OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Inferring Clonal Composition from Multiple Sections of a Breast Cancer

Habil Zare[1☯], Junfeng Wang[2☯], Alex Hu[1], Kris Weber[1], Josh Smith[1], Debbie Nickerson[1], ChaoZhong Song[2], Daniela Witten[3]*, C. Anthony Blau[2]*, William Stafford Noble[1,4]*

1 Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America, **2** Division of Hematology, Department of Medicine, University of Washington, Seattle, Washington, United States of America, **3** Department of Biostatistics, University of Washington, Seattle, Washington, United States of America, **4** Department of Computer Science and Engineering, University of Washington, Seattle, Washington, United States of America

Abstract

Cancers arise from successive rounds of mutation and selection, generating clonal populations that vary in size, mutational content and drug responsiveness. Ascertaining the clonal composition of a tumor is therefore important both for prognosis and therapy. Mutation counts and frequencies resulting from next-generation sequencing (NGS) potentially reflect a tumor's clonal composition; however, deconvolving NGS data to infer a tumor's clonal structure presents a major challenge. We propose a generative model for NGS data derived from multiple subsections of a single tumor, and we describe an expectation-maximization procedure for estimating the clonal genotypes and relative frequencies using this model. We demonstrate, via simulation, the validity of the approach, and then use our algorithm to assess the clonal composition of a primary breast cancer and associated metastatic lymph node. After dividing the tumor into subsections, we perform exome sequencing for each subsection to assess mutational content, followed by deep sequencing to precisely count normal and variant alleles within each subsection. By quantifying the frequencies of 17 somatic variants, we demonstrate that our algorithm predicts clonal relationships that are both phylogenetically and spatially plausible. Applying this method to larger numbers of tumors should cast light on the clonal evolution of cancers in space and time.

Citation: Zare H, Wang J, Hu A, Weber K, Smith J, et al. (2014) Inferring Clonal Composition from Multiple Sections of a Breast Cancer. PLoS Comput Biol 10(7): e1003703. doi:10.1371/journal.pcbi.1003703

Editor: Amos Tanay, Weizmann Institute of Science, Israel

Received November 8, 2013; **Accepted** May 20, 2014; **Published** July 10, 2014

Funding: This work was supported by National Institutes of Health awards 1R01CA135357-01A1 (CAB) and 1DP5OD009145 (DW). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: dwitten@uw.edu (DW); tblau@uw.edu (CAB); william-noble@uw.edu (WSN)

☯ These authors contributed equally to this work.

Introduction

Many clones exist within each cancer, and selective pressure imposed by environmental factors, most notably treatments directed at tumor eradication, favors the emergence of clones that grow increasingly resistant to successive rounds of therapy. Incorporating this intra-tumor heterogeneity into strategies for planning, monitoring, and revising cancer treatment could improve outcomes for oncologists and their patients. Therefore, methods for estimating the number, size and mutational content of clones within a patient's tumor are being explored.

New approaches are being developed to assess the clonal content of a given tumor. Methods based on the interrogation of individual cells have relied on the use of fluorescent markers [1,2] or single cell sequencing [3–6]. Whereas fluorescence-based approaches are inevitably limited by the relatively small number of features they can accommodate, single cell sequencing brings the highest possible resolution to characterizing an individual patient's tumor. Nonetheless, single cell sequencing also faces obstacles to its widespread implementation. Evaluating sufficiently large numbers of single cells to obtain statistical power can be prohibitive, for technical or financial reasons. Additionally, it is often difficult to ascertain the identity of the cells being sequenced, and details regarding the spatial positioning of cells relative to each other and to other cells in the tumor are lost when the single cells are obtained. These disadvantages pose significant challenges to the widespread adoption of single cell sequencing as a means for assessing tumor heterogeneity.

Complementing single cell approaches are efforts to deconvolve clonal subpopulations based on the frequencies of mutated alleles within one or more bulk tumor specimens. Shah et al. [7], who sequenced a breast cancer at the time of diagnosis and nine years later, after metastasis, pointed out that allele frequencies of the mutations shared between the two samples could be used to segregate primary mutations into those that occur in a dominant versus subdominant clone. This insight is the basis for a variety of approaches that apply clustering algorithms to mutation allele frequencies, including kernel density estimation [8] and Dirichlet process modeling applied either to the allele frequencies [9] or to a combination of allele frequency, loss-of-heterozygosity status and copy number [10–13].

Clearly, statistical power to infer variants and, ultimately, clonal composition, is increased if multiple samples are available for analysis. Accordingly, various studies have examined the progression of cancer within one or more patients over time. Sets of variants that exhibit similar allele frequencies within a single sample are suggestive of a clonal population. Hence, clustering methods to identify groups of mutations associated with a single

Author Summary

Cancers arise from a series of mutations that occur over time. As a result, as a tumor grows each cell inherits a distinctive *genotype*, defined by the set of all somatic mutations that distinguish the tumor cell from normal cells. Acertaining these genotype patterns, and identifying which ones are associated with the growth of the cancer and its ability to metastasize, can potentially give clinicians insights into how to treat the cancer. In this work, we describe a method for inferring the predominant genotypes within a single tumor. The method requires that a tumor be sectioned and that each section be subjected to a high-throughput sequencing procedure. The resulting mutations and their associated frequencies within each tumor section are then used as input to a probabilistic model that infers the underlying genotypes and their relative frequencies within the tumor. We use simulated data to demonstrate the validity of the approach, and then we apply our algorithm to data from a primary breast cancer and associated metastatic lymph node. We demonstrate that our algorithm predicts genotypes that are consistent with an evolutionary model and with the physical topology of the tumor itself. Applying this method to larger numbers of tumors should cast light on the evolution of cancers in space and time.

clone have been applied. For example, kernel density estimation has been applied to allele frequencies from tumor-relapse pairs from eight acute myeloid leukemia (AML) patients [14] and from seven secondary AML patients [15].

An orthogonal approach taken by Newberger et al. [16] employs triplet samples of neoplasia, matched normal and carcinoma from six patients to infer lineages of various genetic events. They characterize each locus in terms of a binary vector representing the presence of the mutation across the various samples and then group the loci into classes on the basis of these vectors. After filtering low frequency classes, the classes are used to manually construct a phylogenetic tree. The focus of the study is to identify the shared characteristics of the evolutionary process across six patients with breast cancer.

In the current study, we adopt an alternative approach to identify clonal structure. Rather than measuring allele frequencies in multiple samples from the same patient over time, we physically subdivide a single breast cancer specimen and measure allele frequencies within each subsection (Figure 1). We are aware of two previous studies that have adopted such an approach. Yachida et al. [17] analyzed seven metastatic pancreatic cancers, sequencing from multiple samples per patient. Clones are initially defined relative to sample types (peritoneal, liver and lung metastases). Subsequently, the tumors from two patients are resected and a clonal phylogeny is inferred manually. More recently, Gerlinger et al. [18] carried out exome sequencing followed by targeted deep sequencing on samples from four patients with renal carcinoma. Each primary tumor was divided into 9 regions, and a phylogeny was manually constructed by assuming that higher alternate allele frequencies correspond to earlier mutations. In neither of these studies was an algorithm proposed to automatically infer from such data both the clonal genotypes and the relative frequencies of the clones within each subsection.

The method proposed here bears some similarity to the recently proposed Tree Approach to Clonality (TrAp) method [19]. The TrAp algorithm aims to identify the number, relative frequencies and genotypes of clones within a tumor using a formalism somewhat similar to ours, based on matrix decomposition. However, rather than analyzing data from multiple sections, the authors use as input a single set of variant allele frequencies and then constrain the resulting optimization problem by introducing a series of four assumptions about cancer evolution. It is not clear whether the method can easily generalize to analysis of data from multiple sections or multiple time points.

Here we describe a generative binomial model that incorporates information from multiple sections from a single tumor at a single time point to infer the frequencies and genotypes for a specified number of clones. An implementation of our algorithm is available through Bioconductor as an R package called Clomial (http://www.bioconductor.org/packages/release/bioc/html/Clomial.html). We use Clomial version 1.1.7 to apply this approach to a breast cancer specimen and demonstrate that the results from our model predict relationships that are phylogenetically and spatially plausible.

Results

1 Inferring the clonal architecture of a tumor

We assume that a tumor is comprised of multiple populations of cells (“clones”), each with a unique genotype, and that these populations are heterogeneously distributed within the tumor itself. We collect, from several physical subsections of the tumor, shotgun sequencing reads. We also collect sequencing data from a non-tumor subsection from the same patient. Using the called genotypes from the normal subsection, and restricting ourselves to positions that are homozygous in the normal subsection, each read from a tumor subsection exhibits either a normal allele or a variant allele at each location. We exclude positions that exhibit homozygous normal alleles in all of the tumor subsections. Our goal is to infer, from the remaining N mutated positions, the genotype of each clonal population and their relative frequencies within each physical subsection of the tumor.

Formally, the problem can be stated as follows. Note that we use bold face letters for random variables, and that A_i and A^j respectively denote the i^{th} row and the j^{th} column of matrix A. We are given two primary input matrices $R_{N\times M}$ and $X_{N\times M}$, where N is the number of mutated loci, M is the number of subsections (of which one is normal and $M-1$ are tumor), $R_{i,j}$ is the total number of reads (i.e., the coverage) at locus i in subsection j, and $X_{i,j}$ is the number of cancerous reads (those supporting the mutation) at locus i in subsection j. We assume, without loss of generality, that the first of the M subsections corresponds to normal tissue, and that the remaining $M-1$ subsections are from the tumor. In addition, we consider C, the number of distinct clones in the tumor, as a hyperparameter, and train a model based on a given value of C. We assume that the first clone corresponds to the normal cell population and the tumor is composed of $C-1$ tumor clones. Later, we will discuss whether C can be estimated from the data. Our task is to infer two matrices: a *clone frequency matrix* $P_{C\times M}$ in which $P_{c,j}$ is the proportion of cells of clone c in subsection j, and a *genotype matrix* $Z_{N\times C}$ in which $Z_{i,c}=1$ if clone c has the variant allele at locus i, and $Z_{i,c}=0$ otherwise. The first column of Z contains all zeroes because it represents the “normal clone.” By definition, each column of P sums to 1. Also, by construction, the first column of X corresponds to the normal subsection and hence consists almost entirely of zeroes, although small non-zero counts may be possible due to contamination from tumor or due to sequencing error. If the first column of X consisted entirely of zeroes, then we would expect the first column of P to be of the form $(1,0,\ldots,0)$, but in order to allow for the possibility that the allegedly normal subsection can have slight tumor contamination, we infer the first column of P (as well as the other $M-1$ columns).

