Characterization of tumor heterogeneity by latent haplotypes: a sequential Monte Carlo approach

Oyetunji E Ogundijo ¹, Xiaodong Wang ¹

¹ Department of Electrical Engineering, Columbia University, New York, NY, United States

Corresponding Author: Xiaodong Wang
Email address: wangx@ee.columbia.edu

Tumor samples obtained from a single cancer patient spatially or temporally often consist of varying cell populations, each harboring distinct mutations that uniquely characterize its genome. Thus, in any given samples of tumor, having more than two haplotypes, defined as a scaffold of single nucleotide variants (SNVs) on the same homologous genome, is an evidence of heterogeneity because humans are diploid and we would therefore only observe up to two haplotypes if all cells in a tumor sample were genetically homogeneous. We characterize tumor heterogeneity by latent haplotypes and present state-space formulation of the feature allocation model for estimating the haplotypes and their proportions in the tumor samples. We develop an efficient sequential Monte Carlo (SMC) algorithm that estimates the states and the parameters of our proposed state-space model, which are equivalently the haplotypes and their proportions in the tumor samples. The sequential algorithm produces more accurate estimates of the model parameters when compared with existing methods. Also, because our algorithm processes the variant allele frequency (VAF) of a locus as the observation at a single time-step, VAF from newly sequenced candidate SNVs from next-generation sequencing (NGS) can be analyzed to improve existing estimates without re-analyzing the previous datasets, a feature that existing solutions do not possess.
Characterization of tumor heterogeneity by latent haplotypes: a sequential Monte Carlo approach

Oyetunji E. Ogundijo1 and Xiaodong Wang1

1Department of Electrical Engineering, Columbia University, New York, 10027, USA.

Corresponding author:
Xiaodong Wang1
Email address: wangx@ee.columbia.edu

ABSTRACT

Tumor samples obtained from a single cancer patient spatially or temporally often consist of varying cell populations, each harboring distinct mutations that uniquely characterize its genome. Thus, in any given samples of tumor, having more than two haplotypes, defined as a scaffold of single nucleotide variants (SNVs) on the same homologous genome, is an evidence of heterogeneity because humans are diploid and we would therefore only observe up to two haplotypes if all cells in a tumor sample were genetically homogeneous. We characterize tumor heterogeneity by latent haplotypes and present state-space formulation of the feature allocation model for estimating the haplotypes and their proportions in the tumor samples. We develop an efficient sequential Monte Carlo (SMC) algorithm that estimates the states and the parameters of our proposed state-space model, which are equivalently the haplotypes and their proportions in the tumor samples. The sequential algorithm produces more accurate estimates of the model parameters when compared with existing methods. Also, because our algorithm processes the variant allele frequency (VAF) of a locus as the observation at a single time-step, VAF from newly sequenced candidate SNVs from next-generation sequencing (NGS) can be analyzed to improve existing estimates without re-analyzing the previous datasets, a feature that existing solutions do not possess.

INTRODUCTION

Tumors contain multiple, genetically diverse subclonal populations of cells, each subclone harboring distinct mutations that uniquely characterize its genome (Marusyk and Polyak, 2010; Meacham and Morrison, 2013; Heppner, 1984). Tumor subclones often evolve from a single ancestral population (Hughes et al., 2014; Gerlinger et al., 2012; Visvader, 2011; Nowell, 1976). The genetic diversities that distinguish these subclones are a direct result of evolutionary processes that drive tumor progression, especially the series of somatic genetic variants which arise stochastically by a sequence of randomly acquired mutations (Hanahan and Weinberg, 2011, 2000).

Identifying and characterizing tumor subclonality is crucial for understanding the evolution of tumor cells. The knowledge is important for designing more effective treatments for cancer, especially in avoiding cancer relapse and chemotherapy resistance (Garraway and Lander, 2013). For instance, research has shown the links between the presence of driver mutations within subclones and the adverse clinical outcomes (Landau et al., 2013).

Tumor heterogeneity has been studied using the NGS technology (Lee et al., 2016; Gerlinger et al., 2012). Somatic mutations are quantified using whole exome sequencing (WES) and whole genome sequencing (WGS) of samples (Marusyk et al., 2012), and can be explained by differences in genomes of subclones and the varying proportions of these subclones (Lee et al., 2016; Landau et al., 2013; Russnes et al., 2011; Navin et al., 2010; Marusyk and Polyak, 2010). One method to assess the heterogeneity of a given tumor is to probe individual cell using fluorescent markers (Navin et al., 2010; Irish et al., 2004) and another is to perform single cell sequencing (Xu et al., 2012; Hou et al., 2012; Navin et al., 2011). These approaches have several limitations that prevent their wider usage in examining and quantifying the level of heterogeneity in a given sample. For instance, evaluating large number of cells by probing them...
individually can be expensive. Also, the spatial positioning of the cells relative to other cells in the tumor are lost in the process.

In the literature, a few computational methods have been proposed to explain the inherent structure of tumor heterogeneity. For instance, (Larson and Fridley, 2013) and (Su et al., 2012) viewed a tumor sample as a mixture of tumor cells and normal cells. Although their method can estimate tumor purity levels for paired tumor-normal tissue sample using DNA sequencing data, unpaired and multiple tumor samples are not considered. A more prominent approach is the arrangement of single nucleotide variants (SNVs) in clusters using clustering models such as the Dirichlet Process (DP) (Roth et al., 2014; Jiao et al., 2014; Shah et al., 2009; Ding et al., 2010; Bashashati et al., 2013). Although the clustered SNVs provide some information about tumor heterogeneity, the inference does not directly identify subclones or haplotypes in the tumor samples.

More recently, (Lee et al., 2016; Xu et al., 2015) proposed a feature allocation model for estimating tumor heterogeneity by estimating haplotypes and their proportions in the tumor. This model provides insights on how haplotypes may be distributed within a tumor, using WGS data measuring variant allele frequencies (VAFs) at SNVs. Mathematically, the model can be interpreted as blind matrix factorization. A matrix of expected VAFs at SNVs for different samples is decomposed into a binary matrix of haplotypes (with an unknown number of columns, the exact number to be determined by the data), and a matrix of proportions of haplotypes. This model offers certain modeling advantages over the clustering approach: (i) overlapping SNVs can be shared among different subclones, and (ii) non-overlapping SNV clusters (according to the cellular prevalence) are not used as the building block for subclones i.e., instead of first estimating the SNV clusters and then constructing subclones based on clusters, the model provides a way to infer the subclonal structure based on haplotypes. To make an inference on the haplotypic structure in tumor samples, (Lee et al., 2016) and (Xu et al., 2015) proposed a Markov chain Monte Carlo (MCMC-based) and a maximum a posteriori (MAP-based) asymptotic derivations (MAD) algorithms respectively. However, if more VAFs are available for newly called SNV(s), both algorithms have to be restarted in order to incorporate the newly called SNV(s). Moreover, MCMC approach in general as previously shown in (Nguyen et al., 2016; Jaser et al., 2007), is plagued with some inherent issues which often limit its performance: (i) sometimes, it is difficult to assess when the Markov chain has reached its stationary regime of interest (ii) requirement of burn-in period and thinning interval, and most importantly, (iii) if the target distribution is highly multi-modal, MCMC algorithms can easily become trapped in local modes.

