysis, this cluster was present at less than 5.5% clonal fraction in the primary sample, to then disappear at relapse. Such a cluster would contribute half of its clonal fraction in allele fraction, due to the heterozygosity of the mutations. We do observe that nine of the 257 mutations were not assigned to any clone. Their average allele fraction was 2.2% in the primary sample and close to 0% in the relapse (Supplementary Figure 21). Because they did not fit the dynamics of the other clones, sequencing noise was used to fit them [equation (2.6)].

Because of the high depth of sequencing for this dataset, we additionally ran Cloe with 8 clones. The solution we obtained refines what reported in Figure 11 for 6 clones: clone C5 is divided in two, matching the multiple rising allele fraction patterns, and the missing clone is identified (Supplementary Figures 22 and 23).

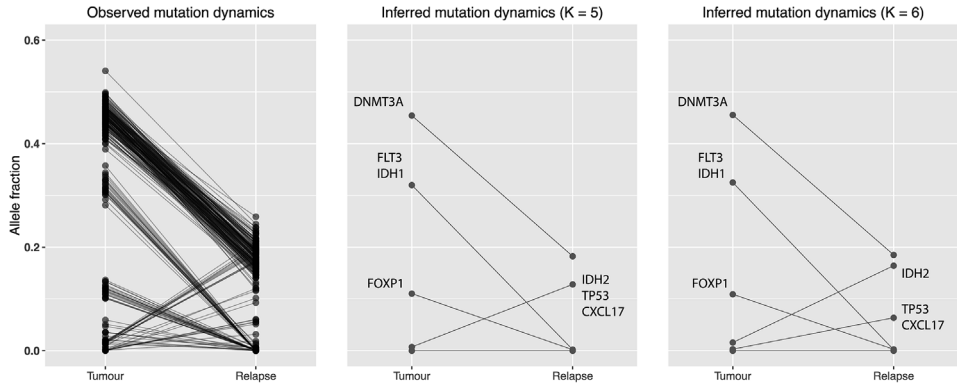

FIG. 10. *Observed and inferred mutation dynamics for* 257 *mutations from the ALLDNA dataset; driver mutations are highlighted. Left: observed allele fractions; centre: predicted allele fractions given the parameters inferred by Cloe using* 5 *clones; right: predicted allele fractions given the parameters inferred by Cloe with* 6 *clones. The allele fraction predicted for a mutation* j *in a sample* t *is given by* $\frac{1}{2}\sum_k z_{jk} f_{kt}$.

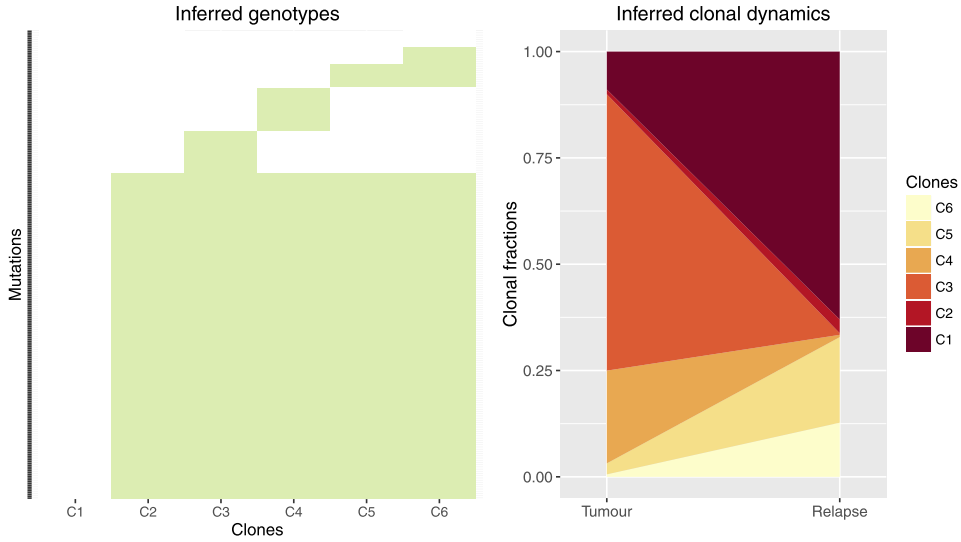

FIG. 11. *Parameters inferred by Cloe running with* 6 *clones on* 257 *mutations from the ALLDNA dataset. Genotypes are shown on the left, where green denotes presence of a mutation; clonal fractions for each clone are shown on the right.* $C1$ *is fixed as the normal contamination. The corresponding phylogeny is shown in Supplementary Figure* 20.