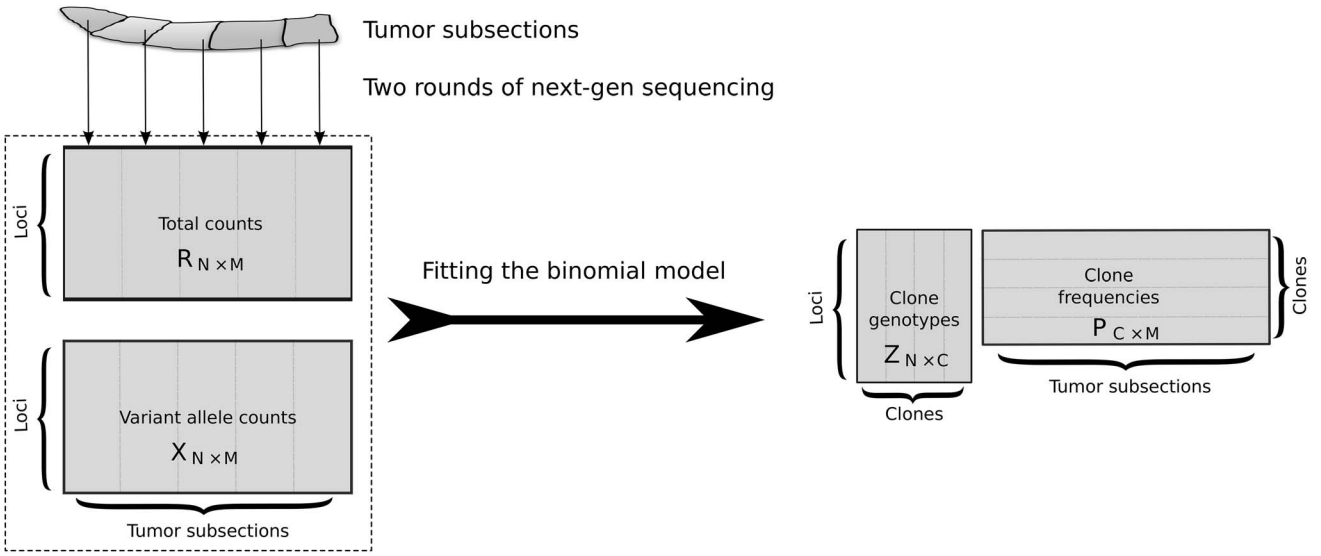

Figure 1. Inference of tumor clonal content. A collection of subsections of a tumor are subjected to next-generation sequencing to measure, across a common set of genomic loci, counts of two alleles—the *normal* allele that was observed in a matched normal sample at that locus, and a *variant allele*. The resulting counts matrices are provided as input to an inference procedure that estimates the clonal genotypes and frequencies.
doi:10.1371/journal.pcbi.1003703.g001

We propose to solve this problem using a generative model whose parameters are learned via expectation-maximization (EM) [20]. Accordingly, we define a matrix of hidden variables $\mathbf{Z}_{N\times C}$ representing the unknown genotypes of the clones; for instance, if $\mathbf{Z}_{i,c}=1$, then the c^{th} clone has a tumor allele at the i^{th} locus. We assume that each $\mathbf{Z}_{i,c}$ follows an independent Bernoulli distribution with parameter $\mu_{i,c}$, i.e.,

$$\mathbf{Z}_{i,c}\sim \text{Bern}(\mu_{i,c}). \tag{1}$$

We also assume that if a mutation is present in a particular clone, then at that locus the clone is heterozygous with copy number equal to 1. Therefore, for subsection j, if clone c has a mutation at locus i ($\mathbf{Z}_{i,c}=1$), then its contribution to the observed count of cancer alleles is by $\frac{1}{2}P_{c,j}$, half of its proportion in the subsection. Conversely, if a clone does not have a mutation at i ($\mathbf{Z}_{i,c}=0$), then it does not contribute to the count of variant alleles. By summing up the contributions of all clones, we obtain the total probability that an observed read corresponds to a variant allele rather than a normal allele. Therefore, the probability that a read contains the variant allele at locus i in subsection j is given by

$$\pi_{i,j}=\frac{1}{2}\mathbf{Z}_i\cdot P^j, \tag{2}$$

where $\mathbf{Z}_i$ is the i^{th} row of $\mathbf{Z}$, and P^j is the j^{th} column of P. Finally, we introduce a matrix $\mathbf{X}_{N\times M}$ of random variables representing the observed data, where $\mathbf{X}_{i,j}$ is the number of reads exhibiting the variant allele at locus i in subsection j. This matrix encodes our primary assumption about the distribution of the data: for each i and j, we observe an independent sample of $\mathbf{X}_{i,j}$ that has a binomial distribution with two parameters $R_{i,j}$ and $\pi_{i,j}$, i.e.,

$$\mathbf{X}_{i,j}\sim \text{Bin}(R_{i,j},\pi_{i,j}). \tag{3}$$

The first parameter of this distribution $R_{i,j}$ is the (known) total number of reads at locus i in subsection j. The second parameter, $\pi_{i,j}$, is the probability of observing a variant allele; it will be inferred by EM.

Given the joint distribution $\Pr(\mathbf{X},\mathbf{Z}|\theta)$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters $\theta=(P,\mu)$, our goal is to maximize the likelihood function $\Pr(\mathbf{X}|\theta)$. We do so using EM, exploiting three assumptions: (1) that each subsection contains non-zero normal contamination, i.e., $P_{1,j}>0$ for all j, (2) independence of the M subsections from each other, and (3) independence of mutations from each other. The first assumption is based on the widely accepted difficulty associated with obtaining perfectly pure samples of tumor cells [21,22]. The two independence assumptions essentially state that each locus and each sample is informative. These assumptions are unavoidable: in the presence of very high dependence, only very limited information about the underlying clonal composition of the tumor would be provided by the loci and samples. Furthermore, it is worth noting that these independence assumptions are made *conditional* on the parameters in the model: that is, the elements of $\mathbf{X}$ are independent conditional on $\mathbf{Z}$ and P. In other words, if we knew the true underlying parameters for the model (that is, the true genotypes for the clones, and the true proportion of each clone present in each sample), then the actual number of "tumor" reads that we would observe for each locus-sample pair would be independent.

While the formulation of our inference problem shows some similarity to well-studied matrix factorization problems [23–25], such techniques cannot be directly applied here. Unlike most matrix factorization techniques, which assume a normal distribution, our observations are binomially distributed. Moreover, the elements of the latent matrix $\mathbf{Z}$ are binary, and each column of P must sum to 1. These constraints required us to develop a customized inference algorithm.

2 Log likelihood

To frame the EM optimization, we consider the following complete-data log likelihood function of the model:

$$\mathcal{L} = \log \Pr(\mathbf{X}, \mathbf{Z} | \theta), \tag{4}$$

which can be computed as follows (for details see Note S4 in Text S1):

$$\mathcal{L} = \sum_{i,j} \left(\log \binom{R_{i,j}}{\mathbf{X}_{i,j}} + \mathbf{X}_{i,j} \log(\pi_{i,j}) + (R_{i,j} - \mathbf{X}_{i,j}) \log(1 - \pi_{i,j}) \right) + \sum_{i,c} \left(\mathbf{Z}_{i,c} \log(\mu_{i,c}) + (1 - \mathbf{Z}_{i,c}) \log(1 - \mu_{i,c}) \right), \tag{5}$$

where $\pi_{i,j} = \frac{1}{2} \mathbf{Z}_i \cdot P^j$.

3 Expectation maximization (EM) algorithm

Our goal is to find the parameters $\theta = (P, \mu)$ which maximize the likelihood. Because our model involves the hidden variable $\mathbf{Z}$, we cannot directly maximize the $\mathcal{L}$ given in Equation 5 with respect to θ. Instead, we use the EM algorithm to fit the model to the data [26]. EM is an iterative algorithm with two steps—E (for expectation) and M (for maximization)—in each iteration. In the E step, we use the current estimates of the parameters, θ^{old}, to compute the conditional expectation of $\mathcal{L}$. In the M step, we find the new parameters θ^{new} that maximize the conditional expectation.

Overview. In this section, we present an overview of the EM algorithm, followed by the specific details of the E and M steps for our application.

1. Randomly initialize the parameters θ^{old}.
2. Repeat the following until a convergence criterion is satisfied (such as insignificant improvement in the log likelihood; see Equation 12).

(a) **E Step.** Evaluate the posterior $\Pr(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ using the current parameter values. Because each locus is independent, we will compute $\Pr(\mathbf{Z}_i|\mathbf{X}_i, \theta^{\text{old}})$ for $1 \le i \le n$. This can be done by Bayes' theorem,

$$\Pr(\mathbf{Z}_i|\mathbf{X}_i, \theta^{\text{old}}) = \frac{\Pr(\mathbf{X}_i|\mathbf{Z}_i, \theta^{\text{old}}) \Pr(\mathbf{Z}_i|\theta^{\text{old}})}{\sum_{z \in \{0,1\}^C} \Pr(\mathbf{X}_i|\mathbf{Z}_i = z, \theta^{\text{old}}) \Pr(\mathbf{Z}_i = z|\theta^{\text{old}})}. \tag{6}$$

(b) **M Step.** Evaluate θ^{new} from

$$\theta^{\text{new}} \leftarrow \operatorname*{argmax}_{\theta} Q(\theta|\theta^{\text{old}})$$

where $Q(\theta|\theta^{\text{old}})$ is the following expected log likelihood with respect to $\mathbf{Z}$ conditioned on $\mathbf{X}$ and θ^{old}:

$$\begin{aligned} Q(\theta|\theta^{\text{old}}) &= \mathbb{E}_{\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}} [\log \Pr(\mathbf{X}, \mathbf{Z}|\theta)] \\ &= \sum_{\mathbf{Z}} \Pr(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \log \Pr(\mathbf{X}, \mathbf{Z}|\theta). \end{aligned} \tag{7}$$

(c) Update the parameters by

$$\theta^{\text{old}} \leftarrow \theta^{\text{new}}.$$

Computation for the E step. To compute the posterior $\Pr(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$, we need to compute $\Pr(\mathbf{Z}_i|\theta^{\text{old}})$ and $\Pr(\mathbf{X}_i|\mathbf{Z}_i, \theta^{\text{old}})$ for the i^{th} locus (see Equation 6). The latter is equal to the product of binomial probabilities because the samples are assumed to be independent. Using Equations 2 and 3, we have

$$\begin{aligned} \Pr(\mathbf{X}_i|\mathbf{Z}_i, \theta^{\text{old}}) &= \prod_{j=1}^{M} \Pr(\mathbf{X}_{i,j}|\mathbf{Z}_i, \theta^{\text{old}}) \\ &= \prod_{j=1}^{M} \binom{R_{i,j}}{\mathbf{X}_{i,j}} \left(\frac{1}{2} \mathbf{Z}_i \cdot P^{j\ \text{old}} \right)^{\mathbf{X}_{i,j}} \left(1 - \frac{1}{2} \mathbf{Z}_i \cdot P^{j\ \text{old}} \right)^{R_{i,j} - \mathbf{X}_{i,j}}. \end{aligned} \tag{8}$$

Also, $\Pr(\mathbf{Z}_i|\theta^{\text{old}})$ is the product of Bernoulli probabilities. From Equation 1, we have that

$$\Pr(\mathbf{Z}_i|\theta^{\text{old}}) = \prod_{c=1}^{C} (\mu_{i,c}^{\text{old}})^{\mathbf{Z}_{i,c}} (1 - \mu_{i,c}^{\text{old}})^{1 - \mathbf{Z}_{i,c}}.$$

Computation for the M step. To get θ^{new}, we maximize $Q(\theta|\theta^{\text{old}})$ defined in Equation 7. First, we split $Q(\theta|\theta^{\text{old}})$ into two terms such that one term depends only on P, and the other term depends only on μ. This simplifies the process of finding the optimal parameters.

$$\begin{aligned} Q(\theta|\theta^{\text{old}}) &= \mathbb{E}_{\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}} [\log \Pr(\mathbf{X}, \mathbf{Z}|\theta)] \\ &= \mathbb{E}_{\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}} [\log(\Pr(\mathbf{X}|\mathbf{Z}, \theta) \Pr(\mathbf{Z}|\theta))] \\ &= \underbrace{\mathbb{E}_{\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}} [\log \Pr(\mathbf{X}|\mathbf{Z}, P)]}_{\Phi(P)} \\ &\quad + \underbrace{\mathbb{E}_{\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}} [\log \Pr(\mathbf{Z}|\mu)]}_{\Psi(\mu)}, \end{aligned} \tag{9}$$

where we let $\Phi(P) := \mathbb{E}_{\mathrm{Z}|\mathrm{X}, \theta^{\text{old}}} [\log \Pr(\mathrm{X}|\mathrm{Z}, P)]$ and $\Psi(\mu) := \mathbb{E}_{\mathrm{Z}|\mathrm{X}, \theta^{\text{old}}} [\log \Pr(\mathrm{Z}|\mu)]$ for simplicity of notation. Similar to Equation S7 in Note S4 in Text S1, we have used the fact that conditional on $\mathbf{Z}$ and P, $\mathbf{X}$ is independent of μ, as well as the fact that conditional on μ, $\mathbf{Z}$ is independent of P.