In this paper, we consider the feature allocation modeling approach in analyzing the WGS data measuring VAFs at SNVs, and present an efficient sequential Monte Carlo (SMC) algorithm (Doucet et al., 2001, 2000) for estimating the binary matrix of haplotypes and the proportions in the samples. Specifically, we formulate the feature allocation problem using a state-space. We consider the following in our state-space framework (i) the rows of the haplotype binary matrix are considered as the states of the system, exploiting the sequential construction of a binary matrix with an unknown number of columns using the Indian Buffet Process (IBP), (ii) the proportions matrix and other parameters are considered as the parameters of the model, (iii) the observed VAF at each SNV are processed, for all samples at a time. SMC is a very powerful algorithm that belongs to a broad class of recursive filtering techniques (Ogundijo et al., 2017; Ogundijo and Wang, 2017). Instead of processing all the observations at once, observations are processed sequentially, one after the other. The posterior probability density function (PDF) of every state is computed every time a measurement is observed. The posterior distributions of the variables of interest are approximated with a set of properly weighted particles (Doucet et al., 2001). With the SMC methods, we can treat, in a principled way, any type of probability distribution, nonlinearity and non-stationarity (Kitagawa, 1998, 1996). We compare the proposed SMC algorithm with the existing MCMC-based and MAP-based algorithms. In terms of the accuracy of estimates of the matrix of haplotypes and the matrix of proportions, denoted as $Z$ and $W$, our proposed SMC method produces better results.

The remainder of this paper is organized as follows. In Section 2, we describe the system model and problem formulation. We describe the general principle of the SMC filtering algorithms, and then derive our proposed SMC algorithm for estimating the mutational profile of each haplotype and their proportions in the samples, in a sequential fashion. In Section 3, we investigate the performance of the proposed method using simulated datasets and the chronic lymphocytic leukemia (CLL) datasets, the real tumor samples obtained from three patients in (Schuh et al., 2012). Finally, Section 4 concludes the paper.
In this paper, we use the following notations:

1. \( p(\cdot) \) and \( p(\cdot|\cdot) \) denote a probability density function (PDF) and a conditional PDF, respectively.
2. \( P(\cdot) \) and \( P(\cdot|\cdot) \) denote a probability and a conditional probability mass function, respectively.
3. \( \mathcal{N}(\mu, \sigma^2) \) denotes a Gaussian distribution with mean \( \mu \) and variance \( \sigma^2 \).
4. Binomial\( (n, p) \) denotes a binomial distribution with \( n \) number of trials and \( p \) probability of success. \( \text{Binomial}(1, p) = \text{Bern}(p) \), i.e., Bernoulli distribution with success probability \( p \).
5. Pois\( (\lambda) \) denotes a Poisson distribution with mean parameter \( \lambda \).
6. Beta\( (\alpha_1, \beta_1) \) denotes a beta distribution with shape parameters \( \alpha_1 \) and \( \beta_1 \).
7. Gam\( (\alpha_0, \beta_0) \) denotes a gamma distribution with shape parameter \( \alpha_0 \) and rate parameter \( \beta_0 \).
8. Dir\( (\alpha) \) denotes a Dirichlet distribution with a vector of concentration parameters \( \alpha \), and \( \hat{\lambda} \) denotes the estimate of variable \( x \).

**SYSTEM MODEL AND PROBLEM FORMULATION**

In an NGS experiment designed to probe the heterogeneity of a tumor sample, two matrices \( Y \) and \( V \), each with dimension \( T \times S \), of count data are often observed, where \( y_{ts} \) and \( v_{ts} \) denote the elements in the \( i^{th} \) row and \( s^{th} \) columns of \( Y \) and \( V \), respectively. At the genomic position of SNV \( t \) for tissue sample \( s \), \( y_{ts} \) denotes the number of reads that bear a variant sequence and \( v_{ts} \) denotes the total number of reads, \( t = 1, \ldots, T, s = 1, \ldots, S \). In summary, the datasets are count data for \( T \) SNVs and \( S \) samples. We follow the binomial sampling framework in (Lee et al., 2016; Zare et al., 2014; Xu et al., 2015) to model the count data:

\[
y_{ts} \overset{\text{ind.}}{\sim} \text{Binomial}(v_{ts}, p_{ts}), \quad t = 1, \ldots, T, \quad s = 1, \ldots, S, \tag{1}
\]

where \( p_{ts} \) are the success probabilities and equivalently the expected VAFs, given by:

\[
p_{ts} = w_{0s}p + \sum_{c=1}^{C} z_{tc}w_{cs}, \quad t = 1, \ldots, T, \quad s = 1, \ldots, S, \tag{2}
\]

where \( C \) denotes the unknown number of distinct haplotypes in the tumor samples, \( z_{tc} \in \{0, 1\} \) denotes an indicator of the event that SNV \( t \) bears a variant sequence for haplotype \( c \) and \( w_{cs} \) denotes the proportion of haplotype \( c \) in sample \( s \). The term \( \sum_{c=1}^{C} z_{tc}w_{cs} \) explains \( p_{ts} \) as arising from sample \( s \) being composed of a mix of hypothetical haplotypes which include a mutation for SNV \( t \) (\( z_{tc} = 1 \)), or do not include a mutation for SNV \( t \) (\( z_{tc} = 0 \)). In addition, there is a background haplotype \( c = 0 \), which includes all SNVs. The background haplotype accounts for experimental noise and haplotypes that appear with negligible abundance. The first term in (2) relates to this background haplotype, where \( p \) denotes the relative frequency of observing a mutation at an SNV due to noise and artifact, assuming equal frequency for all SNVs, and \( w_{0s} \) denotes the proportion in sample \( s \) (Lee et al., 2016). In (2), if we (i) collect the indicators \( z_{tc} \) in an \( T \times C \) binary matrix \( Z \), (ii) collect all \( p \)’s in a \( T \)-dimensional column vector \( p \) and (iii) collect the proportions \( w_{0s} \) and \( w_{cs} \) in an \( C \times S \) matrix \( W \) of probabilities, where \( C = C + 1 \) and each column of \( W \) sums to unity, then we can write (2) as \( p_{ts} = Z' \cdot W \), where \( p_{ts} \) denotes the matrix of success probabilities and equivalently, the matrix of expected VAFs and \( Z' = [p \ Z] \). If the expected VAFs are approximated with the observed VAFs, we can directly solve the matrix factorization problem but instead the observed VAFs are modeled with a probability distribution in (1). However, it should be noted that the number of latent haplotypes \( C \) is unknown, and this leaves the number of columns in \( Z \) and equivalently, the number of rows in \( W \) unknown, left to be estimated from the data.