On the modest amount of data of the TORRENT dataset our model selection produced a more conservative estimate of the number of clones, preferring four clones. Using five or more clones improved the log-likelihood to the same extent. We compare here solutions for $K = 4$ and $K = 5$ (Figure 12).

TABLE 4
Correspondence between Cloe's inferred clones and the clusters in the original analysis by Griffith et al. (2015). While drivers are also present in the children of a clone, here we report the clone in which the mutations first appeared

Clone	Cluster	Drivers
C2	1	*DNMT3A*
C4	4	*FOXP1*
C5	3	*IDH2*
C3	2	*IDH1*
C6	5	*CXCL17*, *TP53*

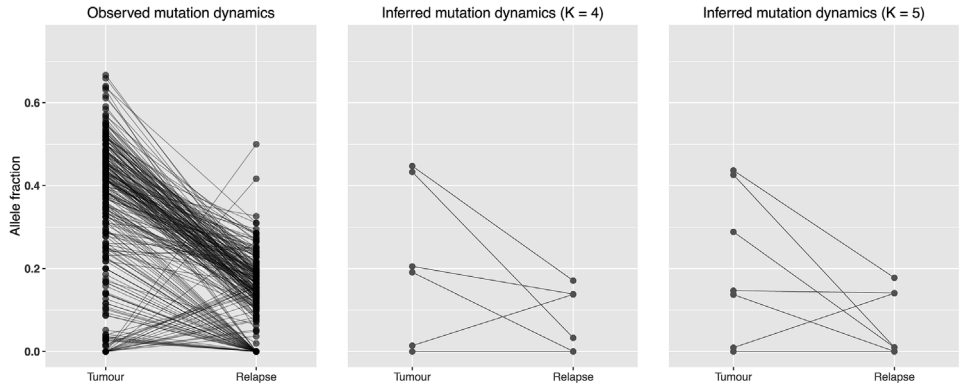

FIG. 12. *Observed and inferred mutation dynamics for* 254 *mutations from the TORRENT dataset. Left*: *observed allele fractions*; *centre*: *allele fractions inferred by Cloe with* 4 *clones*; *right*: *allele fractions inferred by Cloe with* 5 *clones*.

With three tumour clones, our model matched the main trends: two large clones in the primary sample that disappear at relapse, and one growing clone. In addition, tumour content was accurately inferred: 89% for the primary sample and 44% for the relapse sample, compared to the expected values of 91% and 37%. The addition of a fourth tumour clone ($K = 5$) allows a better disambiguation of the clones present in the primary, while the spread of allele fractions in the relapse sample makes it difficult to distinguish two rising clones.

Identifying seven clones including the normal in two samples with a median depth less than $45\times$ is an arduous task. Griffith et al. (2015) show that SciClone detects four clones up to around $100\times$ depth using all mutations on the platinum list. While Cloe prefers four clones using a subset of mutations at a depth of $45\times$, it is capable of splitting the observed dynamics further, obtaining closer approximations of the real clonal structure.

Finally, for the WGS dataset, Cloe's solution with 5 clones obtained the highest posterior probability, while 6 and 7 clones obtained closer fits to the data (Supplementary Figure 24). With four tumour clones, Cloe identified three decreasing groups of mutations and one group that arose at relapse. This matches the observed dynamics, as the low depth at relapse accounts for a larger spread of the allele fractions that confounds the identification of two rising clones (Supplementary Figure 25). Interestingly, the addition of another clone, rather than fitting this low-depth relapse data, matches a fourth group of mutations present only in the primary around 5% clonal fraction. These mutations overlap with the unassigned mutations in the ALLDNA dataset and the inferred clone does not harbour additional driver mutations other than *DNMT3A*, which derives from its parent.

With this case study we applied Cloe to a scenario with two samples, highlighting the difficulties of automatic model selection, especially when trying to identify a large number of clones with a moderate amount of data.