Computing μ^{new} Now that we have separated $Q(\theta|\theta^{\text{old}})$ into two terms, we can first update P by only maximizing $\Phi(P)$. We solve the following constrained optimization problem to get P^{new} (for details see Note S5 in Text S1):

$$\begin{cases} P^{\text{new}} := \operatorname*{argmax}_{P} \Phi(P) \\ \text{such that } \forall j, c : 0 \le P_{c,j} \text{ and } \forall j : \sum_{c=1}^{C} P_{c,j} = 1. \end{cases} \tag{10}$$

We solve our simplified optimization problem using a quasi-Newton method called BFGS-B [27,28]. The original BFGS algorithm uses the gradient to approximate the Hessian matrix of second derivatives; therefore, the algorithm is very efficient when the gradient is available [29,30]. BFGS-B is a variant that can handle simple box constraints. We compute the first derivative of Φ with respect to each entry of V by the chain rule, and provide it to BFGS-B for faster convergence (Note S2 in Text S1).

Computing μ^{new} Recall that $\Psi(\mu) = \mathbb{E}_{\mathbf{Z}|\mathbf{X}, \theta^{old}} [\log \Pr(\mathbf{Z}|\mu)]$ is the only part of the expected log likelihood which is a function of μ (see Equation 9). Therefore, we can compute μ^{new} by maximizing $\Psi(\mu)$. Because we are assuming that conditional on μ, the elements of $\mathbf{Z}$ are independent, we just need to maximize each

term in the following sum:

$$\Psi(\mu) = \mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta^{\text{old}}}[\log \Pr(\mathbf{Z}|\mu)] = \sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{E}_{\mathbf{Z}|\mathbf{X},\theta^{\text{old}}}[\log \Pr(\mathbf{Z}_{i,c}|\mu_{i,c})].$$

The first column of $\mathbf{Z}$ corresponds to the normal cells, $\mathbf{Z}_{i,1}=0$, hence $\mu_{i,1}=0$ for all i. For $1<c\leq C$, we need to maximize, with respect to $\mu_{i,c}$,

$$\begin{aligned}
&\mathbb{E}_{Z|\mathbf{X},\theta^{\text{old}}}[\log \Pr(\mathbf{Z}_{i,c}|\mu_{i,c})] \\
&= \Pr(\mathbf{Z}_{i,c}=1|\mathbf{X},\theta^{\text{old}})\log \Pr(\mathbf{Z}_{i,c}=1|\mu_{i,c}) \\
&+ \Pr(\mathbf{Z}_{i,c}=0|\mathbf{X},\theta^{\text{old}})\log \Pr(\mathbf{Z}_{i,c}=0|\mu_{i,c}) \\
&= \Pr(\mathbf{Z}_{i,c}=1|\mathbf{X},\theta^{\text{old}})\log(\mu_{i,c}) \\
&+ (1-\Pr(\mathbf{Z}_{i,c}=1|\mathbf{X},\theta^{\text{old}}))\log(1-\mu_{i,c}).
\end{aligned} \tag{11}$$

By single-variable calculus, the value of $\mu_{i,c}$ that maximizes (11) is $\Pr(\mathbf{Z}_{i,c}=1|\mathbf{X},\theta^{\text{old}})$. For the c^{th} clone, the probability $\Pr(\mathbf{Z}_{i,c}=1|\mathbf{X},\theta^{\text{old}})$ can be computed by marginalizing $\mathbf{Z}_i$ over all other clones:

$$\Pr(\mathbf{Z}_{i,c}=1|\mathbf{X},\theta^{\text{old}}) = \sum_{\mathbf{z}_i\in\{\{0,1\}^C|\mathbf{Z}_{i,c}=1\}} \Pr(\mathbf{Z}_i|\mathbf{X},\theta^{\text{old}}).$$

Because the posteriors $\Pr(\mathbf{Z}_i|\mathbf{X},\theta^{\text{old}})$ are easy to compute by Bayes' rule (Equation 6), μ^{new} can be updated as follows:

$$\mu_{i,c}^{\text{new}} = \sum_{\mathbf{z}_i\in\{\{0,1\}^C|\mathbf{Z}_{i,c}=1\}} \Pr(\mathbf{Z}_i|\mathbf{X},\theta^{\text{old}}), 1\leq i\leq N, 1<c\leq C.$$

In principle, for each solution, the genotype matrix Z can be obtained by rounding the inferred μ. However, in practice, the inferred values in μ were always exactly 0 or 1 (with observed differences $<10^{-20}$).

Initialization and convergence. We initialize elements of $P_{C\times M}$ with values independently sampled from a Uniform [0,1] distribution. Then we standardize each column such that the sum of the proportions of each clone in a subsection is 1. Similarly, we randomly initialize the matrix $\mu_{N\times C}$ with values independently sampled from a Uniform [0,1] distribution. In practice, we run EM to convergence from multiple random initializations for μ and P, and we choose the run that results in the highest likelihood.

The convergence criterion is based on the change in the expectation of the complete-data log likelihood. Specifically, we stop the EM iterations if:

$$\frac{\mathbb{E}_{\mathbf{Z}|\theta^{\text{new}}}[\mathcal{L}^{\text{new}}] - \mathbb{E}_{\mathbf{Z}|\theta^{\text{old}}}[\mathcal{L}^{\text{old}}]}{\mathbb{E}_{\mathbf{Z}|\theta^{\text{old}}}[\mathcal{L}^{\text{old}}]} < \alpha \tag{12}$$

where α is a small positive number. We set $\alpha=10^{-3}$ in our experiments. Using Equation 5, we can compute $\mathbb{E}_{\mathbf{Z}|\theta}[\mathcal{L}]$ in each iteration as follows:

$$\begin{aligned}
&\mathbb{E}_{Z|\theta}[\mathcal{L}] = \\
&\sum_i\Big[\sum_j\left(\log\binom{R_{i,j}}{\mathbf{X}_{i,j}} + \mathbf{X}_{i,j}\log(\pi_{i,j}) + (R_{i,j}-\mathbf{X}_{i,j})\log(1-\pi_{i,j})\right) \\
&+ \sum_c\left(\mu_{i,c}\log(\mu_{i,c}) + (1-\mu_{i,c})\log(1-\mu_{i,c})\right)\Big].
\end{aligned} \tag{13}$$

where the sums are over locus indices $1\leq i\leq N$, subsection indices $1\leq j\leq M$, and clone indices $1\leq c\leq C$. We used the fact that $\mathbf{Z}_{i,c}$ is binary and $\Pr(\mathbf{Z}_{i,c}=1|\theta)=\mu_{i,c}$ to derive the above equation.

4 Simulation results

To validate our implementation of the EM optimization procedure and to understand our model's behavior, we produced simulated deep sequencing data and measured the extent to which the model successfully recovers the true clonal structure of the data.

For each simulation, we began by randomly generating four matrices. First, we generated a simulated matrix $R_{N\times M}$ of total read counts with respect to a fixed number ($N=20$) of loci and a fixed number ($M\in\{3,\ldots,15\}$) of subsections with a mean coverage of 1000 reads per locus. The matrix was generated by independently sampling each column (corresponding to a single subsection) from a multinomial distribution $\text{Multinomial}(1000N, \frac{1}{N},\ldots,\frac{1}{N})$, where the parameters $1000N$ and $\frac{1}{N}$ correspond to the total number of trials, and the probability of success for each of the N loci, respectively. Second, for any clone number $C\in\{c|3\leq c\leq 5 \text{ and } c\leq M\}$, we generated a corresponding Boolean matrix $Z_{N\times C}$, in which the entry at row i and column c indicates whether locus i exhibits the variant allele in clone c. Entries in Z were generated independently from a Bernoulli distribution with a probability of success $m=0.7$, with the exception of the first ("normal") column of Z, which contains all zeroes. Third, we generated a clone frequency matrix $P_{C\times M}$ as follows: each element of P is independently drawn from a Uniform [0,1] distribution, and then each column of P was divided by the column sum, so that the columns summed to 1. We then set $P^1=(1,0,\ldots,0)$ so that the first column of P corresponds to the normal subsection. Finally, for each locus i and subsection j, we generated the observed number of variant alleles $X_{i,j}$ by sampling from a binomial distribution with parameters $R_{i,j}$ (representing the total number of reads) and $\frac{1}{2}Z_i\cdot P^j$ (representing the probability that a given read corresponds to the variant allele). This last step complies with our primary assumption about the distribution of the data (Equation 3).

We ran the EM algorithm using the simulated data R and X and then evaluated the extent to which the estimated clone frequency matrix $\hat{P}$ and mutation probability matrix $\hat{\mu}$ differed from the corresponding true matrices P and Z. Specifically, we computed the genotype error e_Z, defined as

$$e_Z = \frac{1}{NC}\sum_{i=1}^{N}\sum_{c=1}^{C}\left|\hat{\mu}_{i,c}-Z_{i,c}\right|$$

and the clone frequency error e_P,

$$e_P = \frac{1}{CM}\sum_{c=1}^{C}\sum_{j=1}^{M}\left|\hat{P}_{c,j} - P_{c,j}\right|.$$

Note that, because we did not know which columns of $\hat{\mu}$ correspond to which columns of Z, we compared Z to every permutation of the columns of $\hat{\mu}$ and selected the permutation that resulted in the smallest genotype error. The selected permutation was then also used in the calculation of the clone frequency error.

Our simulation results (Figure 2) exhibit two primary trends. The overall error rate, as measured by either genotype or clone frequency error, decreases systematically as the number of subsections increases, and increases as the number of clones increases. Overall, both error rates are low, especially for $C = 3$. The observed trends are expected: for a fixed number of clones, the availability of more subsections leads to more accurate estimation of the true parameter values; and for a fixed number of subsections, the presence of more clones leads to a greater number of parameters that must be inferred, leading to greater error in estimation.

To assess the affect of sequencing error on the performance of Clomial, we added noise to the simulated data and repeated the above experiments. Specifically, we modeled noise by Bernoulli random variables with probability of success interpreted as the probability that a non-tumor allele is read as a tumor allele or vica versa. Running the EM algorithm on the noisy data revealed that Clomial is robust with respect to noise for all reasonable levels of sequencing error (Figure S6) in Text S1.

