Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing

Yuchao Jiang[a,b], Yu Qiu[c,d,e,f], Andy J. Minn[c,d,e,f], and Nancy R. Zhang[b,1]

[a]Genomics and Computational Biology Graduate Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; [b]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104; [c]Abramson Family Cancer Research Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; [d]Department of Radiation Oncology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; [e]Abramson Cancer Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; and [f]Institute of Immunology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

Edited by David O. Siegmund, Stanford University, Stanford, CA, and approved July 12, 2016 (received for review November 10, 2015)

Cancer is a disease driven by evolutionary selection on somatic genetic and epigenetic alterations. Here, we propose *Canopy*, a method for inferring the evolutionary phylogeny of a tumor using both somatic copy number alterations and single-nucleotide alterations from one or more samples derived from a single patient. Canopy is applied to bulk sequencing datasets of both longitudinal and spatial experimental designs and to a transplantable metastasis model derived from human cancer cell line MDA-MB-231. Canopy successfully identifies cell populations and infers phylogenies that are in concordance with existing knowledge and ground truth. Through simulations, we explore the effects of key parameters on deconvolution accuracy and compare against existing methods. Canopy is an open-source R package available at https://cran.r-project.org/web/packages/Canopy/.

intratumor heterogeneity | cancer evolution | clonal deconvolution | cancer genomics | phylogeny inference

It has been long recognized that cancer is a disease driven by genetic and epigenetic alterations (1–3). These alterations confer upon its carrier cell selective advantage, and rounds of Darwinian selection produce tumor cell populations with aggressive phenotypes. High-throughput sequencing technologies have made possible the large-scale, high-resolution analysis of tumor genomes. A recurring finding of these studies is the high degree of heterogeneity—both intertumor heterogeneity among patients with the same clinical diagnosis (4, 5), as well as intratumor heterogeneity between tumor cells derived from the same patient (summarized in *SI Appendix*, Table S1) (6–12). Heterogeneity, at all levels, confound diagnosis and treatment. Most large-scale studies to date, for example, those led by the Cancer Genome Atlas Research Network (4) and the International Cancer Genome Consortium (5), have focused on intertumor heterogeneity. These studies typically collect and sequence bulk tissue data, usually one sample per patient, and compare the mutation profiles across patients. This study design is not optimized for the study of intratumor heterogeneity, which has thus received, until recently, comparatively less attention.

When only one sample from a tumor is sequenced, early analyses of intratumor heterogeneity started with the estimation of normal cell contamination and tumor ploidy (13, 14). For example, ABSOLUTE (13), one of the earliest methods, classifies mutations as clonal or subclonal after adjusting for the estimated purity and ploidy of the sample. Most approaches for the detection of subclonal mutations treat point mutations and copy number aberrations separately (15–18). In the case of point mutations, that is, single-nucleotide alterations (SNAs) and small insertions and deletions (indels), most methods rely on mixture models for the variant allele frequency (VAF) under the assumption that mutations carried by the same set of cells have the same VAF. However, the VAF is also affected by the copy number of the region where the point mutation resides, and copy number aberrations (CNAs) are prevalent in cancer. Recently, Li and Li (19) and Deshwar et al. (20) proposed models for joint inference of SNAs and CNAs. Li and Li (19) further gave important insight into the identifiability of the underlying parameters, if one were to analyze each mutation locus separately. The many unknowns, including the number of subpopulations in the tumor, the mutation profile of each subpopulation and its contributing proportion to the sample, and the phasing of aberrations that affect the same genome locus make the estimation problem challenging if one were to sequence only one bulk DNA sample from the tumor. We will discuss these underlying challenges through a more thorough literature review after giving a more detailed formulation of the problem.

Ultimately, tumor evolution occurs at the single-cell level, and single-cell methods provide a powerful approach to assess tumor heterogeneity without the confounding effects of mixed cell populations (6, 21). Despite its promise, single-cell DNA sequencing data are much noisier than bulk sequencing data due to allele dropout events and amplification errors (22), and furthermore, the per-cell coverage is still limited due to constraints

Significance

Cancer is a disease driven by rounds of genetic and epigenetic mutations that follow Darwinian evolution. The tumor for a given patient is often a mixture of multiple genotypically and phenotypically distinct cell populations. This contributes to failures of targeted therapies and to drug resistance, and thus it is important to study intratumor heterogeneity. Here, we propose Canopy, a statistical framework to reconstruct tumor phylogeny by next-generation sequencing data from temporally and/or spatially separated tumor resections from the same patient. We show that such analyses lead to the identification of potentially useful prognostic/diagnostic biomarkers and successfully recover the tumor's evolutionary history, validated by single-cell sequencing. Canopy provides a rigorous foundation for statistical analysis of repeated sequencing data from evolving populations.