Our goal is to perform a joint inference on \( C, Z, W \) and \( p \), all of which explain the heterogeneity in the tumor samples, using the observed VAFs of SNVs described by the matrices \( Y \) and \( V \), the input data. To do this, we describe the system using a state-space model and then derive an efficient SMC algorithm to estimate all the hidden states and the model parameters in our model, in a sequential fashion. Our analysis is restricted to mutations in copy-number neutral regions.
**State-Space Formulation**

Our state-space formulation of the problem exploits the sequential construction of $Z$. Specifically, we consider the $t^{th}$ row of the data matrix $Y$ and $V$ as the new observation at time $t$ of our state-space model, treat the $t^{th}$ row of the binary matrix $Z$ as the hidden state at time $t$, and $W$ and $p$ as the model parameters. Before explicitly stating the state transition and the observation models, we succinctly describe the prior distribution on a “left-ordered” binary matrix (i.e., ordering the columns of the binary matrix from left to right by the magnitude of the binary expressed by that column, taking the first row as the most significant bit) with a finite number of rows and an unknown number of columns. The prior distribution as detailed in (Griffiths and Ghahramani, 2011; Ghahramani and Griffiths, 2006) is given by:

$$P(Z) = \frac{\alpha^{C_+}}{\prod_{h=1}^{C_+} C_h} \exp\{-\alpha H_T\} \prod_{c=1}^{C_h} \frac{(T-m_c)!(m_c-1)!}{T!},$$  

where $C_+$ denotes the number of columns of $Z$ with non-zero entries, $m_c$ denotes the number of 1’s in column $c$, $T$ denotes the number of rows in $Z$, $H_T = \sum_{t=1}^{T} 1/t$ denotes the $T^{th}$ harmonic number, and $C_h$ denotes the number of columns in $Z$ that when read top-to-bottom form a sequence of 1’s and 0’s corresponding to the binary representation of the number $h$.

The distribution in (3) can be derived as the outcome of a **sequential generative process** called the **Indian buffet process** (Griffiths and Ghahramani, 2011; Doshi-Velez, 2009). Imagine that in an Indian buffet restaurant, we have $T$ customers who arrive at the restaurant sequentially, one after the other. The first customer walks into the restaurant and loads her plate from the first $c_1$ dishes, where $c_1 = \text{Pois}(\alpha)$ (as is similar to the dispersion parameter in the Chinese Restaurant Process (Zhang, 2008)). The $t^{th}$ customer will choose a particular dish according to the popularity of the dish, i.e., choosing a dish with probability $m_c/t$, where $m_c$ denotes the number of people who have previously chosen the $c^{th}$ dish, and in addition, chooses $\text{Pois}(\alpha/t)$ new dishes as well. Now, if we record the choices of each customer on each row of a matrix, where each column corresponds to a dish on the buffet (1 if the dish is chosen, and 0 if not), then such a binary matrix is a draw from the distribution in (3) (Ghahramani and Griffiths, 2006). The entire process is sequential because the choices made by the $t^{th}$ customer are dependent only on the choices made by the $t-1$ preceeding customers and not on the remaining $T-t$ customers.

In our case, the dishes in the IBP are the haplotypes in the tumor samples, the SNVs are the customers who arrive at the restaurant sequentially, one after the other. The algorithm to sample from (4) is presented in **Algorithm 1** in the Supplementary Material. Note that in the algorithm, $Z_{t-1}$ is implicitly constructed from $Z_t$ and if in the process, new non-zero column(s) is/are introduced in $Z_t$ ($\text{Pois}(\alpha/t) > 0$), then new row(s) will be added to $W$ as well. On the other hand, if the numbers of non-zero columns in $Z_{t-1}$ and $Z_t$ are the same, then the number of rows in $W$ does not change between $t-1$ and $t$. To account for any possible change of dimension in $W$, we re-parameterize matrix $W$. Specifically, we rewrite $w_{cs} = \theta_{cs}/\sum_{d=0}^{T-1} \theta_{cd}$, which implies that we estimate $\theta_{cs}$ and compute $w_{cs}$ from the estimates of $\theta_{cs}$. This procedure ensures that each column of $W$ sums to unity at any point in time during the process.

Moreover, since we are interested in the final estimates of the model parameters $W$ and $p$, we create artificial dynamics for these parameters using the random walk model, i.e.,

$$\phi_t \sim p(\phi_t | \phi_{t-1}) = N(\phi_{t-1}, \sigma^2),$$

$$\phi_t \in \{p, \theta_{cs}, c = 0, 1, ..., C, s = 1, ..., S\},$$

where $\sigma$ denotes the standard deviation. Hence, (4)-(5) fully describe the system state transition.

Similarly, the observation at time $t$ is given by:

$$y_t \sim P(y_t | Z_{1:t}, W, p) = P(y_t | z_t, W, p) = \prod_{j=1}^{S} \text{Binomial}(y_{ts} | v_{ts}, p_{ts}),$$
where \( y_t \) denotes the observation at time \( t \) (which is conditionally independent of the previous observations \( Y_{t-1} \) given the state \( z_t \)), i.e., the \( t^{th} \) row of \( Y \). (6) fully describes the measurement model for the system. Finally, (4) - (6) completely describe our proposed state-space model for estimating the mutational profile and the proportion of each haplotype in the tumor samples.

The SMC Algorithm

In this section, we briefly describe the SMC filtering framework that will be employed to estimate the states and the parameters of our state-space model (Doucet et al., 2000, 2001). Consider the general dynamic system with hidden state variable \( x_t \), in our case, consisting of discrete variables \( z_t \) and continuous variables \( \phi_t, \phi_t \in \{ p_0, \phi_t, c = 0, \ldots, C, s = 1, \ldots, S \} \), and measurement variable \( y_t \), where there is an initial state model \( p(x_0) \), and \( \forall t \geq 1 \), a state transition model given in (4) - (5) and an observation model given in (6). The sequence \( X_t = \{x_1, x_2, \ldots, x_t\} \) is not observed and we want to estimate it for each time \( t \), given that we have the observations \( Y_t = \{y_1, y_2, \ldots, y_t\} \).