5 Application to a primary breast cancer

We obtained breast cancer tissue from a 44 year old premenopausal female with infiltrative ductal carcinoma (IDC) with ductal carcinoma in situ (DCIS), stage pT1c pN1, Grade II/III, estrogen receptor (ER) positive, progesterone receptor (PR) positive and Her2 negative. Axillary lymph node dissection revealed that one out of 13 nodes was positive for metastatic disease. A total of 6 tissue sections were obtained, including 2 sections from adjacent normal breast tissue, 3 from the primary breast cancer, and 1 from the positive lymph node. The tumor content, including both IDC and DCIS, ranged from 40% to 55% in the primary tumor and axillary lymph node tissue sections based on pathological examination. For subsequent analysis, each tissue section was subdivided into subsections (Figure 3).

To identify mutations and quantify allele frequencies, we performed two rounds of DNA sequencing. Initially, DNA was extracted from each individual subsection and subjected to exome capture followed by Illumina sequencing. Variants were detected independently in each subsection using the SeattleSeq Annotation Server. We focused on single nucleotide variants and short indels that exhibited a coverage of >15 reads in at least one of the subsections, ranking them using DeepSNV [31] and Fisher's exact test (Methods). This analysis produced an initial set of 281 variants (Dataset S1).

To better quantify the allele frequencies at these loci, we designed primer pairs surrounding each locus and used these primers to perform a second round of targeted DNA sequencing. This experiment successfully sequenced 244 of the 281 loci, with a mean and median coverage of 1615 and 1118, respectively, reads per locus. Each of these loci was individually validated by visual inspection using the Integrative Genomics viewer (IGV). Manual inspection showed that many of the initially identified mutations were flanked by homopolymer repeats, suggesting that the alternate alleles were read calling errors, rather than true mutations [32]. For all downstream analysis we focused on a set of 17 confirmed somatic variants. For clarity of presentation, we refer to each somatic variant by the chromosome where it resides, appending a letter if more than one somatic variant occurred within a chromosome (Table S1 in Text S1). The targeted sequencing thus produced two 17-by-12 matrices containing, respectively, the total coverage and the tumor allele count at each locus (Table S1 in Text S1). Visual inspection of the allele frequency profiles shows, not surprisingly, a markedly different pattern of allele frequencies among the subsections from primary and metastatic sites (Figure 3). In addition, several of the samples (e.g., P1-4 and P3-1) exhibit consistently lower frequencies across all loci, presumably indicating a higher prevalence of normal cells within these samples.

We applied our EM optimization procedure to the two counts matrices, varying the number of assumed clones from C = 3 up to C = 6. For each value of C, we ran EM 100,000 times from different random initializations, and we selected the solution with the highest likelihood (Figure 4). The resulting three-clone solution identifies two mutations, chr4a and chr9b, that occur in both the primary and metastatic samples and segregate the remaining mutations into nine that occurred in the primary tumor and six that occurred in the metastatic lymph node. The four- and five-clone solutions further subdivide the primary tumor mutations, and the six-clone solution separates the two metastatic mutations into distinct clones.

To better understand the inferred clonal landscape, we investigated the relationship between clone frequencies and the anatomy of the three primary and one metastatic tumor sections. We hypothesized that clone frequencies should vary smoothly between adjacent subsections, reflecting the physical spread of successful clonal populations. This hypothesis is supported by the data (Figure 5 and Figure S1 in Text S1). The trends are most striking in sections P1 and P2, for which we obtained four separate subsections. In each case, the primary clone frequencies vary in a monotonic fashion as we traverse the sample. Given that the EM inference procedure was provided with no information about which subsection was derived from which section, nor the relative orientation of the subsections to one another, the smoothly varying frequencies among adjacent subsections provides evidence that the method has successfully identified true clonal variation.

6 Tumor phylogeny

Cancer progression is an evolutionary process in which clones accrue mutations over time, forming new clones. Accordingly, it should be possible to organize the clonal progression of a tumor into a phylogenetic tree with the founder clone at the root. We therefore investigated whether the clones inferred by our EM procedure obey some simple phylogenetic constraints, with two complementary goals. First, because our EM procedure makes no use of phylogenetic constraints, this analysis can provide further evidence for the validity of our inferred solutions. Second, the phylogenetic analysis has the potential to provide significant insights into the clonal and mutational history of this specific cancer.

We started with the C = 3 solution to our EM algorithm, manually constructing a phylogenetic tree in which each node is a clonal population, and edges are marked with the mutations that occurred in the evolution from the parent clone to the offspring (Figure 6A). This particular tree shows two founder mutations, chr4a and chr9b, occurring prior to metastasis, six mutations occurring along the metastatic lineage, and nine along the primary

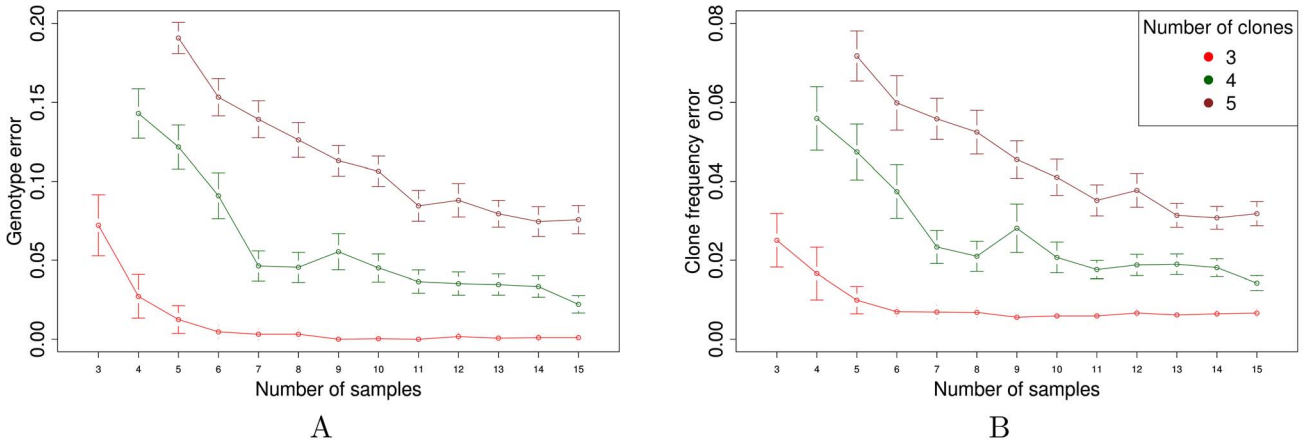

Figure 2. Simulation results. The figure plots the mean (A) genotype error e_Z and (B) clone frequency error e_P as a function of the number of subsections. Each mean is computed over 100 simulated data sets. For each data set, the EM optimization is repeated from 10 different random initializations, and the results corresponding to the largest log likelihood are reported.
doi:10.1371/journal.pcbi.1003703.g002

lineage. This is the only phylogenetic tree that is consistent with the inferred clonal genotypes.

In contrast, for the solutions inferred from the EM algorithm assuming C = 4 through 6, we found that it is not possible to construct a tree without requiring that the same mutation occur independently along multiple branches. We therefore considered all possible "nearby" trees (where "nearby" means that, among the distinct rows of the genotype matrix, the two trees differ by only one bit) that produce a valid phylogenetic tree with no repeated mutations. For example, for the C = 4 solution, we evaluated the likelihood of six nearby trees, yielding log-likelihoods of −28482, −21282, −7500, −6692, −5659, and −4333 (Table S2 in Text S1). The highest of these likelihoods is −4333, compared to −4244 for the solution initially inferred by EM. The selected solution requires changing only one bit in the genotype matrix from "0" to "1" (indicated by asterisks in Figure 4). The resulting phylogenetic tree (Figure 6B) closely resembles the C = 3 tree, except that one mutation initially assigned to the metastatic clone C3 is instead assigned to clone C2 in the C = 4 tree. Also, the nine mutations associated with the primary section in the C = 3 tree are further subdivided into three that occur shortly after metastasis and six that lead to clone C1. Reassuringly, the C = 5 and C = 6 solutions, constructed in a similar fashion (Figure 6C–D), are largely consistent with this story, each introducing a subdivision among the existing sets of mutations to produce a larger set of clones. Among these trees, the only inconsistencies concern (1) three mutations (chr5, chr9a and chr20b) that occur later according to the C = 4 solution than according to the C = 5 or C = 6 solutions and (2) two mutations (chr1 and chr4b) that are assigned their own branch, directly off the normal clone, in the C = 5 and C = 6 solutions. In practice, the chance that a randomly generated genotype matrix would produce a valid phylogenetic tree is vanishingly small (Note S3 in Text S1). Therefore, the fact that each of our inferred solutions very nearly produce a valid phylogenetic tree provides evidence for the validity of these solutions.

We also investigated the extent to which the observed mutation frequencies obey the phylogenetic tree. In principle, a mutation that occurs earlier in the evolution of the cancer should have a higher frequency than mutations that occur later along the same lineage because a child clone necessarily contains all of the mutations belonging to its parent clone. This investigation is hampered, however, by copy number variation. In practice, we cannot directly compare the allele frequencies of two distal sites because the observed allele frequencies are actually the product of mutation frequency and copy number. Empirically, we observe variation in copy number along the genome and differences in copy number variation from one subsection to the next (Figure S2 in Text S1). A consistent duplication of a large portion of chromosome 8 is known to occur commonly in breast cancer [33]. We were lucky, however, that two of our mutated loci occur quite close to one another on chromosome 9 (chr9a and chr9b, separated by only 3.3 Mbp). Given the observed data, the likelihood that a change in copy number occurring between these two loci is small, thereby allowing us to safely compare the corresponding mutation frequencies. Across all nine primary tumor subsections, we observe that the frequency of the parent mutation (chr9b) is higher than that of the child mutation (chr9a). Hence, these mutation frequencies are consistent with the inferred phylogeny.

To assess the stability of our inference, we performed leave-one-out analysis and compared the inferred phylogenies as follows. We held out each of the 12 tumor subsections one at a time and trained the model using the data from only 11 subsections for the case of C = 4. When samples p1-1 or p1-3 were excluded, the inferred genotypes were exactly the same as the genotype obtained from the full data. Excluding any of the other of 10 subsections resulted in a genotype which was different only in one bit; namely, the mutation chr4a was predicted to be present in all clones. However, this difference did not affect the inferred phylogeny because the change of this bit was in fact required to build a valid phylogenetic tree (Figure 4). In other words, by excluding any of the 12 tumor subsections, the inferred genotype always led to the same valid phylogenetic tree, which suggests that our algorithm is stable.