Author contributions: Y.J. and N.R.Z. formulated the model; Y.J. developed the algorithm and methods; Y.J. and N.R.Z. planned and executed the simulations studies and the analysis of the human breast cancer xenograft dataset; Y.J., Y.Q., A.J.M., and N.R.Z. generated and analyzed the breast cancer cell line dataset; and Y.J. and N.R.Z. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The whole-exome sequencing data of the transplantable metastasis model derived from MDA-MB-231 have been deposited in the BioProject database, www.ncbi.nlm.nih.gov/bioproject (accession no. PRJNA315318).

[1]To whom correspondence should be addressed. Email: nzh@wharton.upenn.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1522203113/-/DCSupplemental.

on budget and labor. Although these single-cell sequencing studies have improved our understanding of intratumor heterogeneity, most current tumor studies still sequence the DNA at the bulk tissue level.

Recently, there have been increasing efforts to sequence the tumor from the same patient at multiple time points and/or from multiple spatially separated resections (7–12). Multiple snapshots of the same tumor have proved invaluable for identifying subclonal populations and for inferring the tumor's evolutionary history. Multidimensional scatterplots of VAFs allow higher resolution for cluster detection than the one-dimensional histogram in the single-sample case. Recent methods, such as Pyclone (15) and SciClone (16), apply Bayesian mixture models to detect these clusters. LICHeE (23) and SCHISM (24) infer phylogeny from VAFs as an acyclic directed graph network. Another recent work, Clomial (25), showed that it is possible to obtain precise and informative estimates of the underlying subpopulations through a matrix deconvolution framework. One practical drawback of Clomial (25) is that it takes only SNA input and assumes that all mutational loci are heterozygous from copy number-neutral regions. SCHISM (24), BitPhylogeny (26), PhyloWGS (20), and SPRUCE (27) adjust for CNAs in their model in different ways, but these methods still require limiting assumptions and do not make full use of the data, as we discuss in detail a bit later.

Here, we focus on the analysis of intratumor heterogeneity by multisample bulk DNA sequencing of tumor samples. We propose copy number and single-nucleotide alteration analysis of tumor phylogeny (Canopy), a statistical framework and computational procedure for identifying the subpopulations within a tumor, determining the mutation profiles of these subpopulations, and inferring the tumor's phylogenetic history. The input to Canopy are VAFs of somatic SNAs along with allele-specific coverage ratios between the tumor and matched normal sample for somatic copy number calls. These quantities can be directly taken from the output of existing software (28–31). Canopy provides a general mathematical framework for pooling data across samples and sites to infer the underlying phylogeny. For SNAs that fall within CNA regions, Canopy infers their temporal ordering and resolves their phase. When there are multiple evolutionary configurations consistent with the data, Canopy attempts to explore all configurations and assess their confidence.

Identifiability of the underlying evolutionary process and confidence in its reconstruction is an important aspect of consideration. The Bayesian framework for Canopy allows assessment of the quality of inference. The resolution at which clones can be differentiated depends on the data, and in particular, on how many slices of the tumor are taken, how genotypically different these slices are to each other, and sequencing depth. As the number of clones increase, the proportion of cells attributable to at least some of the subclones would necessarily decrease, and thus, the higher sequencing depth would be needed to detect mutations present in those clones. Under the Bayesian framework, the resolution of our estimates and the confidence in our conclusions can be quantified by the posterior distribution.