Our goal is to approximate the posterior distribution of states \( p(X_t|Y_t) \) using particles drawn from it. However, getting such particles from \( p(X_t|Y_t) \) is usually not feasible. We can still implement an estimate using \( N \) particles, \( \{X_i^t\}_{i=1}^N \), taken from another distribution, \( q(X_t|Y_t) \), whose support includes the support of \( p(X_t|Y_t) \) (importance sampling theorem). For the approximation, the weights associated with the particles are calculated as follows:

\[
\tilde{w}_i^t = \frac{p(X_i^t|Y_t)}{q(X_i^t|Y_t)} \quad \text{and} \quad w_i^t = \frac{\tilde{w}_i^t}{\sum_{m=1}^N \tilde{w}_m^t}, \quad i = 1, \ldots, N. \tag{7}
\]

Thus, the pair \( \{X_i^t, w_i^t\}_{i=1}^N \) is said to be properly weighted with respect to the distribution \( p(X_t|Y_t) \), and the approximation \( \tilde{p}(X_t|Y_t) \) is then given by:

\[
\tilde{p}(X_t|Y_t) = \sum_{i=1}^N w_i^t \delta(X_t - X_t^i), \quad \text{where} \quad \delta(u) = \begin{cases} 1, & \text{if } u = 0 \\ 0, & \text{otherwise}. \end{cases} \tag{8}
\]

Similar to the above importance sampling theory, a sequential algorithm can be obtained as follows. First, we express the full posterior distribution of states \( X_t \) given the observations \( Y_t \) as follows:

\[
p(X_t|Y_t) \propto p(y_t|X_t, Y_{t-1})p(X_t|Y_{t-1}) = p(y_t|X_t, Y_{t-1})p(X_{t-1}|Y_{t-1}p(X_t|Y_{t-1}). \tag{9}
\]

At time \( t \), we desire to obtain \( N \) weighted particles from \( p(X_t|Y_t) \), which is not feasible. Instead, we define an importance distribution \( q(X_t|Y_t) = q(x_t|X_{t-1}, Y_t)q(X_{t-1}|Y_{t-1}) \), where particles can be obtained from, and then calculate the associated unnormalized importance weights as follows:

\[
\tilde{w}_i^t = \frac{p(y_t|X_i^t, Y_{t-1})p(x_i^t|Y_{t-1})}{q(x_i^t|X_i^t, Y_t)} \tag{10}
\]

Assuming that at time \( t - 1 \), we have already drawn the particles \( \{X_i^{t-1}\}_{i=1}^N \) from the importance distribution \( q(X_{t-1}|Y_{t-1}) \) and the corresponding normalized weights written as follows:

\[
w_{i-1}^t \propto \frac{p(X_i^{t-1}|Y_{t-1})}{q(X_i^{t-1}|Y_{t-1})}, \quad i = 1, \ldots, N, \tag{11}
\]

we can now draw particles \( \{X_i^t\}_{i=1}^N \) from the importance distribution \( q(X_t|Y_t) \) by drawing the new state particles for the time step \( t \) as \( X_t^i \sim q(x_t|X_{t-1}, Y_t) \), and write \( \{X_i^t\}_{i=1}^N = \{x_t^i, X_i^t\}_{i=1}^N \). If we substitute (11) into (10), the weights at time \( t \) satisfy the recursion:

\[
\tilde{w}_i^t \propto \frac{p(y_t|X_i^t, Y_{t-1})p(x_i^t|Y_{t-1})}{q(x_i^t|X_i^t, Y_t)}, \quad i = 1, \ldots, N, \tag{12}
\]

and then the weights are normalized to sum to unity.

So far, we have presented a generic sequential sampling algorithm. We obtain the optimal importance distribution by setting \( q(x_i^t|X_{t-1}, Y_t) = p(x_i^t|X_{t-1}, Y_t) \), and the weights in (12) become \( \tilde{w}_i^t \propto p(y_t|X_{t-1}, Y_t) \).
Algorithm 1 SMC Algorithm for Characterizing Tumor Heterogeneity

Input: $Y, V$

1: Initialize $N$ particles $\{z_i^0, p_i^0, W_i^0\}_{i=1}^N$
2: for $t = 1, ..., T$ do
3:   for $i = 1, ..., N$ do
4:     Sample $z_i^t$ from $Z_{i-1}^t$ using Algorithm 1 in the Supplementary Material.
5:     $n_1 \leftarrow$ number of columns in $Z_{i-1}^t$
6:     $n_2 \leftarrow$ length of $z_i^t$
7:     $d \leftarrow (n_2 - n_1)$
8:     if $d = 0$ then
9:       \[
10:       Z_i^t \leftarrow \begin{bmatrix} Z_{i-1}^t & z_i^t \end{bmatrix}
11:       \]
12:     else
13:       Sample $W_i^t$ using (5)
14:       \[
15:       Z_i^t \leftarrow \begin{bmatrix} Z_{i-1}^t & 0 \end{bmatrix}
16:       \]
17:     end if
18:     Sample $W_i^t$ using (5)
19:     Sample new rows of $W_i^t$ from the priors in (14)
20:   end for
21: end for
22: Approximations of posterior estimates of all the unknown variables are obtained from the final particles and weights, using the procedures highlighted in (Lee et al., 2016) and discussed in the Supplementary Material.

\[w_{i-1}^t p(y_i | X_{i-1}^t, Y_{i-1})\] (Ristic et al., 2004) i.e., if the distributions $p(y_i | X_{i}, Y_{i-1})$ and $p(x_i | X_{i-1}, Y_{i-1})$ are conjugates, then closed form solutions can be obtained for $p(x_i | X_{i-1}, Y_{i})$, and hence, $p(y_i | X_{i-1}, Y_{i-1})$. However, if no such conjugacy exists, which is the case for our state-space model, the most popular choice and equally efficient solution (Van Der Merwe, 2004) is to set $q(x_i | X_{i-1}^t, Y_{i-1}) = p(x_i | X_{i-1}^t)$ (in (4)-(5)) (Wood and Griffiths, 2007; Särkkä, 2013). Considering the assumed independence in our model, i.e., $p(x_i | X_{i-1}^t, Y_{i-1}) = p(x_i | X_{i-1}^t)$ and $p(y_i | X_{i}^t, Y_{i-1}) = p(y_i | x_i^t)$, then (12) becomes:

\[\tilde{w}_i^t \propto w_{i-1}^t p(y_i | x_i^t) = w_{i-1}^t p(y_i | z_i^t, W_i^t),\] (13)

and the weights are normalized. Such implementation is commonly referred to as a bootstrap filter in the literature (Särkkä, 2013).

However, the variance of the weights increases over time, a condition referred to as degeneracy in the literature (Doucet et al., 2001). To avoid this, we perform resampling, at every time step, owing to the choice of the importance distribution (Wood and Griffiths, 2007; Särkkä, 2013), discarding the ineffective particles and multiplying the effective ones. The resampling procedure (Särkkä, 2013) is described in the Supplementary Material.