Discussion

Once a tumor has been resected, clinicians pay a great deal of attention to characterizing its anatomy. Features such as necrosis, extension beyond normal anatomical boundaries, and microvascular invasion convey important prognostic information. In addition, the cancer cells within any given tumor are frequently heterogeneous with respect to features such as differentiation state, the fraction of cells undergoing mitosis (as determined by Ki67

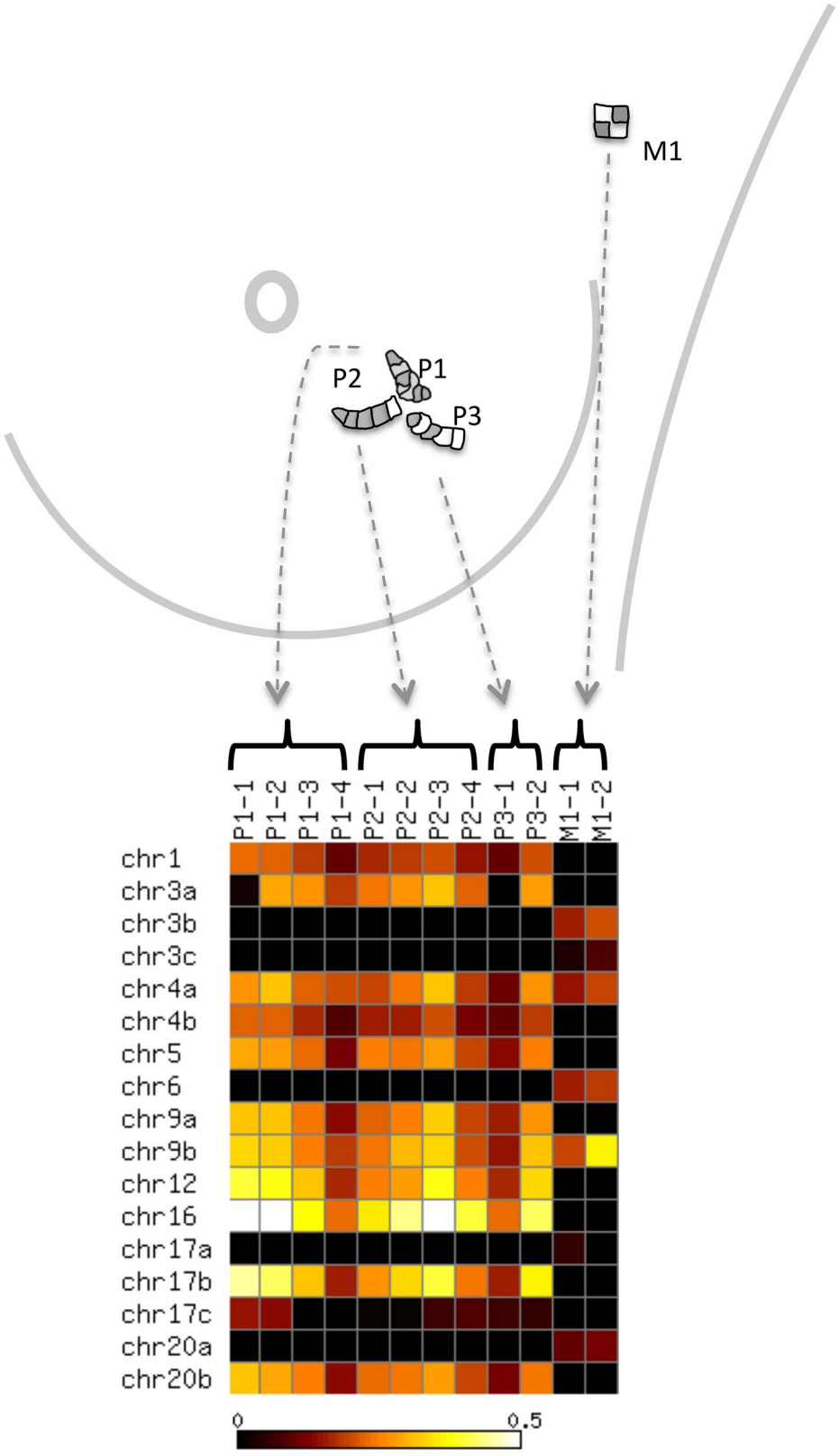

Figure 3. Anatomic locations of the sections, and corresponding allele frequencies. The figure shows (top) the anatomic locations of the three primary and one metastatic sections and (bottom) the corresponding alternative allele frequencies for each subsection. The full coordinates for each of the 17 loci are provided in Dataset S2.
doi:10.1371/journal.pcbi.1003703.g003

	C=3		C=4			C=5				C=6				
	C1	C2	C1	C2	C3	C1	C2	C3	C4	C1	C2	C3	C4	C5
chr4a	1	1	1	1*	1	1	1	0	1	1	1	0	1	1
chr9b	1	1	1	1	1	1	1	0	1	1	1	0	1	1
chr1	1	0	1	0	0	0	0	1	0	0	0	1	0	0
chr4b	1	0	1	0	0	0	0	1	0	0	0	1	0	0
chr12	1	0	1	1	0	1	1	0	0	1	1	0	0	0
chr16	1	0	1	1	0	1	1	0*	0	1	1	0*	0	0
chr17b	1	0	1	1	0	1	1	0	0	1	1	0	0	0
chr5	1	0	1	0	0	1	1	0	0	1	1	0	0	0
chr9a	1	0	1	0	0	1	1	0	0	1	1	0	0	0
chr20b	1	0	1	0	0	1	1	0	0	1	1	0	0	0
chr3a	1	0	1	0	0	1	0	0	0	1	0	0	0	0
chr17c	0	1	0	1	0	0	1	0	0	0	1	0	0	0
chr3c	0	1	0	0	1	0	0	0	1	0	0	0	1	0
chr17a	0	1	0	0	1	0	0	0	1	0	0	0	1	0
chr20a	0	1	0	0	1	0	0	0	1	0	0	0	1	0
chr3b	0	1	0	0	1	0	0	0	1	0	0	0	0	1
chr6	0	1	0	0	1	0	0	0	1	0	0	0	0	1

		N1-1	P1-1	P1-2	P1-3	P1-4	P2-1	P2-2	P2-3	P2-4	P3-1	P3-2	M1-2	M1-1
C=3	C0	1.00	0.47	0.43	0.57	0.75	0.77	0.51	0.61	0.55	0.44	0.63	0.77	0.84
	C1	0.00	0.51	0.55	0.43	0.25	0.21	0.48	0.38	0.44	0.54	0.36	0.00	0.00
	C2	0.00	0.02	0.03	0.00	0.00	0.01	0.01	0.00	0.00	0.01	0.02	0.22	0.16
C=4	C0	0.99	0.31	0.31	0.54	0.74	0.72	0.44	0.59	0.50	0.37	0.56	0.75	0.82
	C1	0.00	0.37	0.44	0.40	0.23	0.16	0.41	0.36	0.40	0.47	0.29	0.00	0.00
	C2	0.00	0.32	0.24	0.06	0.03	0.12	0.15	0.05	0.10	0.17	0.15	0.00	0.00
	C3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.18
C=5	C0	0.99	0.10	0.07	0.29	0.60	0.61	0.22	0.36	0.27	0.11	0.40	0.74	0.82
	C1	0.00	0.07	0.42	0.44	0.26	0.02	0.42	0.37	0.44	0.48	0.28	0.00	0.00
	C2	0.00	0.48	0.16	0.00	0.00	0.22	0.07	0.02	0.02	0.08	0.10	0.00	0.00
	C3	0.00	0.35	0.34	0.27	0.14	0.15	0.29	0.25	0.28	0.33	0.22	0.00	0.00
	C4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.18
C=6	C0	0.99	0.10	0.07	0.29	0.60	0.61	0.22	0.36	0.27	0.11	0.40	0.52	0.66
	C1	0.00	0.07	0.42	0.44	0.26	0.02	0.42	0.37	0.44	0.48	0.28	0.00	0.00
	C2	0.00	0.48	0.16	0.00	0.00	0.22	0.07	0.02	0.02	0.08	0.10	0.00	0.00
	C3	0.00	0.35	0.34	0.27	0.14	0.15	0.29	0.25	0.28	0.33	0.22	0.00	0.00
	C4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.10
	C5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.36	0.24

Figure 4. Inferred clonal genotypes and frequencies. The top table lists, for each of the 17 loci, the inferred clonal genotypes using the EM procedure, assuming C = 3, 4, 5 and 6. In each case, the normal clone (C0) is omitted from the inferred matrix Z, because its genotype consists entirely of zeroes by construction. For reference, each distinct genotype pattern per locus is assigned a unique color according to the scheme from Figure 6. In the table, bits with asterisks were flipped based on the phylogenetic analysis. The corresponding inferred clonal frequencies are listed in the bottom table, where each block shows a matrix P for a value of C, and C0 denotes the normal clone.
doi:10.1371/journal.pcbi.1003703.g004

staining), or (for breast cancer) the fraction of cells expressing HER-2 or estrogen receptor. The method described here provides a framework for linking a tumor's molecular anatomy to its structural anatomy as well as its phylogenetic evolution.

Several lines of evidence support the validity of the clonal genotypes and relative frequencies inferred by our model. One prediction from our phylogenetic reconstruction is that somatic variants at the trunk will be present at higher frequencies throughout all tumor subsections than variants appearing at the branches. While copy number variation across the somatic genome complicates these comparisons, one of two closely juxtaposed somatic variants (chr9b) is positioned at the trunk of our phylogenetic tree, while its neighbor (chr9a) arises in one of the branches. Consistent with this representation, the variant allele