The running parameters for the two case studies differ from the ones listed in Table 1 in that we used five chains with $\Delta T = 0.25$. In addition, for the AML datasets we ran 50,000 iterations of our sampler with $\mu = 0.2$, $\rho = 0.04$ and $\varepsilon = 0.001$.

5. Discussion. As tumour sequencing data grow in depth and breadth, the question of tumour heterogeneity will continue to be focal. In this study we presented Cloe, a novel latent feature model for direct clonal reconstruction. Our model discovers genotypes in the data by assigning observed mutations to latent features (clones) guided by a latent phylogeny. This phylogenetic deconvolution sets Cloe apart from other direct reconstruction methods [Fischer et al. (2014); Zare et al. (2014); Sengupta et al. (2015)]. Compared to indirect reconstruction methods, our algorithm can handle multiple primary tumours, the loss of mutations and convergent evolution. In particular, to our knowledge, this is the first method to allow and penalise convergent evolution.

Our study on simulated data showed a good performance of our MCMCMC algorithm. However, tuning the MCMCMC parameters in order to correctly explore the spiked posterior landscape is not trivial. We empirically found parameters that would allow the chains to mix well. Regions of high posterior probability are quickly reached, yet finding the right peak may be a slow process, complicated by each biological constraint on the model. Many parameters can be tuned in our model. We sought values that would work well for both simulated data and our validation data. Tuning the MCMCMC parameters to each dataset independently, thus optimising the exploration of the posterior space, might further improve results.

In our definition of the tree we assume that multiple primary tumours are less likely to occur than tumours with a single origin. If our understanding of clonal

evolution were to suggest otherwise, the definition of the tree may be simplified to a discrete uniform distribution, giving equal weight to a single origin or multiple ones.

Limitations. The main limitation of our method is the restriction to mutations from copy-number neutral regions. Whereas this may be amenable to certain types of cancer (e.g., mutation-driven rather than copy-number-driven cancers), it may preclude the analysis of more genomically rearranged tumours.

In contrast to some models described in the literature, our method does not include the number of clones as a parameter. Instead, Cloe must be run for various choices of K, and the best solution in terms of posterior probability will indicate the number of clones with good accuracy. On our simulation and validation datasets our model was indeed able to identify the correct number of clones in 58/59 cases.

As shown in the case studies, model selection may not be trivial. We thus recommend manual review of the inferred parameters for various model sizes to ensure that the results of the inference are robust.

Analysing hundreds of mutations can result in a high computational burden. This limitation could be alleviated by preprocessing the input data, grouping mutations that exhibit similar dynamics throughout the samples. One way to do this is via a Chinese Restaurant Process with a product of binomials; mutant read counts and depths for all mutations in a cluster could then be summed and analysed as a single unit.

Extensions. We see several avenues for future extensions. At the theoretical level, future work should focus on optimising the inference and extending this framework to arbitrary copy numbers. Also, to address the model selection problem, the phylogenetic latent feature model could be rephrased in a nonparametric perspective. In terms of applications, our model could also be applied to epigenetics: by appropriately changing the likelihood function, Cloe could deconvolute methylation data into evolutionarily related epigenotypes.

In summary, Cloe is a rigorous and flexible framework for clonal deconvolution of cancer genomes that achieves high accuracy in benchmarking studies and leads to important insights into tumour evolution in clinical case studies.

Acknowledgements. We would like to acknowledge the support of The University of Cambridge, Cancer Research UK and Hutchison Whampoa Limited.

We wish to thank Marta Grzelak and James Hadfield for their assistance with sequencing, and Malvina Josephidou for discussions on the CRP clustering of mutations.

SUPPLEMENTARY MATERIAL

Supplement A: Supplementary information (DOI: 10.1214/16-AOAS986SUPPA; .pdf). Supplementary text and figures.

Supplement B: Source code of the analyses (DOI: 10.1214/16-AOAS986SUPPB; .zip). This package contains scripts, data (in the form of matrices of mutant read counts and depths) analysed in this article and a version of Cloe to reproduce the findings.