Finally, our proposed SMC algorithm for estimating the mutational profiles and the proportions of the haplotypes in the tumor samples i.e., the states and the parameters of our state-space model, is presented in Algorithm 1. The algorithm is initialized by taking particles from the prior distributions of the parameters.
We produced simulated datasets with average sequencing depth $r \in \{20, 50, 200, 100, 10000\}$ per locus. We fixed the number of haplotypes $C = 4$ and number of samples $S \in \{1, 3, 5, 10, 20\}$. For all combinations of $r$ and $S$, we generated the variants count matrix $Y$ and the total count matrix $V$ for different number of SNVs $T \in \{20, 60\}$. Specifically, we generated each entry of $V$, i.e., $v_{rs}$ from Pois($r$) and to generate each entry of $Y$, i.e., $y_{rs}$, we did the following: (i) generate each column of $W$ from $\text{Dir}(a_0, a_1, \ldots, a_4)$, where $a_0 = 0.2$ and $a_c, c \in \{1, \ldots, 4\}$ is randomly chosen from the set $\{2, 4, 5, 6, 7, 8\}$.

| $r$  | $e_{p_{0i}}$ | $e_{Z}$ | $e_{W}$ | $e_{p_{0i}}$ | $e_{Z}$ | $e_{W}$ | $e_{p_{0i}}$ | $e_{Z}$ | $e_{W}$ |
|-----|--------------|---------|---------|--------------|---------|---------|--------------|---------|---------|
| 20  | 0.0200       | 0.1033  | 0.0416  | 0.1240       | 0.1200  | 0.1001  | 0.1021       | 0.1400  | 0.0900  |
| 40  | 0.0137       | 0.0033  | 0.0316  | 0.0638       | 0.0536  | 0.0422  | 0.0490       | 0.0505  | 0.0388  |
| 50  | 0.0148       | 0.0017  | 0.0173  | 0.0745       | 0.0404  | 0.0601  | 0.0820       | 0.0500  | 0.0428  |
| 200 | 0.0122       | 0.0000  | 0.0107  | 0.0321       | 0.0325  | 0.0200  | 0.0301       | 0.0305  | 0.0211  |
| 1000| 0.0100       | 0.0000  | 0.0199  | 0.0302       | 0.0100  | 0.0324  | 0.0411       | 0.0250  | 0.0220  |
| 10000| 0.0000      | 0.0000  | 0.0020  | 0.0100       | 0.0050  | 0.0100  | 0.0110       | 0.0105  | 0.0111  |

We assume the following:

\[
\begin{align*}
\theta_{0s} & \sim \text{Gamma}(a_0, 1), \ s = 1, \ldots, S, \ p \sim \text{Beta}(a_{00}, b_{00}) \\
\theta_{cs} & \sim \text{Gamma}(a_1, 1), \ s = 1, \ldots, S, c = 1, \ldots, C,
\end{align*}
\]

such that $w_{cs} = \frac{\theta_{cs}}{\sum_{c=1}^{C} \theta_{cs}}$ and consequently, $\sum_{c=1}^{C} w_{cs} = 1$ and assume that $a_{00} \ll b_{00}$ to impose a small $p$. We report the posterior estimates of all the unknown variables using the procedure highlighted in (Lee et al., 2016), with the details discussed in the Supplementary Material.

**RESULTS AND DISCUSSION**

In this section, we demonstrate the performance of the proposed SMC algorithm using both simulated datasets and the CLL datasets obtained from three different patients (Schuh et al., 2012). In addition, we compare the estimates obtained from the proposed SMC algorithm with the MCMC-based algorithm proposed in (Lee et al., 2016) and the MAP-based algorithm proposed in (Xu et al., 2015). For the MCMC-based algorithm, the algorithm parameters are set as in (Lee et al., 2016), running a simulation over 40,000 iterations, discarding the first 15,000 iterations as burn-in. For the MAD-based algorithm, we ran 1000 random initializations for each dataset.

**Simulated data**

We produced simulated datasets with average sequencing depth $r \in \{20, 40, 50, 200, 100, 10000\}$ per locus. We fixed the number of haplotypes $C = 4$ and number of samples $S \in \{1, 3, 5, 10, 20\}$. For all combinations of $r$ and $S$, we generated the variants count matrix $Y$ and the total count matrix $V$ for different number of SNVs $T \in \{20, 60\}$. Specifically, we generated each entry of $V$, i.e., $v_{rs}$ from Pois($r$) and to generate each entry of $Y$, i.e., $y_{rs}$, we did the following: (i) generate each column of $W$ from $\text{Dir}(a_0, a_1, \ldots, a_4)$, where $a_0 = 0.2$ and $a_c, c \in \{1, \ldots, 4\}$ is randomly chosen from the set $\{2, 4, 5, 6, 7, 8\}$. 

![Figure 1. Plot of $e_{p_{0i}}$ versus sample size $S$ for SNVs $T = 60$, sequencing depth averages $r \in \{20, 50\}$, and haplotypes $C = 4$.](image1)

![Figure 2. Plot of $e_{Z}$ versus sample size $S$ for SNVs $T = 60$, sequencing depth averages $r \in \{20, 50\}$, and haplotypes $C = 4$.](image2)

![Figure 3. Plot of $e_{W}$ versus sample size $S$ for SNVs $T = 60$, sequencing depth averages $r \in \{20, 50\}$, and haplotypes $C = 4$.](image3)
Figure 4. CLL003: Plot of the estimates of the proportions of the haplotypes in each sample. Samples a, b, c, d, e are designated as 1, 2, 3, 4, 5, respectively.

Table 2. $e_Z$, $e_W$ and $e_{p_{ts}}$ computed for the proposed SMC algorithm for $C = 4$, $S = 3$ and $T \in \{1000, 2000\}$.

| Number of loci ($T$) | Number of samples ($S$) | $e_Z$   | $e_W$   | $e_{p_{ts}}$ |
|----------------------|-------------------------|---------|---------|--------------|
| 1000                 | 3                       | 0.0000  | 0.0060  | 0.0073       |
| 2000                 | 3                       | 0.0080  | 0.0048  | 0.0057       |

(ii) generate entries of $Z$ independently from Bern$(0.6)$, (iii) set $p = 0.02$, (iv) compute $p_{ts}$ using (2), and finally, (v) generate $y_{ts}$ as an independent sample from Binomial$(y_{ts}, p_{ts})$.

Next, we run the proposed SMC-based, MCMC-based and the MAP-based algorithms on the simulated $Y$ and $V$ datasets. To quantify the performance of the algorithms, we define the following metrics: haplotype error ($e_Z$), proportion error ($e_W$) and the error of the success probabilities ($e_{p_{ts}}$) as follows:

$$e_Z = \frac{1}{TC} \sum_{t=1}^{T} \sum_{c=1}^{C} |\hat{z}_{tc} - z_{tc}|,$$

$$e_W = \frac{1}{CS} \sum_{c=0}^{C} \sum_{s=1}^{S} |\hat{w}_{cs} - w_{cs}|,$$

and

$$e_{p_{ts}} = \frac{1}{TS} \sum_{t=1}^{T} \sum_{s=1}^{S} |\hat{p}_{ts} - p_{ts}|,$$

where $\hat{p}_{ts} = \hat{p}_{0s} + \sum_{c=1}^{C} \hat{z}_{tc} \hat{w}_{cs}$.

However, since this is a blind decomposition, one does not know a priori which column of $\hat{Z}$ corresponds to which column of $Z$. To resolve this, we calculate $e_Z$ with every permutation of the columns of $\hat{Z}$ and then select the permutation that results in the smallest $e_Z$. The selected permutation is then used in computing $e_W$ and $e_{p_{ts}}$. 