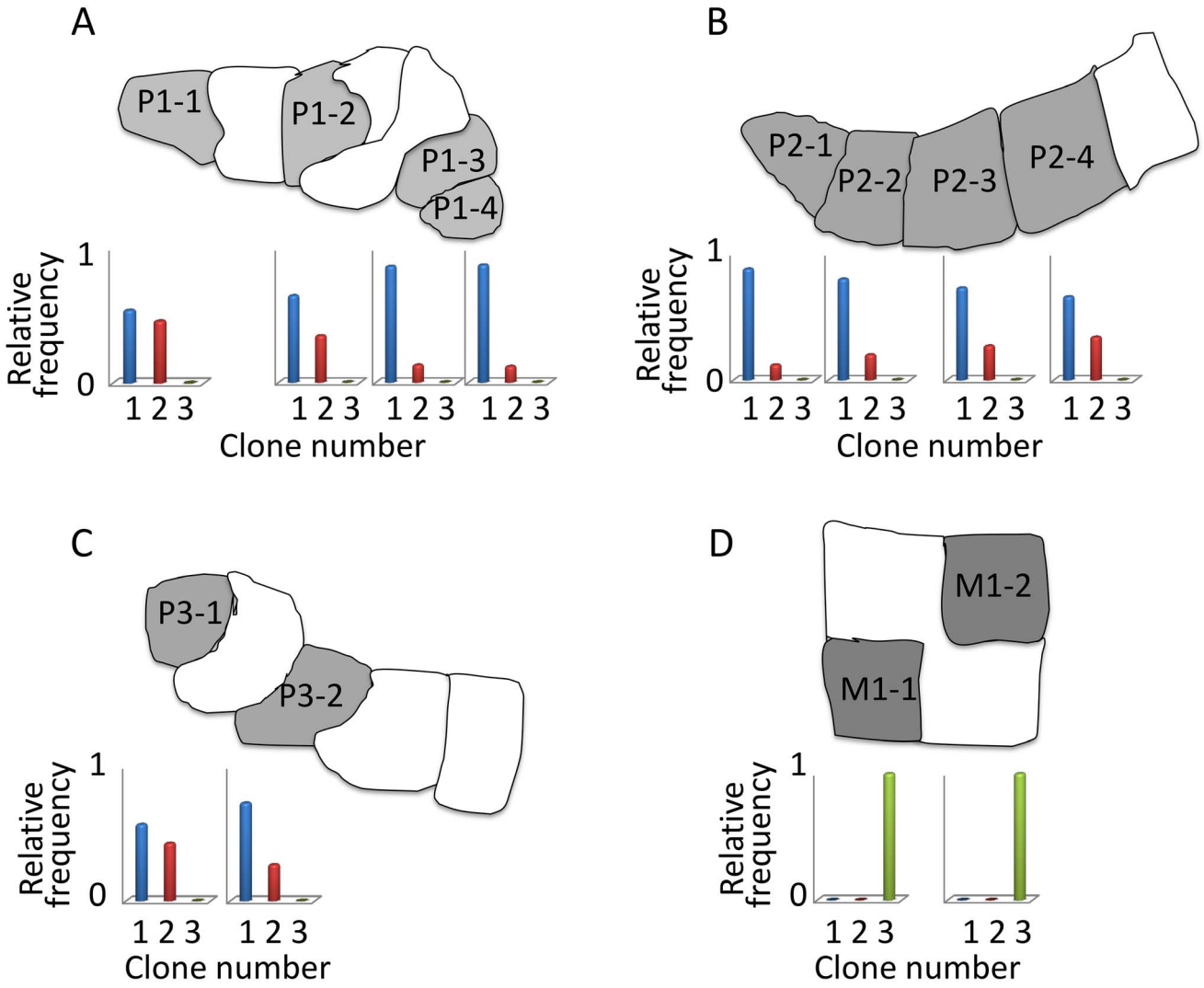

Figure 5. Clone frequencies vary smoothly across adjacent subsections. The panels display the pattern of inferred clone frequencies across subsections (A) P1, (B) P2, (C) P3 and (D) M1. Each bar plot shows the relative frequencies of tumor clones in the corresponding subsection after accounting for normal contamination. Clones are numbered as in Figure 4, and the normal clone, C0, is not shown. This figure shows the $C=4$ solution; Figure S1 in Text S1 shows the $C=5$ and $C=6$ solutions.
doi:10.1371/journal.pcbi.1003703.g005

frequencies for chr9b are consistently higher than for chr9a in all ten tumor subsections examined.

Interestingly, phylogenies can be built from the inferred genotypes even given the relatively low purity of the tumor sections: contamination with normal tissue was $>50\%$ in 9 out of 12 subsections in our data (Figure 4, $C=3$). In particular, although we estimate that the metastatic subsections contained $<20\%$ tumor cells in M1-1 and $<30\%$ in M1-2, the corresponding branch of the phylogenies is stable and consistent.

Similar to phylogenetic analysis, reassembly of the tumor subsections indicates that our assignment of mutations to clones produces spatial representations that are anatomically reasonable. With further refinements, our method should enable reconstructions that layer a tumor's phylogeny on top of its spatial organization.

While our results underscore the potential power of this new method, our study also has several limitations. Our assessments were confined to heterozygous somatic variants, and did not take into account the many chromosomal structural changes that were present in the tumor we examined. A comparison of exome copy numbers between primary tumor and lymph node indicates that the vast majority of these chromosomal changes preceded the divergence shown in our phylogenetic tree (Figure S2 in Text S1). In theory, one could imagine generalizing our generative model to take copy number variations into account by replacing the 2 in the denominator of Equation 2 with a hidden random variable for each locus, but without some form of aggressive regularization, this formulation would lead to a prohibitively complex and overfit model.

Additionally, a key characteristic of our method is the requirement to specify the number of clones C prior to the EM inference procedure. It is important to recognize that this choice should depend upon properties of the data set itself, rather than fundamental properties of the cancer. After all, each cell division results in multiple mutations, such that every cancer cell constitutes a distinct clone. Consequently, a picture of the full clonal history of a cancer would consist of a phylogenetic tree with one leaf for each cancer cell. In practice, such a tree would be of

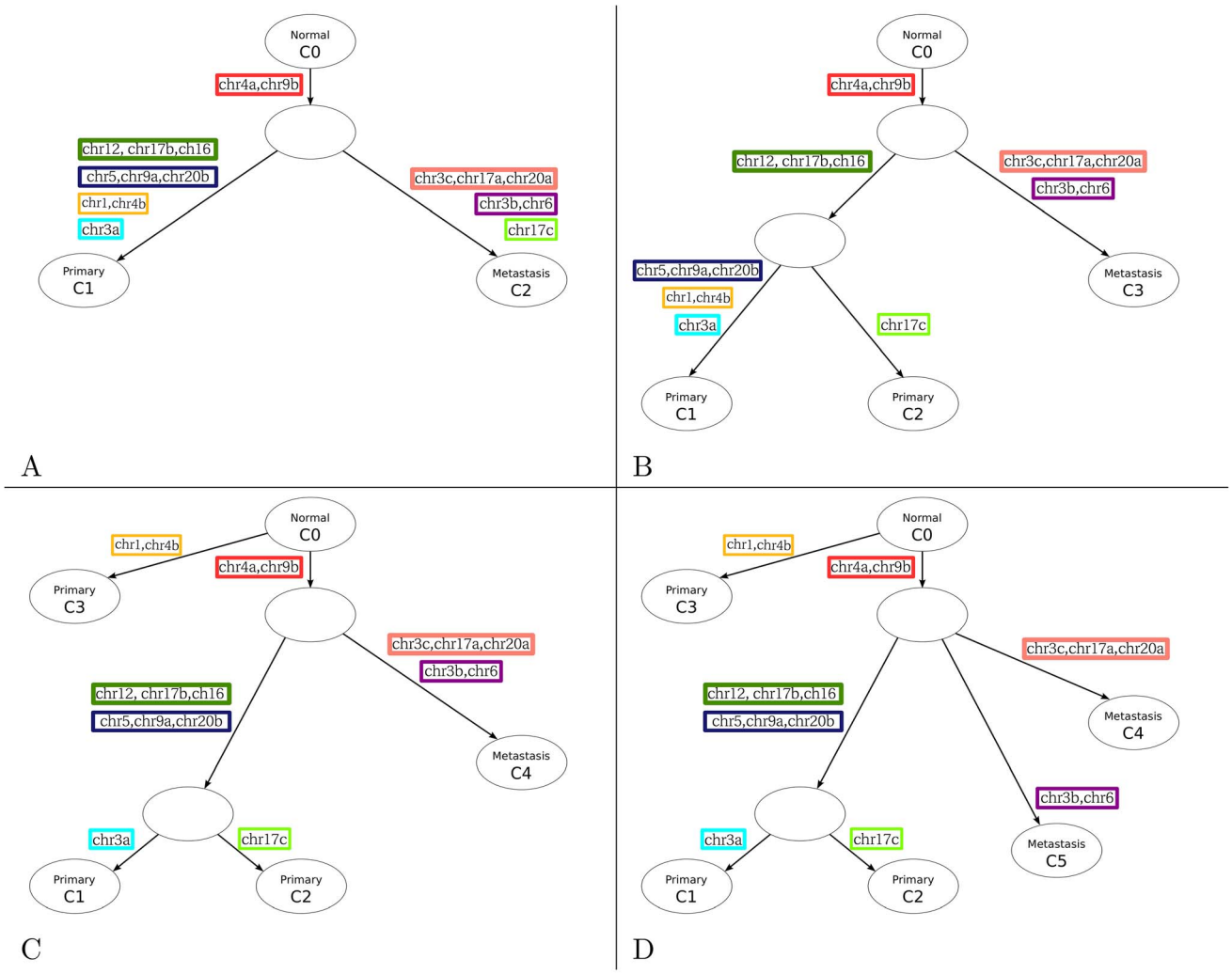

Figure 6. Cancer phylogenies. Each panel shows the inferred clonal phylogeny assuming (A) $C=3$, (B) $C=4$, (C) $C=5$ and (D) $C=6$ clones, where C0 corresponds to the normal clone. Nodes correspond to inferred clonal populations, and edges are annotated with mutations that occur between the parent and child clones. Two mutations are grouped into a colored box if they both occur on the same branch in all four phylogenies.
doi:10.1371/journal.pcbi.1003703.g006

limited utility and, more importantly, could not be accurately estimated from any reasonably sized data set. Perhaps the most useful definition of a tumor clone is a population of cells that exhibit distinct spatial or functional properties. Our approach allows the user to specify the number of clones and, hence, the resolution at which the clonal history is viewed.

Because Clomial does not impose any assumption on the distribution of mutation frequencies, the number of inferred clones may not exceed the number of samples; otherwise, the resulting optimization problem will be under-constrained.

In the particular cancer studied here, the three-clone solution appears to provide an inaccurate view of the clonal history. The placement of the chr17c mutation along the path leading to metastatic clone C2 is surprising, given that this particular locus has such low counts for both metastatic subsections (2 counts for subsection M1-1 and 0 counts for M1-2, Table S1 in Text S1). This apparent anomaly can be explained by the small counts associated with chr17c in four out of the 10 primary tumor subsections (3 counts in P1-3, 4 in P1-4, and 21 in each of P2-1 and P2-2). Faced with the choice of what genotype profile to assign to this particular locus, the inference procedure selected a solution in which only two subsections, rather than four, are inconsistent. However, given the flexibility of a 4-clone model, the anomaly is resolved, and chr17c defines a novel clone C2 that occurs in the primary tumor samples and is completely absent from the metastatic samples.

In practice, it may be possible to estimate how many clones the data set can resolve using a method such as the Bayesian Information Criterion (BIC), with a smaller BIC value indicating a better fit to the data [34–36]. This approach has been used previously for estimating tumor clonal composition [37,38]. BIC analysis of our model on simulated data suggests that, on average, the BIC accurately estimates the true number of clones, even in the presence of sequencing noise (Figure S3A–B in Text S1).

We also computed the BIC for models trained on our real breast cancer data (Figure S3C in Text S1) and observed a large decrease in BIC (45%) when C increases from 3 to 4, suggesting that the $C=3$ model is too simple to describe the data. However, the subsequent improvements of the BIC are smaller: 29%, 20%, 9%, and 3% respectively, as C grows from 4 to 8. In general, one should avoid increasing the complexity of the model when the BIC improvement is small because, in such situations, adding to the

number of free parameters of the model can potentially lead to over-fitting [39–47]. Note that, as an alternative to a BIC approach, one could instead take an approach motivated by cross-validation, as has been explored in the context of matrix factorization models [48–50].

Running the EM algorithm is very fast. In practice, using a 2.40 GHz processor with 2 GB memory, training a single EM instance on the real data set takes a few seconds up to several minutes, depending on the value of the hyperparameter C (Figure S4 in Text S1). However, because the optimization problem in the M step is non-convex, many EM instances must be trained from different random initializations to avoid local optima.

We first noted that Clomial achieved good results on simulated data using only 10 random initializations when $C=3$ (Figure 2). Then, to further assess the appropriate number of EM instances to run, we revisited the solutions from all of our 100,000 EM instances, counting how many instances are required to achieve the best observed model (Figure S5 in Text S1). In practice, while 1000 EM instances is sufficient to find the optimum solution when $C=2$ or 3, a larger number of random initializations is required as the number of clones grows. This is an expected phenomenon because the complexity of the model grows significantly with C, resulting in an optimization surface with many more local optima. Consequently, despite the highly parallel nature of the computation, scaling up to analysis of larger data set with larger numbers of clones will likely require improved EM training strategies, such as noise injection or regularization.