Results

We will start by giving a more precise formulation of the clonal decomposition problem along with a more in-depth discussion of existing methods and their key assumptions. We will show that, under our formulation, the likelihood of the observed sequencing data can be written in matrix form and be decomposed into terms that reflect the tumor's phylogenetic history, the phasing of overlapping SNAs and CNAs, and the contributing proportions of the admixed cell populations. Canopy assumes noninformative priors for the unknowns in the model, and explores their possible values by Markov chain Monte Carlo (MCMC). Through simulations, we explore the effects of various parameters on deconvolution accuracy and compare Canopy against existing methods. Canopy is then applied to four datasets with different sequencing designs: the whole-exome sequencing (WES) of a heterogeneous triple-negative breast carcinoma cell line MDA-MB-231 and its derived sublines with single and mixed cell populations, the whole-genome sequencing (WGS) of breast cancer patient xenografts from Eirew et al. (10), the WGS of leukemia patients, each at two time points from Ding et al. (7) (*SI Appendix, SI Results*), and the multiregion sequencing of an ovarian cancer patient from Bashashati et al. (8) (*SI Appendix, SI Results*).

Modeling of SNAs, CNAs, and Clonal Tree. Fig. 1*A* shows the phylogeny of an evolving tumor, which starts from a diploid normal cell and progresses through waves of somatic mutations. The tumor's evolution is depicted as a bifurcating tree, with the ancestral normal cell population at the root, and accumulating mutations along its branches. Time runs vertically down the tree from the root, and when a sample of the tumor is taken at any point in time, the tree is sliced horizontally, cutting the branches to form leaves. The subpopulations within the sample are represented by the "leaves" in that slice. Each subpopulation contributes a fraction of cells to the sample, which, taken together, are represented by a vector of nonnegative numbers that sum to 1. To model normal cell contamination, we restrict the left-most branch of the tree to be nonbifurcating and mutation-free. Thus,

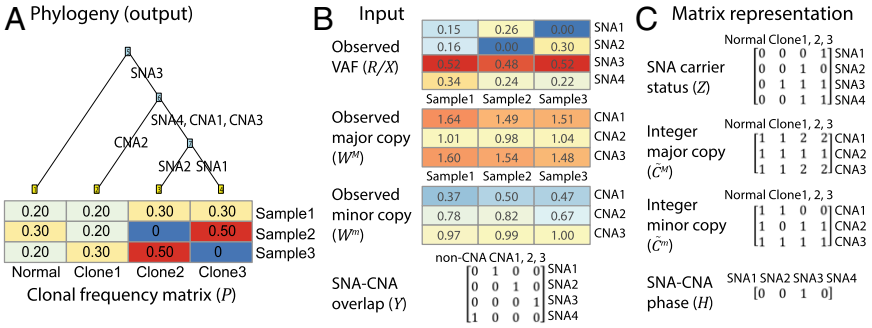


Fig. 1. Tumor phylogeny, observed input, and inferred output of Canopy. (*A*) Phylogeny of tumor progression as a bifurcating tree with SNAs and CNAs along the branches. Longitudinal and/or spatial samples offer different snapshot of subpopulations, represented by tree leaves. The lengths of the branches are arbitrary—because without further strong assumptions, we cannot infer branch length from these data. (*B*) Observed VAFs, major copies, and minor copies across samples. Matrix *Y* indicates whether an SNA resides in a CNA. (*C*) Matrix decomposition by Canopy. Genotyping matrix *Z* represents the positions of the SNAs in the phylogeny. $\tilde{C}^M$ and $\tilde{C}^m$ encode major and minor copy number of each clone. *H* specifies SNA–CNA phasing—whether SNAs reside in major or minor copies. Clonal frequency matrix *P* is shown as part of *A*.

the proportion of normal cells within any sample is simply the first entry in its mixture proportion vector. Multiple samples collected for the same tumor are represented by multiple horizontal slices of the phylogeny, each receiving its own vector of proportions.

The observed data are summarized in Fig. 1*B*. We let N be the number of samples, and S and T be the number of somatic SNAs and CNAs, respectively, that were identified across all samples. For SNAs, let the matrices $R \in \mathbb{R}^{S\times N}$ and $X \in \mathbb{R}^{S\times N}$ be, respectively, the number of reads containing the mutant allele and the total number of reads covering each of the S loci in each of the N samples. The ratio R/X is the proportion of reads supporting the mutant allele, known as the VAF. For CNAs, Canopy directly takes output from FALCON (28), FALCON-X, or other allele-specific copy number estimation methods (29). These outputs are in the form of estimated major and minor copy number ratios, denoted by $W^M \in \mathbb{R}^{T\times N}$ and $W^m \in \mathbb{R}^{T\times N}$, respectively, with their corresponding standard errors $\varepsilon^M \in \mathbb{R}^{T\times N}$ and $\varepsilon^m \in \mathbb{R}^{T\times N}$. See *Methods* for details regarding these quantities. For each SNA and each CNA, we also know whether they overlap. This information is represented by the matrix $Y \in \mathbb{R}^{S\times(T+1)}$: for column $j+1$, Y has 1's for SNAs that lie within CNA j and 0's for all other SNAs; as first column, Y has 1's for SNAs that do not reside in any CNAs and 0's otherwise (see example in Fig. 1*B*).