8/15
Figure 5. CLL077: Plot of the estimates of the proportions of the haplotypes in each sample. Samples a, b, c, d, e are designated as 1, 2, 3, 4, 5, respectively.

The results obtained from the analyses of the simulated datasets are presented in Table 1, Figures 1 - 3 and Table 1 in the Supplementary Material. Table 1 shows $e_{p_{Z1}}$, $e_Z$ and $e_W$ obtained for the datasets from $T = 60$ SNVs, $C = 4$ haplotypes and $S = 10$ samples for all the average sequencing depth $r \in \{20, 40, 50, 200, 1000, 10000\}$. Similar results are presented in the Supplementary Material with $T = 20$ SNVs and $S = 5$ samples. From the results obtained for the different number of average sequencing depth $r$ and number of SNVs, the proposed SMC algorithm yields more accurate estimates of the model parameters when compared with the other two algorithms. Specifically, the SMC-based algorithm produced lower error values $e_{p_{Z1}}$, $e_Z$ and $e_W$ in all the datasets analyzed. Moreover, we investigated the effect of sample size on the proposed SMC algorithm. As shown in Figures 1 - 3, the results show slight improvement as the number of samples increased. It can be observed that the results are less sensitive to sample size when $S > 5$. Also noticed is a slight improvement in the results when the average sequencing depth $r$ is increased.

By construction, the proposed SMC algorithm can handle datasets with any number of loci since the VAF of each loci is processed as an observation at every time step. Apart from the ability to process any number of loci, this property allows the proposed algorithm to process VAFs data from newly sequenced candidate SNVs to improve existing estimates without re-analyzing the previous datasets. To validate this, we generated datasets for 1000 and 2000 loci respectively and these datasets are analyzed with the proposed SMC algorithm with the results presented in Table 2. In fact, the proposed SMC algorithm benefits from larger number of loci because the larger the number of loci, the better the estimate of the proportions. This result is evident from the proportion errors in Table 2.

Lastly, we record the runtime ($t_r$) for the two algorithms on a 3.5 Ghz Intel 8 processors running MATLAB when analyzing some of the datasets described in Table 1 (i.e. $T = 60$, $C = 4$, $S = 10$ and $r = 20$). Observed $t_r$ was 311 seconds and 585 seconds for the proposed SMC and the MCMC-based algorithms, respectively. For the MAP-based algorithm, a single run is 5 seconds but for a set of input data, the algorithm requires different random initializations.
Figure 6. CLL006: Plot of the estimates of the proportions of the haplotypes in each sample. Samples a,b,c,d,e are designated as 1,2,3,4,5, respectively.

Real Tumor Samples: CLL Datasets

We evaluate the proposed SMC algorithm on the datasets for the B-cell chronic lymphocytic leukemia (CLL), obtained from three patients: CLL003, CLL006, and CLL077 (Schuh et al., 2012). These datasets represent the molecular changes in pre-treatment, post-treatment, and relapse samples in the three selected patients, i.e., the samples were taken temporally (see the Supplementary Material for the summary of data pre-processing). The datasets are analyzed with the proposed SMC algorithm and the two other algorithms.

CLL003

The CLL dataset obtained from patient CLL003 has 20 distinct loci, shown in the first column in Table 3, and the dataset is analyzed with the proposed SMC algorithm. In Table 3, we present the posterior point estimate of the mutational profiles of the haplotypes in each of the 5 samples, where 1 and 0 denote the variant and the reference sequence, respectively. Moreover, in Figure 4, we present a graphical representation of how the haplotypes are distributed across the samples. For instance, haplotype C2 with approximately 40 percent abundance in the first sample has reduced to approximately 3 percent after the last treatment. In the Supplementary Material, we present the table of proportions. The first row on the table and equivalently C0 in Figure 4 comprises of the proportion of the background haplotype, which accounts for experimental noise in each sample. From Table 3, we found that each sample consists of at least 2 dominant haplotypes. For instance, tumor sample a is dominated by haplotypes C2 and C6, each with a proportion of approximately 0.4. Also, we analyzed the same dataset with the other two algorithms and the results are in the Supplementary Material.

CLL077

The CLL dataset obtained from patient CLL077 has 16 distinct loci, shown in the first column in Table 4, and the dataset is analyzed with the proposed SMC algorithm. In Table 4, we present the posterior point estimate of mutational profiles of the haplotypes in each of the 5 samples. Moreover, in Figure 5, we
Table 3. **CLL003**: Estimates of the mutational profiles of haplotypes, $Z$ in the samples.

| Gene    | C1 | C2 | C3 | C4 | C5 | C6 |
|---------|----|----|----|----|----|----|
| ADAD1  | 1  | 1  | 1  | 0  | 0  | 0  |
| AMTN   | 0  | 1  | 0  | 0  | 0  | 0  |
| APBB2  | 0  | 1  | 0  | 0  | 0  | 0  |
| ASXL1  | 1  | 0  | 0  | 1  | 0  | 0  |
| ATM    | 0  | 1  | 0  | 0  | 1  | 0  |
| BPIL2  | 0  | 1  | 0  | 0  | 0  | 0  |
| CHRN2B | 1  | 0  | 0  | 1  | 0  | 0  |
| CHTF8  | 1  | 1  | 1  | 0  | 0  | 0  |
| FAT3   | 1  | 0  | 0  | 1  | 0  | 0  |
| HERC2  | 1  | 1  | 1  | 0  | 0  | 0  |
| IL11RA | 1  | 1  | 1  | 0  | 0  | 0  |
| MTUS1  | 0  | 1  | 0  | 0  | 0  | 0  |
| MUSK   | 1  | 0  | 0  | 1  | 0  | 0  |
| NPY    | 1  | 0  | 0  | 1  | 0  | 0  |
| NRG3   | 1  | 0  | 0  | 1  | 0  | 0  |
| PLEKHG5| 0  | 1  | 0  | 0  | 0  | 0  |
| SEMA3E | 1  | 0  | 0  | 1  | 0  | 0  |
| SF3B1  | 1  | 1  | 1  | 0  | 0  | 0  |
| SHROOM1| 1  | 1  | 1  | 0  | 0  | 0  |
| SPTAN1 | 0  | 1  | 0  | 0  | 0  | 0  |

The genes where the mutations are found are shown in the first column.

Present a graphical representation of how the haplotypes are distributed across the samples, with a table of proportions presented in the Supplementary Material. From our analysis results, we found that each sample consists of at least 2 dominant haplotypes. Also, we analyzed the same dataset with the other two algorithms and the results are in the Supplementary Material.

**CLL006**

Here, we analyze the CLL dataset obtained from patient **CLL006**. The dataset comprises of 11 loci as shown in Table 5 in the first column, and is analyzed with the proposed SMC algorithm. Table 5 and Figure 6 show the estimates of mutational profiles and proportions of the haplotypes, respectively. Also, in the Supplementary Material, we present the results obtained from analyzing the dataset with the two other algorithms.