Finally, although we used a simple phylogenetic tree construction procedure to evaluate the quality of our inferred clonal genotypes, the EM inference procedure described here does not explicitly model tumor evolution. Ultimately, we aim to produce a model that automatically infers not only clonal genotypes and clonal frequencies, but also the number of clones and the phylogenetic tree relating them.

Our method differs significantly from other approaches. A recent characterization of 21 breast cancers defined clones by clustering mutations with similar variant allele frequencies [9]. The success of this strategy hinges on characterizing the frequencies of large numbers (hundreds or thousands) of somatic variants. In contrast, our method can reconstruct clonal phylogenies based on accurately measuring alleles of much smaller numbers of somatic variants. The view afforded by our method may provide novel insights into tumor biology. In particular, results from Nik-Zainal and colleagues [9] were interpreted to indicate that cancers become clinically apparent only after one of the competing clones has achieved clonal dominance. In contrast to this "winner takes all" hypothesis, our model suggests that some cancers might be more accurately regarded as ecosystems, in which clones may be subject to spatial influences that affect their competitive fitness, or may even collaborate to support tumor growth.

An important difference between our method and many other methods based on clustering [8,9,12] is our explicit probabilistic modeling of the random selection of normal and variant alleles during sequencing, according to a binomial distribution. By taking into account not just the relative frequency of the two alleles but the separate counts of normal and variant alleles, our model automatically assigns less importance to a locus with lower coverage, even if the locus yields the same variant allele frequency as a high-coverage locus.

While this manuscript was under review, two methods called PyClone [13] and PhyloSub [51] were published, which do model allele counts using a binomial distribution. These methods attempt to simultaneously infer not only clonal genotypes and frequencies, as Clomial does, but also infer the number of clones and their phylogeny. Furthermore, PyClone and PhyloSub are not limited, as Clomial is, to situations in which the number of inferred clones is less than or equal to the number of available samples. How is this possible? To make these inferences feasible, these clustering methods must make certain distributional assumptions about the data. Specifically, PyClone assumes a Dirichlet Process prior for clone frequencies, where the base distribution is Uniform [0,1] and the concentration parameter is Gamma distributed with shape and scale parameters equal to 1 and 0.001, respectively. PhyloSub extends PyClone by using a tree-structured stick-breaking process [52] to directly account for phylogenetic relationships during the inference. In principle, these assumptions enable PyClone and PhyloSub to infer information about a large number of clones from only a single sample. On the other hand, when multiple samples are available, Clomial can draw accurate inferences without requiring these distributional assumptions. In practice, our comparison showed that Clomial and PhyloSub produce similar results on three previously described chronic lymphocytic leukemia (CLL) cases [53] (Tables S3–S5 in Text S1).

We note that if e_i is the sequencing error rate at locus i, then the probability of observing a variant allele at this locus in subsection j is estimated by $\pi_{i,j}(1-e_i)+(1-\pi_{i,j})e_i$. In principle, sequencing noise could be incorporated into our model by replacing $\pi_{i,j}$, defined in Equation 2, with $\pi_{i,j}(1-2e_i)+e_i$ in the likelihood and EM algorithm. However, given the robustness of the current method to noise (Figures S6 and S3C in Text S1), we opted to keep our model simple. In future applications, it may be beneficial to model noise in data produced by sequencing technologies that exhibit high error rates ($>0.02\%$) such as PacBio RS [54].

The EM algorithm is not the only option for maximizing the log-likelihood for the observed data. In particular, one could instead treat both Z and P as optimization variables and seek to maximize $\mathcal{L}(\mathbf{X}|Z,P)$ with respect to Z and P. This would amount to iteratively updating Z and then updating P until convergence, similar to the iterative algorithms typically used for matrix factorization models [23–25,50]. However, this alternative approach would not have any computational advantage in terms of the update for P, which would still not have a closed-form solution, and would need to be solved using BFGS-B or an equivalent approach. Furthermore, the update for Z would be very complicated under the constraint that Z is a binary matrix. Therefore, we developed a customized inference algorithm based on EM.

Whereas genetic testing for cancer patients today focuses on mutations affecting a relatively small number of cancer-associated genes, most cancers are sustained by networks of aberrantly regulated genes that collaborate to promote tumor growth. The ability to assign mutations to clones, and to layer a tumor's clonal content on top of its structural anatomy in space and over time, can provide new insights into the mechanisms that enable cancers to invade, metastasize and escape treatment.

Materials and Methods

Ethics statement

This research was reviewed and approved by the Cancer Consortium Institutional Review Board (IRB) located at the Fred Hutchinson Cancer Research Center (FHCRC). The FHCRC has an approved Federalwide Assurance on file with the Office for Human Research Protections (number 00001920). The Federalwide Assurance is a formal written, binding commitment that assures that the FHCRC promises to comply with the regulations and ethical guidelines governing research with human subjects, as

stipulated by the U.S. Department of Health and Human Services under 45 CFR 46. Because this study involved the use of de-identified specimens obtained from an IRB-approved repository, we did not interface with patients. Patient consent was administered, in compliance with 45 CFR 46, by investigators who maintain the repository. Patients gave their consent for their specimens to be stored in the repository and subsequently used for research in cancer. The FHCRC IRB deemed that our research was in concordance with the purpose of the registry and the patient informed consent.

Breast cancer tissue sample

We obtained breast cancer tissues from the Breast Cancer Biospecimen Repository of Fred Hutchinson Cancer Research Center after IRB approval. The patient was a 44 year old pre-menopausal woman diagnosed with infiltrative ductal carcinoma (IDC) and ductal carcinoma in situ (DCIS), stage pT1c pN1, Grade II/III, ER positive, PR positive and Her2 negative. Axillary lymph node dissection revealed that one out of 13 nodes was positive for metastatic disease. A total of 5 pieces were obtained from surgical samples including 1 tissue section from adjacent normal breast tissue (N1), 3 tissue sections from the primary breast cancer (P1, P2, P3), and 1 tissue section from the positive axillary lymph node (M1). Each section is about 1 cm by 1 cm by 0.5 cm. The tumor content, including both IDC and DCIS, ranges from 40% to 55% in the primary tumor and axillary lymph node tissue sections based on pathological examination (P1 55% IDC, P2 45% IDC, P3 40% IDC and 15% DCIS, M1 50% IDC).

Tissue DNA extraction

Each individual section was subdivided into multiple subsections, and the anatomic locations of all the subsections were recorded (Figure 3). Using Qiagen AllPrep DNA/RNA Micro Kit, DNA was extracted from one normal subsection (N1-1), seven primary subsections (P1-2, P1-3, P1-5, P2-1, P2-3, P3-3, P3-4) and one metastatic subsection (M1-1). After quantification, all the DNA samples were subjected to exome capture followed by Illumina sequencing.

Whole exome sequencing

Next generation sequencing was carried out at the Northwest Genome Center at University of Washington on the normal subsection, seven primary subsections, and one metastatic subsection. For each subsection, one microgram of genomic DNA was used to construct the random-shearing library per standard protocol with Covaris acoustic sonication. Libraries then underwent exome capture using the ~ 36.5 Mb target from Roche/Nimblegen SeqCap EZ v2.0 ($\sim 80{,}000$ exons and flanking sequence). Since each library was uniquely barcoded, samples were performed in multiplex. Massively parallel sequencing was carried out on the HiSeq sequencer.

Read processing

Sequence reads were processed with a pipeline consisting of the following elements: (1) base calls generated in real-time on the HiSeq instrument (RTA 1.12.4.2); (2) Perl scripts developed in-house to produce demultiplexed fastq files by lane and index sequence; (3) demultiplexed BAM files aligned to a human reference (hg19) using BWA (Burrows-Wheeler Aligner; v0.5.9) [55]. Read-pairs not mapping within ± 2 standard deviations of the average library size ($\sim 125 \pm 15$ bp for exomes) are removed. All aligned read data were subjected to the following steps: (1) "duplicate removal" was performed, (i.e., the removal of reads with duplicate start positions; Picard MarkDuplicates; v1.14); (2) indel realignment was performed (GATK IndelRealigner; v1.0-6125) resulting in improved base placement and lower false variant calls; (3) base qualities were recalibrated (GATK TableRecalibration; v1.0-6125). All sequence data then underwent a previously described quality control protocol [56].

Variant detection

Variant detection and genotyping were performed using the UnifiedGenotyper tool from GATK (v1.0-6125). Variant data for each sample were formatted (variant call format) as "raw" calls that contain individual genotype data for one or multiple samples, and flagged using the filtration walker (GATK) to mark sites that are of lower quality/false positives, e.g., low quality scores (≤ 50), allelic imbalance (≥ 0.75), long homopolymer runs (>3) and/or low quality by depth (QD <5).

Calling single nucleotide variants (SNVs) and indels

Most of the commonly used software for calling SNVs and indels, including SNVMix [57] and VarScan [58], requires tumor content $>80\%$. To allow identification of low frequency alleles that occur in only one or a few subsections, we did not pool all of the data together. Instead, we designed a method that is appropriate for multiple samples from one patient, with relatively low tumor content, ranging from 45% to 55%. At each chromosomal position (locus), we considered six mutually exclusive possible outcomes: A, C, G, T, deletion, and unknown. The counts of these six outcomes at each locus between normal and each of the multiple tumor subsections were compared with a 2×6 Fisher's exact test. To correct for multiple testing, we used the **qvalue** R package to convert $p-values$ to $q-values$. Only those chromosomal loci with $q<0.1$ in at least one comparison between normal and tumor samples were accepted for downstream analysis. This analysis identified 6310 loci.

For each accepted locus, we used a heuristic procedure to identify which of the six alleles differed between the tumor and normal sample. For each subsection, we carried out six 2×2 Fisher's exact tests, one for each of the six possible alleles. Thus, each such test compared one allele's counts to the sum of the counts for the other five alleles. Using a p-value threshold of 0.01, an allele was declared to be increased, decreased, or unchanged in the tumor subsection as compared to the normal sample. The changes that were classified as "increased" and had a normal count of zero were called tumor-specific mutations. This procedure identified a total of 268 such tumor-specific mutations, with a mean and median sequencing depth of 92 and 75, respectively. Corresponding annotations were obtained from SeattleSeq (http://snp.gs.washington.edu/SeattleSeqAnnotation137).

In parallel, we also analyzed our data using deepSNV [31] by comparing the normal subsection to the 8 tumor subsections. We ran deepSNV on the loci with total coverage across all samples more than 50, which resulted in the identification of 29 loci with $q \leq 0.1$. The union of the two lists yielded 281 loci for further validation (Dataset S1).

Targeted deep sequencing

Mutations were validated by targeted deep sequencing of DNA derived from one normal subsection (N1-1), 10 primary subsections (P1-1, P1-2, P1-3, P1-4, P2-1, P2-2, P2-3, P2-4, P3-1, P3-2) and two metastatic subsections (M1-1 and M1-2). The subsections were selected to have low normal content and to span the tumor anatomy. Genomic DNA was prepared as described for the initial exome sequencing. A HaloPlex probe capture library for selective capture of 281 target loci was generated with SureDesign (Agilent

Technologies). Target enrichment for deep sequencing was carried out with the HaloPlexTM Target Enrichment System from Agilent Technologies following the manufacturer's protocol. Triplicate enrichments were performed for each sample. Target-enriched samples were sequenced using a MiSeq (Illumina). Of the 281 target loci, 244 were successfully sequenced with coverage more than 100 reads for the normal sample. The mean, median, and the standard deviation of the coverage were 1615, 1118, and 1600, respectively (Dataset S2).