Each sample contains a mixture of the clones that comprise the tumor, and thus these observed VAFs and copy number ratios rely on the mixture proportions as well as the genomic profiles of the clones, as embodied by the underlying phylogenetic tree that is shared across all samples collected for the same tumor.

Relationship to Existing Work. Many existing studies of tumor evolution by multiregion or multi-time point bulk tumor DNA sequencing rely on laborious manual history reconstruction (7, 8). There have been much recent progress in the development of computational approaches for the analysis of such data. These approaches differ in the types of mutations that are modeled and the assumptions that are made. The main differences are summarized in Table 1 and discussed below.

TITAN (17) and THetA (18) focus on estimating cell population structure and recovering clonal evolutionary history for the case where somatic CNAs and loss of heterozygosity (LOH) distinguish subpopulations. These methods use allelic read coverage at germline heterozygous SNP loci to distinguish clonal versus subclonal CNA events. They ignore SNAs and do not pool data across multiple samples from the same tumor.

Many programs focus specifically on SNAs. For example, SciClone (16) clusters the VAFs of SNAs in copy-number neutral and LOH-free portions of the genome using a Bayesian beta mixture model. Pyclone (15) is an extension of SciClone that adds prior information elicited from copy number estimates obtained from either genotyping arrays or WGS to its Bayesian nonparametric clustering method. Neither SciClone (16) nor Pyclone (15) infers the phylogenetic relationship between subclones. LICHeE (23) and SCHISM (24) take VAFs of SNAs as input and construct a phylogenetic tree via an acyclic directed graph. Clomial (25), another program designed exclusively for SNAs, performs mixture deconvolution assuming that all mutational loci are heterozygous from copy number-neutral regions. Clomial decomposes the VAF matrix into a product of sample proportions and population genotypes, and uses expectation maximization (EM) to estimate both matrices.

ABSOLUTE (13) was the first software to infer subclonal heterozygosity from both SNAs and CNAs. However, taking data from only one sample, it determines whether each event is clonal or subclonal, but does not attempt to genotype or quantify the underlying subclones. In a similar fashion, Lönnstedt et al. (32) took a two-step approach using both SNA and CNA input, first estimating CNAs and then comparing VAF of SNA to its local copy number estimate to classify the somatic point mutation as clonal or subclonal. Recent approaches such as BitPhylogeny (26) and PhyloSub (33) detect major subclonal lineages by sampling the subclonal proportions via a tree-structured stick-breaking (TSSB) process, adjusting for overlapping CNAs. BitPhylogeny further adapts the nonparametric Bayesian mixture model to DNA methylation data from multiple microdissections from different regions of the same tumor.

As mentioned earlier, the VAF, which quantifies the proportion of alleles in the sample carrying a somatic mutation in the sample's DNA pool, is not the same as the proportion of cells in the sample carrying the somatic mutation. We call the latter, which is not directly observed in sequencing data, the mutant cell frequency (MCF). A similar quantity that is sometimes used in literature is cancer cell fraction (CCF), which is the proportion of cells among all cancer cells carrying the mutation. Given the tumor purity ϕ_C, $\mathrm{MCF} = \mathrm{CCF} \times \phi_C$. The MCF of a mutation directly reflects the total contributing proportion of the clone(s) that carry it, but to compute MCF from VAF, one needs to compensate for any CNAs that affect the locus. The existing methods differ by how this compensation is done. ABSOLUTE (13), EXPANDS (14), Pyclone (15), and PhyloSub (33) assume that, when a CNA event overlaps a SNA, the point mutation resides in a region with homogeneous aneuploidy, a scenario where no subclonal CNA events are allowed. Li and Li (19) conducted a detailed analysis of the complete set of scenarios covering the possible order and phase of overlapping SNAs and CNAs in developing their software CHAT. However, CHAT

Table 1. Properties and assumptions of cancer clonal phylogeny reconstruction methods