As presented in the Supplementary Material, the results obtained from the other two algorithms for all the patients are similar. When the estimated haplotypes are compared with some methods that estimate the mutational profiles of tumor subclones, Phylousub proposed in (Jiao et al., 2014) and the manual method proposed in (Schuh et al., 2012), we found that some of the haplotypes, specifically, C1, C2, C3 in **CLL003**; C3, C4, C5 in **CLL077** and C1, C2, C3 in **CLL006**, carry the same set of unique mutations that are present in distinct genomes of subclones. The complete results of the clonal analyses of these methods are presented in the Supplementary Material.

Finally, the results presented so far indicate that each heterogeneous tumor sample is made up of more than two haplotypes: usually a few dominant haplotypes and other minor types. The multiple number of haplotypes in a tumor is an indication of the presence of heterogeneity in the sample.

**DISCUSSION**

Tumor samples that are obtained from cancer patients often comprise of genetically diverse populations of cells and this defines the heterogeneous nature of the samples. Most time, to explain the inherent heterogeneity in the tumor tissues, biologists obtain VAFs datasets via the NGS technology and fit the data into an appropriate model. In this paper, to analyze the observed VAFs data, we employed the feature allocation model. The model, which describes the distribution of haplotypes within the tumor samples, posits that because humans are diploids, having more than two haplotypes in the tumor sample is an...
Table 4. **CLL077**: Estimates of the mutational profiles of haplotypes, $Z$ in the samples.

| Gene   | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 |
|--------|----|----|----|----|----|----|----|----|----|
| BCL2L13| 1  | 0  | 1  | 1  | 0  | 1  | 0  | 0  | 0  |
| COL24A1| 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  |
| DAZAP1 | 0  | 0  | 0  | 1  | 0  | 0  | 0  | 1  | 0  |
| EXOC6B | 0  | 0  | 0  | 1  | 1  | 0  | 0  | 0  | 0  |
| GHDC   | 0  | 0  | 0  | 1  | 1  | 0  | 0  | 0  | 1  |
| GPR158 | 1  | 0  | 1  | 1  | 0  | 0  | 0  | 0  | 0  |
| HMCN1  | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  |
| KLHDC2 | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  |
| LRRC16A| 0  | 0  | 0  | 0  | 1  | 0  | 0  | 0  | 0  |
| MAP2K1 | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  |
| NAMPT  | 1  | 0  | 1  | 1  | 1  | 0  | 1  | 0  | 0  |
| NOD1   | 0  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  |
| OCA2   | 0  | 0  | 0  | 1  | 1  | 0  | 0  | 0  | 1  |
| PLA2G16| 0  | 0  | 0  | 1  | 1  | 0  | 1  | 0  | 0  |
| SAMHD1 | 0  | 1  | 1  | 1  | 1  | 0  | 1  | 0  | 0  |
| SLC12A1| 0  | 1  | 1  | 1  | 1  | 0  | 0  | 0  | 0  |

Table 5. **CLL006**: Estimates of the mutational profiles of haplotypes, $Z$ in the samples.

| Gene   | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 |
|--------|----|----|----|----|----|----|----|----|----|-----|
| ARHGAP29| 1  | 1  | 1  | 1  | 0  | 1  | 0  | 0  | 0  | 0   |
| EGFR   | 1  | 1  | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 0   |
| IRF4   | 1  | 0  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0   |
| KIAA0182| 1  | 1  | 1  | 0  | 1  | 0  | 1  | 0  | 0  | 0   |
| KIAA0319L| 1  | 0  | 1  | 1  | 0  | 0  | 0  | 1  | 0  | 0   |
| KLHL4  | 1  | 1  | 1  | 0  | 1  | 1  | 1  | 0  | 1  | 0   |
| MED12  | 1  | 1  | 1  | 1  | 1  | 1  | 1  | 0  | 0  | 1   |
| PILRB  | 1  | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0   |
| RBPJ   | 1  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0   |
| SIK1   | 1  | 1  | 1  | 1  | 0  | 0  | 0  | 0  | 0  | 0   |
| U2AF1  | 1  | 1  | 1  | 0  | 1  | 0  | 0  | 0  | 0  | 0   |

The genes where the mutations are found are shown in the first column.

Evidence of heterogeneity within the sample. According to this model, haplotypes in the tumor samples are the features and SNVs are the experimental units that select the features. So given the observed VAFs of the SNVs, estimating the unknown latent features and the proportions in the samples completely described the inherent heterogeneity in the data.

To estimate the unknown variables in the model, we reformulate the latent feature model into state-space model and present an efficient SMC algorithm. The SMC algorithm takes advantage of the sequential construction of the latent binary matrix, with an unknown number of columns, using the IBP, and treating other variables in the latent feature model as the parameters of our newly formed state-space model. This way, we are able to analyze the VAFs of a single SNV at each iteration. We evaluated our proposed SMC algorithm on simulated datasets, specifically, by varying the average sequencing depth ($r$), the number of tumor samples ($S$) and the number of SNVs ($T$). Also, we analyzed real datasets, i.e., the CLL datasets obtained from (Schuh et al., 2012) for 3 patients. The proposed SMC algorithm produced satisfying results on all categories of datasets analyzed.

Further, we compared the estimates obtained from the proposed SMC-based, MCMC-based and MAP-based algorithms. In terms of the accuracy of estimates, the proposed SMC algorithm yields an improved performance over the two competing algorithms. In the proposed SMC-based algorithm, it is possible to easily incorporate datasets from newly sequenced SNVs (when available) so as to refine...
the existing estimates. However, in the competing algorithms, to incorporate the new datasets, the entire datasets (old and new) need be analyzed.

In our experiments, we set $N = 500$ particles for all the simulated datasets and for the tumor datasets, we set $N = 1000$. Also, we run the SMC algorithm 5 times for the simulated data and 10 times for the CLL datasets. Multiple runs allow the VAFs of each SNV to be visited more than once, and we noticed that this, in a way, improves the results of the SMC algorithm.

Finally, we have demonstrated the efficacy of the SMC algorithm, an algorithm that can effectively handle any type of probability distribution, nonlinearity and non-stationarity, particularly in analyzing VAFs of SNVs from tumor samples.

ACKNOWLEDGMENTS

REFERENCES

Bashashati, A., Ha, G., Tone, A., Ding, J., Prentice, L. M., Roth, A., Rosner, J., Shumansky, K., Kalloger, S., Senz, J., Yang, W., McConney, M., Melnyk, N., Anglesio, M., Luk, M. T., Tse, K., Zeng, T., Moore, R., Zhao, Y., Marra, M. A., Gilks, B., Yip, S., Huntsman, D. G., McAlpine, J. N., and Shah, S. P. (2013). Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. The Journal of pathology, 231(1):21–34.