All 244 loci were visualized using the Integrative Genomics Viewer [59,60]. A set of 17 loci were selected based upon three criteria: (1) at least 3 reads cover the locus in the normal sample, (2) the variant allele is not present in the normal tissue (allowing for a few variant counts, which may reflect sequencing error) and (3) there are no nearby clustered mutations, indicative of sequencing or mapping error. Independently, the data were also analyzed using deepSNV. Applying a $q-value$ threshold of $p \leq 10^{-6}$ yielded 19 loci, including all 17 of the initially selected loci. The 17 loci were retained for downstream analysis (Table S1 in Text S1).

Bayesian Information Criterion analysis

We computed BIC using the following formula:

$$\mathcal{BIC} = -2\mathcal{L}^* + (NC + MC - M)\log(\|R\|_1), \qquad (14)$$

where $\mathcal{L}^*$ is the expectation of the complete-data log likelihood, which is maximized in the last M step (see Equations 7 and 13). Also, $(NC+MC-M)$ represents the total number of free parameters, and $\|R\|_1 = \sum_{i,j} R_{i,j}$ is the total number of counts.

Supporting Information

Dataset S1 Dataset S1 includes, for each of the 281 targeted loci, the following information: (1) the chromosomal coordinates of the locus, (2) the variant allele, (3) for the 268 targeted loci identified by Fisher's exact test, the number of reads supporting the mutation and the coverage of the locus in each subsection, (4) the $q-value$ from the 2×6 Fisher's exact test, (5) the minimum $p-value$ from the six 2×2 Fisher's exact tests, and (6) the q-value for each of the 29 target regions identified by deepSNV.
(XLS)

Dataset S2 Dataset S2 lists, for each of the 281 sequenced target regions, the following information: (1) the chromosomal coordinates of the locus, (2) the integer counts for each of the five possible alleles (A, C, G, T, -, where "-" denotes deletion or insertion) in each of the ten primary subsections, two metastatic subsections, and the normal subsection, (3) for each subsection, the deepSNV $p-value$ for the test that the subsection has a mutation on each specific locus, and (4) the mnemonic for each of the 17 mutations used in our inference procedure.
(XLS)

Software S1 Software S1 is an R package called *Clonality* that implements the EM algorithm described in this paper.
(GZ)

Text S1 Text S1 includes five supplementary notes, six supplementary figures and five supplementary tables.
(PDF)

Acknowledgments

The authors would like to thank Peggy Porter and Barbara Stein for providing access to the breast cancer specimen through Northwest BioTrust.

Author Contributions

Conceived and designed the experiments: HZ JW DW CAB WSN. Performed the experiments: JW JS CS. Analyzed the data: HZ JW AH JS KW. Wrote the paper: HZ JW DN DW CAB WSN.

References

1. Irish JM, Hovland R, Krutzik P, Perez OD, Bruserud O, et al. (2004) Single cell profiling of potentiated phospho-protein networks in cancer cells. Cell 118: 217–228.
2. Navin N, Krasnitz A, Rodgers L, Cook K, Meth J, et al. (2010) Inferring tumor progression from genomic heterogeneity. Genome research 20: 68–80.
3. Xu X, Yong Y, Yin X, Bao L, Tang A, et al. (2012) Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell 148: 886–895.
4. Hou Y, Song L, Zhu P, Zhang B, Tao Y, et al. (2012) Single-cell exome sequencing and monoclonal evolution of a *JAK2*-negative myeloproliferative neoplasm. Cell 148: 873–885.
5. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, et al. (2011) Tumour evolution inferred by single-cell sequencing. Nature 472: 90–94.
6. Potter NE, Ermini L, Papaemmanuil E, Cazzaniga G, Vijayaraghavan G, et al. (2013) Single cell mutational profiling and clonal phylogeny in cancer. Genome research 23: 2115–25. doi: 10.1101/gr.159913.113.
7. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, et al. (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature 461: 809–813.
8. Welch JS, Ley TJ, Link DC, Miller CA, Larson DE, et al. (2012) The origin and evolution of mutations in acute myeloid leukemia. Cell 150: 264–278.
9. Nik-Zainal S, Alexandrov LB, Wedge DC, Van LP, Greenman CD, et al. (2012) Mutational processes molding the genomes of 21 breast cancers. Cell 149: 979–993.
10. Ding L, Ellis MJ, Li S, Larson DE, Chen K, et al. (2010) Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature 464: 999–1005.
11. Bashashati A, Ha G, Tone A, Ding J, Prentice LM, et al. (2013) Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. The Journal of pathology 231: 21–34.
12. Shah SP, Roth A, Goya R, Oloumi A, Ha G, et al. (2012) The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature 486: 395–399.
13. Roth A, Khattra J, Yap D, Wan A, Laks E, et al. (2014) Pyclone: statistical inference of clonal population structure in cancer. Nature methods 11: 396–8. doi: 10.1038/nmeth.2883.
14. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, et al. (2012) Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature 481: 506–510.
15. Walter MJ, Shen D, Ding L, Shao J, Koboldt DC, et al. (2012) Clonal architecture of secondary acute myeloid leukemia. New England Journal of Medicine 366: 1090–1098.
16. Newburger DE, Kashef-Haghighi D, Weng Z, Salari R, Sweeney RT, et al. (2013) Genome evolution during progression to breast cancer. Genome research 23: 1097–108. doi: 10.1101/gr.151670.112.
17. Yachida S, Jones S, Bozic I, Antal T, Leary R, et al. (2010) Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature 467: 1114–1117.
18. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, et al. (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. New England Journal of Medicine 366: 883–892.
19. Strino F, Parisi F, Micsinai M, Kluger Y (2013) TrAp: a tree approach for fingerprinting subclonal tumor composition. Nucleic Acids Research 41: e165.
20. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B (Methodological) 39: 1–22.
21. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, et al. (2012) Absolute quantification of somatic dna alterations in human cancer. Nature biotechnology 30: 413–421.
22. Mwenifumbo JC, Marra MA (2013) Cancer genome-sequencing study design. Nature Reviews Genetics 14: 321–332.
23. Rennie JDM, Srebro N (2005) Fast maximum margin matrix factorization for collaborative prediction. In: Proceedings of the 22nd international conference on Machine learning. ACM, pp. 713–719.
24. Wu J (2009) Binomial matrix factorization for discrete collaborative filtering. In: Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on. IEEE, pp. 1046–1051.
25. Engelhardt BE, Stephens M (2010) Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS genetics 6: e1001117.

26. Bishop C (2006) Pattern Recognition and Machine Learning. New York, NY, USA: Springer Science+Business Media, LLC.
27. Fletcher R (1981) Practical methods of optimization: Vol. 2: Constrained optimization. Wiley.
28. Dennis JE, Schnabel RB (1987) Numerical Methods for Unconstrained Optimization and Nonlinear Equations, volume 16. Society for Industrial Mathematics.
29. Sun W, Yuan Y (2006) Optimization Theory and Methods: Nonlinear Programming, volume 1. Springer.
30. Wriggers P (2008) Nonlinear Finite Element Methods. Springer.
31. Gerstung M, Beisel C, Rechsteiner M, Wild P, Schraml P, et al. (2012) Reliable detection of subclonal single-nucleotide variants in tumor cell populations. Nature Communications 3: 811.
32. Moore M, Dhingra A, Soltis P, Shaw R, Farmerie W, et al. (2006) Rapid and accurate pyrosequencing of angiosperm plastid genomes. BMC Plant Biology 6: 17.
33. Curtis C, Shah SP, CHin SF, Turashvili G, Rueda OM, et al. (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486: 346–352.
34. Schwarz G (1978) Estimating the dimension of a model. Annals of Statistics 6: 461–464.
35. Watkins LP, Yang H (2005) Detection of intensity change points in time-resolved single-molecule measurements. The Journal of Physical Chemistry B 109: 617–628.
36. Powers DA, Xie Y (2008) Statistical methods for categorical data analysis. Emerald Group Publishing.
37. Oesper L, Mahmoody A, Raphael BJ (2013) Theta: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. Genome biology 14: R80–R80.
38. Chen M, Gunel M, Zhao H (2013) Somatica: identifying, characterizing and quantifying somatic copy number aberrations from cancer genome sequencing data. PloS one 8: e78143.
39. Speed TP, Yu B (1993) Model selection and prediction: normal regression. Annals of the institute of statistical mathematics 45: 35–54.
40. Shibata R (1989) Statistical aspects of model selection. Springer.
41. Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97: 611–631.
42. Zhao Q, Xu M, Franti P (2008) Knee point detection on bayesian information criterion. In: Tools with Artificial Intelligence, 2008. ICTAI'08. 20th IEEE International Conference on. IEEE, volume 2, pp. 431–438.
43. Zhao Q, Hautamaki V, Fränti P (2008) Knee point detection in bic for detecting the number of clusters. In: Advanced Concepts for Intelligent Vision Systems. Springer, pp. 664–673.
44. Satopaa V, Albrecht J, Irwin D, Raghavan B (2011) Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In: Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on. IEEE, pp. 166–171.
45. Zang C, Chen B (2010) Automatic estimation the number of clusters in hierarchical data clustering. In: Mechatronics and Embedded Systems and Applications (MESA), 2010 IEEE/ASME International Conference on. IEEE, pp. 269–274.
46. Singh DK, Ku C, Wichaidit C, Steininger RJ, Wu LF, et al. (2010) Patterns of basal signaling heterogeneity can distinguish cellular populations with different drug sensitivities. Molecular systems biology 6: 369. doi: 10.1038/msb.2010.22.
47. Lo K, Hahne F, Brinkman RRR, Gottardo RG (2009) flowclust: a bioconductor package for automated gating of flow cytometry data. BMC Bioinformatics 10: 145.
48. Wold V (1978) Cross-validatory estimation of the number of components in factor and principal components models. Technometrics 20: 397–405.
49. Owen AB, Perry PO (2009) Bi-cross-validation of the svd and the nonnegative matrix factorization. The Annals of Applied Statistics : 564–594.
50. Witten DM, Tibshirani R, Hastie T (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10: 515–534.
51. Jiao W, Vembu S, Deshwar AG, Stein L, Morris Q (2014) Inferring clonal evolution of tumors from single nucleotide somatic mutations. BMC Bioinformatics 15: 35.
52. Adams RP, Ghahramani Z, Jordan MI (2010) Tree-structured stick breaking for hierarchical data. In: NIPS. pp. 19–27.
53. Schuh A, Becq J, Humphray S, Alexa A, Burns A, et al. (2012) Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. Blood 120: 4191–4196.
54. Quail MA, Smith M, Coupland P, T D Otto T, Harris SR, et al. (2012) A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina miseq sequencers. BMC genomics 13: 341.
55. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 26: 589–595.
56. Tennessen JA, Bigham AW, OConnor TD, Fu W, Kenny E, et al. (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337: 64–69.
57. Goya R, Sun MG, Morin RD, Leung G, Ha H, et al. (2010) SNVMix: Predicting single nucleotide variants from next-generation sequencing of tumors. Bioinformatics 26: 730–736.
58. Koboldt DC, K KC, Wylie T, Larson DE, McLellan MD, et al. (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25: 2283–2285.
59. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in bioinformatics 14: 178–192.
60. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, et al. (2011) Integrative genomics viewer. Nature biotechnology 29: 24–26.