REFERENCES

APARICIO, S. and CALDAS, C. (2013). The implications of clonal genome evolution for cancer medicine. *N. Engl. J. Med.* **368** 842–851.

BEERENWINKEL, N., SCHWARZ, R. F., GERSTUNG, M. and MARKOWETZ, F. (2015). Cancer evolution: Mathematical models and computational inference. *Syst. Biol.* **64** e1–e25.

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.

EL-KEBIR, M., OESPER, L., ACHESON-FIELD, H. and RAPHAEL, B. J. (2015). Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. *Bioinform.* **31** i62–i70.

FEARON, E. R. and VOGELSTEIN, B. (1990). A genetic model for colorectal tumorigenesis. *Cell* **61** 759–767.

FISCHER, A., VÁZQUEZ-GARCÍA, I., ILLINGWORTH, C. J. and MUSTONEN, V. (2014). High-definition reconstruction of clonal composition in cancer. *Cell Rep.* **7** 1740–1752.

FORSHEW, T., MURTAZA, M., PARKINSON, C., GALE, D., TSUI, D. W. Y., KAPER, F., DAWSON, S.-J., PISKORZ, A. M., JIMENEZ-LINAN, M., BENTLEY, D., HADFIELD, J., MAY, A. P., CALDAS, C., BRENTON, J. D. and ROSENFELD, N. (2012). Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci. Transl. Med.* **4** 136ra68.

GERLINGER, M., ROWAN, A. J., HORSWELL, S., LARKIN, J., ENDESFELDER, D., GRONROOS, E., MARTINEZ, P., MATTHEWS, N., STEWART, A., TARPEY, P., VARELA, I., PHILLIMORE, B., BEGUM, S., MCDONALD, N. Q., BUTLER, A., JONES, D., RAINE, K., LATIMER, C., SANTOS, C. R., NOHADANI, M., EKLUND, A. C., SPENCER-DENE, B., CLARK, G., PICKERING, L., STAMP, G., GORE, M., SZALLASI, Z., DOWNWARD, J., FUTREAL, P. A. and SWANTON, C. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366** 883–892.

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

GHAHRAMANI, Z. and GRIFFITHS, T. L. (2005). Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems* 475–482.

GRIFFITH, M., MILLER, C. A., GRIFFITH, O. L., KRYSIAK, K., SKIDMORE, Z. L., RAMU, A., WALKER, J. R., DANG, H. X., TRANI, L., LARSON, D. E., DEMETER, R. T., WENDL, M. C., MCMICHAEL, J. F., AUSTIN, R. E., MAGRINI, V., MCGRATH, S. D., LY, A., KULKARNI, S., CORDES, M. G., FRONICK, C. C., FULTON, R. S., MAHER, C. A., DING, L., KLCO, J. M., MARDIS, E. R., LEY, T. J. and WILSON, R. K. (2015). Optimizing cancer genome sequencing and analysis. *Cell Syst.* **1** 210–223.

HEAUKULANI, C., KNOWLES, D. A. and GHAHRAMANI, Z. (2014). Beta diffusion trees and hierarchical feature allocations. Preprint. Available at arXiv:1408.3378.

JI, Y. (2016). Biostatistics and Bioinformatics Lab—Software. Available at http://health.bsd.uchicago.edu/yji/soft.html. Accessed: 2016-02-05.

JIAO, W., VEMBU, S., DESHWAR, A. G., STEIN, L. and MORRIS, Q. (2014). Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinform.* **15** 1.

MARASS, F., MOULIERE, F., YUAN, K., ROSENFELD, N. and MARKOWETZ, F. (2016). Supplement to "A phylogenetic latent feature model for clonal deconvolution." DOI:10.1214/16-AOAS986SUPPA, DOI:10.1214/16-AOAS986SUPPB.

MILLER, K. T., GRIFFITHS, T. and JORDAN, M. I. (2012). The phylogenetic Indian buffet process: A non-exchangeable nonparametric prior for latent features. Preprint. Available at arXiv:1206.3279.

MILLER, C. A., WHITE, B. S., DEES, N. D., GRIFFITH, M., WELCH, J. S., GRIFFITH, O. L., VIJ, R., TOMASSON, M. H., GRAUBERT, T. A., WALTER, M. J. et al. (2014). SciClone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol.* **10** e1003665.