Ding, L., Ellis, M. J., Li, S., Larson, D. E., Chen, K., Wallis, J. W., Harris, C. C., McLellan, M. D., Fulton, R. S., Fulton, L. M., Abbott, R. M., Hoog, J. D., Dooling, D. J., Koboldt, D. C., Schmidt, H., Kalicki, J., Zhang, Q., Chen, L., Lin, L., Wend, M. C., McMichael, J. F., Magrini, V. J., Cook, L., McGrath, S. D., Vickery, T. L., Appelbaum, E., DeSchryver, K., Davies, S., Gintonti, T., Lin, L., Crowder, R., Tao, Y., Snider, J. E., Smith, S. M., Dukes, A. F., Sanderson, G. E., Pohl, C. S., Delehaunty, K. D., Fronick, C. C., Pape, K. A., Reed, J. S., Robinson, J. S., Hodges, J. S., Schierding, W., Dees, N. D., Shen, D., Locke, D. P., Wiechert, M. E., Eldred, J. M., Peck, J. B., Oberkfell, B. J., Loloife, J. D., Du, F., Hawkins, A. E., O’Laughlin, M. D., Bernard, K. E., Cunningham, M., Elliott, G., Mason, M. D., Thompson Jr, D. M., Ivanovich, J. L., Goodfellow, P. J., Perou, C. M., Weinstock, G. M., Aft, R., Watson, M., Ley, T., Wilson, R. K., and Mardis, E. R. (2010). Genome remodeling in a basal-like breast cancer metastasis and xenograft. Nature, 464(7291):999.

Doshi-Velez, F. (2009). The indian buffet process: Scalable inference and extensions. Master’s thesis, The University of Cambridge.

Doucet, A., De Freitas, N., and Gordon, N. (2001). Sequential monte carlo methods in practice springer. New York.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential monte carlo sampling methods for bayesian filtering. Statistics and computing, 10(3):197–208.

Garraway, L. A. and Lander, E. S. (2013). Lessons from the cancer genome. Cell, 153(1):17–37.

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. New England journal of medicine, 366(10):883–892.

Ghahramani, Z. and Griffiths, T. L. (2006). Infinite latent feature models and the indian buffet process. In Advances in neural information processing systems, pages 475–482.

Griffiths, T. L. and Ghahramani, Z. (2011). The indian buffet process: An introduction and review. Journal of Machine Learning Research, 12(Apr):1185–1224.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. cell, 100(1):57–70.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. cell, 144(5):646–674.

Heppner, G. H. (1984). Tumor heterogeneity. Cancer research, 44(6):2259–2265.

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., Wu, H., Ye, X., Ye, C., Wu, R., Jian, M., Chen, Y., Xie, W., Zhang, R., Chen, L., Liu, X., Yao, X., Zheng, H., Yu, C., Li, Q., Gong, Z., Mao, M., Yang, X., Yang, L., Yang, H., Li, Y., Zhang, X., and Wang, J. (2012). Single-cell exome sequencing and monoclonal evolution of a jak2-negative myeloproliferative neoplasm. Cell, 148(5):873–885.
Shah, S. P., Morin, R. D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Schuh, A., Becq, J., Humphray, S., Alexa, A., Burns, A., Clifford, R., Feller, S. M., Grocock, R., Henderson, S., Khroutkova, I., Kingsbury, Z., Luo, S., McBride, D., Murray, L., Menju, T., Timbs, A., Ross, M., Taylor, J., and Bentley, D. (2012). Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood*, 120(20):4191–4196.

Shah, S. P., Morin, R. D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K.,...
Guliany, R., Senz, J., Steidl, C., Holt, R. A., Jones, S., Sun, M., Leung, G., Moore, R., Severson, T., Taylor, G. A., Teschendorff, A. E., Tse, K., Turashvili, G., Varhol, R., Warren, R. L., Watson, P., Zhao, Y., Caldas, C., Huntsman, D., Hirst, M., Marra, M. A., and Aparicio, S. (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature*, 461(7265):809.

Su, X., Zhang, L., Zhang, J., Meric-Bernstam, F., and Weinstein, J. N. (2012). Purityest: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics*, 28(17):2265–2266.

Van Der Merwe, R. (2004). Sigma-point kalman filters for probabilistic inference in dynamic state-space models.

Visvader, J. E. (2011). Cells of origin in cancer. *Nature*, 469(7330):314.

Wood, F. and Griffiths, T. L. (2007). Particle filtering for nonparametric bayesian matrix factorization. In *Advances in Neural Information Processing Systems*, pages 1513–1520.

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., He, W., Zeng, L., Xing, M., Wu, R., Jiang, H., Liu, X., Cao, D., Guo, G., Hu, X., Gui, Y., Li, Z., Xie, W., Sun, X., Shi, M., Cai, Z., Wang, B., Zhong, Meiming, L. J., Lu, Z., Gu, N., Zhang, X., Goodman, L., Bolund, L., Wang, J., Yang, H., Kristiansen, K., Dean, M., Li, Y., and Wang, J. (2012). Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, 148(5):886–895.

Xu, Y., Müller, P., Yuan, Y., Gulukota, K., and Ji, Y. (2015). Mad bayes for tumor heterogeneity? feature allocation with exponential family sampling. *Journal of the American Statistical Association*, 110(510):503–514.

Zare, H., Wang, J., Hu, A., Weber, K., Smith, J., Nickerson, D., Song, C., Witten, D., Blau, C. A., and Noble, W. S. (2014). Inferring clonal composition from multiple sections of a breast cancer. *PLoS computational biology*, 10(7):e1003703.

Zhang, X. (2008). A very gentle note on the construction of dirichlet process. *The Australian National University, Canberra.*
Figure 1 (on next page)

Plot of $e_p$ts versus sample size $S$ for SNVs $T = 60$, sequencing depth averages $r$ in $\{20, 50\}$, and haplotypes $C = 4$. 
Figure 2 (on next page)

Plot of $e_z$ versus sample size $S$ for SNVs $T = 60$, sequencing depth averages $r$ in $\{20, 50\}$, and haplotypes $C = 4$. 
The graph shows the relationship between samples (S) and $e^z$ for two different values of $r$: 20 and 50.

- The blue line with square markers represents $r = 20$.
- The red dashed line with plus markers represents $r = 50$.

As the number of samples increases, the value of $e^z$ decreases for both values of $r$.

The x-axis represents the number of samples, and the y-axis represents $e^z$. The graph illustrates how the effect of $e^z$ diminishes with an increasing number of samples.
Figure 3 (on next page)

Plot of $e_w$ versus sample size $S$ for SNVs $T = 60$, sequencing depth averages $r$ in \{20, 50\}, and haplotypes $C = 4$. 
Figure 4 (on next page)

CLL003: Plot of the estimates of the proportions of the haplotypes in each sample. Samples a,b,c,d,e are designated as 1,2,3,4,5, respectively.
**Figure 5** (on next page)

CLL077: Plot of the estimates of the proportions of the haplotypes in each sample. Samples a, b, c, d, e are designated as 1, 2, 3, 4, 5, respectively.
Figure 6 (on next page)

CLL006: Plot of the estimates of the proportions of the haplotypes in each sample. Samples a,b,c,d,e are designated as 1,2,3,4,5, respectively.