	TITAN	SciClone			SCHISM		PhyloWGS	
					PhyloSub			
Property/assumption	THetA	Clomial	PyClone	LICHeE	BitPhylogeny	CHAT	SPRUCE	Canopy
Takes raw SNAs calls as input	N	Y	Y	Y	Y	Y	Y	Y
Takes raw copy number estimates as input	Y	N	Y	N	Y	Y	N	Y
Allows CNAs to be subclonal	Y	N	N	N	N	Y	Y	Y
Resolves SNA and CNA that overlap	N	N	N	N	N	Y	Y	Y
Resolves overlapping CNAs with different endpoints	N	N	N	N	N	N	N	Y
Pools information across samples	N	Y	Y	Y	Y	N	Y	Y
Pools information across sites	Y	Y	Y	Y	Y	N	Y	Y
Reconstructs phylogeny	N	N	N	Y	Y	N	Y	Y
Quantifies uncertainty in phylogeny	N	N	N/A	N	Y	N/A	Y	Y

TITAN (17); THetA (18); SciClone (16); Clomial (25); PyClone (15); LICHeE (23); SCHISM (24); PhyloSub (33) BitPhylogeny (26); CHAT (19); PhyloWGS (20); SPRUCE (27). N, no; N/A, not applicable; Y, yes.

does not pool information across sites or across samples. PhyloWGS (20) also conducts a detailed breakdown of the possible configurations of overlapping SNA and CNA and is the first method to integrate both types of mutations when reconstructing cancer phylogenies using a TSSB. However, for each CNA region, PhyloWGS requires as input the integer absolute copy number of each allele and treats CNA events as pseudo-SNA events to compute its MCF. Because knowing integer-valued copy number is akin to knowing its MCF, in essence, in essence, PhyloWGS requires a two-step procedure where the underlying clones are first identified with their absolute copy numbers estimated using CNA data only, and this information is then used to compute the MCF of SNAs. SPRUCE (27) is another recent method that analyzes both SNAs with CNAs and characterizes the tumor phylogeny as a restricted class of spanning trees. Like PhyloWGS, SPRUCE takes processed CNA calls, for example, from THetA, and assumes known MCF for CNA events. Unlike these existing approaches, Canopy takes as input raw copy number ratios estimated by existing segmentation programs, and uses SNAs and CNAs to jointly infer the underlying clones and their evolutionary history. Because the same clonal admixture underlie CNAs and SNAs, this integrated approach achieves more accurate estimates in complex scenarios, as we illustrate later through examples.

As with all phylogenetic inference, assumptions are needed to resolve ambiguity. The perfect phylogeny model (23, 34) assumes that all subclones share the same phylogenetic tree and that mutations do not recur independently in different subclones. That is, each mutation appears only once and once it appears, it does not revert back to its original state. This no-homoplasy assumption, also referred to as the infinite-sites assumption (20, 35), is adopted by most methods to allow model identifiability. For example, it is possible to assert that under the infinite-sites assumption, mutations with lower CCFs cannot be ancestral to mutations with higher CCFs. To deal with copy number changes, El-Kebir et al. (27) proposed instead an infinite-alleles assumption, or the multistate perfect phylogeny, where a mutation may change state more than once on the tree due to gain or loss of copy number, but changes to the same state at most once. Furthermore, Deshwar et al. (20) introduces the "weak parsimony" assumption, which posits that mutations with similar CCFs across all samples lie on the same branch segment in the phylogeny. Canopy relies on both the infinite-sites assumption and the weak parsimony assumption, but takes a different approach from El-Kebir et al. (27) in modeling CNAs: Canopy extends the infinite-sites assumption to CNAs by assuming that copy number events with the same breakpoints and the same copy number across all samples must be the result of a single CNA event that occurs exactly once in the tumor's evolution. CNAs that overlap but have different breakpoints or different copy number states are treated as separate events. For example, a homozygous deletion nested within a heterozygous deletion, or a series of nested amplifications, are treated as separate events rather than separate alleles of the same mutation. This assumption allows Canopy to, with appropriate data, resolve the evolutionary relationship between overlapping copy number events, as we show in the whole-exome study of breast cancer cell line MDA-MB-231.