NIK-ZAINAL, S., LOO, P. V., WEDGE, D. C., ALEXANDROV, L. B., GREENMAN, C. D., LAU, K. W., RAINE, K., JONES, D., MARSHALL, J., RAMAKRISHNA, M., SHLIEN, A., COOKE, S. L., HINTON, J., MENZIES, A., STEBBINGS, L. A., LEROY, C., JIA, M., RANCE, R., MUDIE, L. J., GAMBLE, S. J., STEPHENS, P. J., MCLAREN, S., TARPEY, P. S., PAPAEMMANUIL, E., DAVIES, H. R., VARELA, I., MCBRIDE, D. J., BIGNELL, G. R., LEUNG, K., BUTLER, A. P., TEAGUE, J. W., MARTIN, S., JÖNSSON, G., MARIANI, O., BOYAULT, S., MIRON, P., FATIMA, A., LANGERØD, A., APARICIO, S. A. J. R., TUTT, A., SIEUWERTS, A. M., BORG, Å., THOMAS, G., SALOMON, A. V., RICHARDSON, A. L., BØRRESEN-DALE, A.-L., FUTREAL, P. A., STRATTON, M. R., CAMPBELL, P. J. and BREAST CANCER WORKING GROUP OF THE INTERNATIONAL CANCER GENOME CONSORTIUM (2012). The life history of 21 breast cancers. *Cell* **149** 994–1007.

NOWELL, P. C. (1976). The clonal evolution of tumor cell populations. *Science* **194** 23–28.

RONQUIST, F., HUELSENBECK, J. P. and TESLENKO, M. (2005). MrBayes version 3.2 Manual: Tutorials and Model Summaries. http://mrbayes.sourceforge.net/mb3.2_manual.pdf. [Online; accessed 29 June 2016].

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

SCHUH, A., BECQ, J., HUMPHRAY, S., ALEXA, A., BURNS, A., CLIFFORD, R., FELLER, S. M., GROCOCK, R., HENDERSON, S., KHREBTUKOVA, I. et al. (2012). Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood* **120** 4191–4196.

SCHWARZ, R. F., NG, C. K., COOOKE, S. L., NEWMAN, S., TEMPLE, J., PISKORZ, A. M., GALE, D., SAYAL, K., MURTAZA, M., BALDWIN, P. J., ROSENFELD, N., EARL, H. M., SALA, E., JIMENEZ-LINAN, M., PARKINSON, C. A., MARKOWETZ, F. and BRENTON, J. D. (2015). Spatial and temporal heterogeneity in high-grade serous ovarian cancer: A phylogenetic reconstruction. *PLoS Med* **12** e1001789.

SENGUPTA, S., WANG, J., LEE, J., MÜLLER, P., GULUKOTA, K., BANERJEE, A. and JI, Y. (2015). BayClone: Bayesian nonparametric inference of tumor subclones using NGS data. In *Pacific Symposium on Biocomputing* **20** 467. World Scientific.

STRATTON, M. R., CAMPBELL, P. J. and FUTREAL, P. A. (2009). The cancer genome. *Nature* **458** 719–724.

YUAN, K., SAKOPARNIG, T., MARKOWETZ, F. and BEERENWINKEL, N. (2015). BitPhylogeny: A probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biol.* **16** 36.

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

F. MARASS
F. MOULIERE
N. ROSENFELD
F. MARKOWETZ
CANCER RESEARCH UK CAMBRIDGE INSTITUTE
UNIVERSITY OF CAMBRIDGE
LI KA SHING CENTRE
ROBINSON WAY
CAMBRIDGE, CB2 0RE
UNITED KINGDOM
E-MAIL: francesco.marass@cruk.cam.ac.uk
florent.mouliere@cruk.cam.ac.uk
nitzan.rosenfeld@cruk.cam.ac.uk
florian.markowetz@cruk.cam.ac.uk

K. YUAN
SCHOOL OF COMPUTING SCIENCE
SIR ALWYN WILLIAMS BUILDING
UNIVERSITY OF GLASGOW
GLASGOW, G12 8RZ
SCOTLAND
UNITED KINGDOM
E-MAIL: ke.yuan@glasgow.ac.uk