Matrix Representation of a Tumor's Clonal Composition. We use K to denote the total number of clones of the tumor that have representation among the cells in our sample(s). As shown in Fig. 1*A*, the tumor's evolutionary history is denoted by τ_K, a bifurcating tree with K leaves and with point mutation and copy number events assigned to its branch segments. Any τ_K gives us three matrices reflecting the mutation profiles of the underlying clones, shown in Fig. 1*C*: The SNA genotypes $Z \in \mathbb{R}^{S\times K}$, where Z_{sk} is the indicator of whether the sth SNA is present at the kth clone, and the major and minor copy numbers $\tilde{C}^M \in \mathbb{R}^{T\times K}$ and $\tilde{C}^m \in \mathbb{R}^{T\times K}$, where $\tilde{C}^M_{tk}$ and $\tilde{C}^m_{tk}$ are integer-valued major and minor copy numbers of the tth CNA in the kth clone. Phylogenetic restrictions are imposed by Canopy in that there is a one-to-one mapping between the positions of SNAs and CNAs on the tree as well as the major and minor copies of CNA events and the matrix Z, $\tilde{C}^M$, and $\tilde{C}^m$. Furthermore, because the left-most clone in the tree represents the normal cells, the first column of Z contains all zeros and the first columns of $\tilde{C}^M$ and $\tilde{C}^m$ contain all 1's. We do not directly observe the clones; instead, the samples we sequence are mixtures. We define $P \in \mathbb{R}^{K\times N}$ as the clonal frequency matrix, where P_{kj} $(1 \le k \le K, 1 \le j \le N)$ is the fraction of cells in the jth sample that belong to the kth clone (P shown in Fig. 1*A* is transposed to be aligned to the phylogeny). Each column of P sums up to 1 with the first row corresponding to the normal cell contamination. The matrices Z, $\tilde{C}^M$, $\tilde{C}^m$, and P are all unobserved, as well as the number of clones K. Our goal is to estimate them from the observed data, that is, the VAFs and the major and minor copy number ratios.

SNA–CNA Phase and Combined Likelihood. Here, we derive the likelihood for the data, given the model parameters $\{Z, \tilde{C}^M, \tilde{C}^m, P, K\}$. First, consider the CNA events. For CNAs, multiplication (denoted by $\times$) of the clonal integer copy number matrices $(\tilde{C}^M, \tilde{C}^m)$ and the sample proportion matrix (P) gives us the continuous-valued major and minor copy numbers (C^M and C^m) for each sample:

$$\begin{aligned}\tilde{C}^M \times P = C^M \in \mathbb{R}^{T\times N}\\ \tilde{C}^m \times P = C^m \in \mathbb{R}^{T\times N}.\end{aligned}$$

Because the observed copy number ratios are usually computed by averaging over a large number of loci (for microarrays), exons (for WES), or bins (for WGS), we assume that they are normally distributed with the given standard errors, that is,

$$\begin{aligned}W^M \sim N\left(C^M, \left(\varepsilon^M\right)^2\right)\\ W^m \sim N\left(C^m, \left(\varepsilon^m\right)^2\right).\end{aligned}$$

For SNAs, $Z \times P$ gives the MCF of each SNA in each sample, which we denote by the matrix $\mathrm{MCF} \in \mathbb{R}^{S\times N}$. The observed number of mutant reads R_{sj} follows a binomial distribution with total count X_{sj} and probability of success being the VAF, which we denote by VAF_{sj} $(1 \le s \le S,\ 1 \le j \le N)$. That is,

$$R_{sj} \sim \mathrm{Binomial}\left(X_{sj}, \mathrm{VAF}_{sj}\right).$$

Therefore, we need to convert MCF to VAF to calculate the binomial likelihood for SNAs.

If all SNAs are heterozygous from copy number neutral regions, as assumed by SciClone (16) and Clomial (25), then $\mathrm{VAF} = 1/2 \times \mathrm{MCF} = 1/2 \times \mathrm{CCF} \times \phi_C$, where ϕ_C is the cancer cell purity, MCF is the fraction of cells that have the SNA, and CCF is the fraction of cancer cell that have the mutation. Pyclone (15), PhyloSub (33), and EXPANDS (14) account for CNAs but make the assumption that was first introduced by ABSOLUTE (13), namely, that there are no subclonal CNA events. Therefore,

$$\begin{aligned}\mathrm{VAF} &= \frac{C_{\mathrm{mut}}}{2\times(1-\phi_c)+C_{\mathrm{total}}\times\phi_C}\mathrm{MCF}\\ &= \frac{C_{\mathrm{mut}}\times\phi_C}{2\times(1-\phi_c)+C_{\mathrm{total}}\times\phi_C}\mathrm{CCF},\end{aligned}$$

where ϕ_C is the purity of cancer cells, which have a homogeneous CNA state with total copy number C_{total} and mutant-allele copy number C_{mut}.

To more accurately quantify the relationship between VAF and MCF, which accounts for the possible phases and temporal orders of overlapping CNAs and SNAs, we consider separately each of the three possible underlying scenarios, which were first delineated by CHAT (19) and PhyloWGS (20): (*i*) the CNA is ancestral to the SNA (Fig. 2*A*); (*ii*) the CNA and SNA occur in separate branches of the tree and thus affect separate clones (Fig. 2*B*); (*iii*) the SNA is ancestral to the CNA (Fig. 2*C*). To compute the VAF, we separately calculate the numerator (copy number of the affected allele at the mutational locus), and the denominator (total copy number at the locus) for each SNA across all samples. The denominator can be generalized and is the same for all three cases, being simply the total copy number at the SNA locus. Therefore, the denominator is, in matrix notation, as follows:

$$\left(Y\times\begin{bmatrix}\mathbb{1}\\ \cdots\\ \tilde{C}^{M}\end{bmatrix}+Y\times\begin{bmatrix}\mathbb{1}\\ \cdots\\ \tilde{C}^{m}\end{bmatrix}\right)\times P\in\mathbb{R}^{S\times N}$$

$$=\begin{array}{c}\\ \text{SNA1}\\ \text{SNA2}\\ \text{SNA3}\\ \text{SNA4}\end{array}\begin{array}{c}\begin{array}{ccc}\text{Sample 1} & \text{Sample 2} & \text{Sample 3}\end{array}\\ \begin{bmatrix}2 & 2 & 2\\ 1.8 & 1.8 & 1.7\\ 2.6 & 2.5 & 2.5\\ 2 & 2 & 2\end{bmatrix}\end{array},$$

where $\mathbb{1}$ is a vector of ones augmented to the first row of $\tilde{C}^{M}$ and $\tilde{C}^{m}$ representing the major and minor copy number for the "non-CNA" SNA loci. The numerator differs across the three cases.

Case 1: The CNA is ancestral to the SNA. Only one allele of the locus is affected (e.g., SNA1 in Fig. 2*A*). Therefore, the copy number of the affected allele for SNA s in each clone is $Z_{s,:}\in\mathbb{R}^{1\times K}$ (the s row of the Z matrix). The numerator, which is the copy number of the affected allele in each sample, is thus the matrix product of $Z_{s,:}$ and P. For SNA1 in Fig. 2*A*, this evaluates to the following:

$$(Z_{1,:}\times P)=\text{SNA1}\begin{array}{c}\begin{array}{ccc}\text{Sample 1} & \text{Sample 2} & \text{Sample 3}\end{array}\\ \begin{bmatrix}0.3 & 0.5 & 0\end{bmatrix}\end{array}.$$

Note that the numerator in this case is the same as the row corresponding to the SNA in the MCF matrix ($\text{MCF}=Z\times P$) because each variant cell has only one variant allele.

Case 2: The CNA and SNA occur in two nonoverlapping lineages. The SNA is not affected by the CNA, which lies on a different branch of the tree (e.g., SNA2 in Fig. 2*B*). Therefore, the numerator is the same as is in the previous case. For SNA2 in Fig. 2*B*, this evaluates to the following:

$$(Z_{2,:}\times P)=\text{SNA2}\begin{array}{c}\begin{array}{ccc}\text{Sample 1} & \text{Sample 2} & \text{Sample 3}\end{array}\\ \begin{bmatrix}0.3 & 0 & 0.5\end{bmatrix}\end{array}.$$

Case 3: The SNA is ancestral to the CNA. This is usually the most interesting case. For example, copy number loss or LOH following an SNA may delete the normal allele, or copy number gain may amplify the mutated allele (e.g., SNA3 in Fig. 2*C*)—both scenarios having potential phenotypic consequences. Of the two underlying alleles at the SNA locus, we need to distinguish which has been lost and/or gained. That is, if the CNA confers allelic imbalance, we need to distinguish whether the major or the minor allele is the mutated allele. For SNA3 in Fig. 2*C*, the major allele has the SNA and the copy number of the mutant allele in each sample, that is, the numerator, is as follows:

$$\left\{Z_{3,:}\cdot\left(Y_{3,:}\times\begin{bmatrix}\mathbb{1}\\ \cdots\\ \tilde{C}^{M}\end{bmatrix}\right)\right\}\times P\in\mathbb{R}^{1\times N}$$

$$=\text{SNA3}\begin{array}{c}\begin{array}{ccc}\text{Sample 1} & \text{Sample 2} & \text{Sample 3}\end{array}\\ \begin{bmatrix}1.4 & 1.2 & 1.3\end{bmatrix}\end{array}.$$

Here, we define the notation $\cdot$ as elementwise matrix multiplication. If the mutant allele lands on the minor copy, the numerator is simply the above with $\tilde{C}^{M}$ replaced by $\tilde{C}^{m}$.

A general formula for the numerator encompassing all three cases is derived in *Methods*. Division of the numerator by the denominator gives us the VAF, and the likelihood can then be expressed as follows:

$$L\left(Z,P,\tilde{C}^{M},\tilde{C}^{m},\tilde{H},\tilde{Q},\tau_K|W^{M},W^{m},\varepsilon^{M},\varepsilon^{m},R,X,Y\right)$$

$$=\prod_{j=1}^{N}\prod_{s=1}^{S}\prod_{t=1}^{T}\left\{\text{pNorm}\left(W_{tj}^{M},\left(\tilde{C}^{M}\times P\right)_{tj},\left(\varepsilon_{tj}^{M}\right)^{2}\right)\right.$$

$$\left.\text{pNorm}\left(W_{tj}^{m},\left(\tilde{C}^{m}\times P\right)_{tj},\left(\varepsilon_{tj}^{m}\right)^{2}\right)\text{pBinomial}\left(VAF_{sj},R_{sj},X_{sj}\right)\right\},$$

where $\text{pNorm}(x,\mu,\sigma^2)$ is the likelihood for observing x from a Gaussian distribution with mean μ and standard deviation σ, $\text{pBinomial}(p,R,X)$ is the binomial likelihood for observing R successes from X trials with success probability p, $\tilde{H}$ indicates the phasing of the SNAs with overlapping CNAs (whether an SNA precedes a CNA), $\tilde{Q}$ is a vector of the ordering of the SNA–CNA pair that can be directly obtained from the tree, and VAF is derived in *Methods*.

For cases where nested CNAs are observed, Canopy samples the temporal and spatial orders of the CNAs together with the affected SNAs in the phylogenic tree (*SI Appendix*, *SI Methods* and Figs. S1 and S2). Resolving overlapping and nested CNA events is not trivial, because in real dataset analysis we only

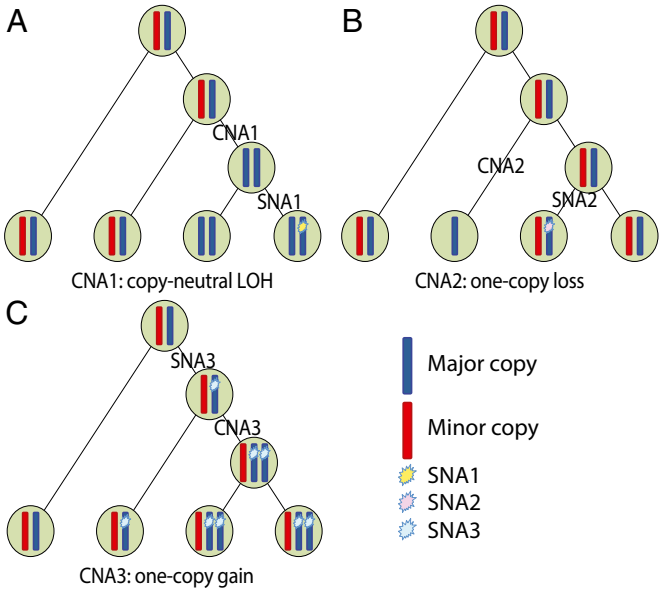


Fig. 2. Three cases of SNA–CNA phase and order. Different phases and orders of CNA and the SNA it affects are shown with clonal histories concordant with Fig. 1. Major and minor copies are in blue and red, respectively; SNA mutational loci are shown as stars. (*A*) CNA precedes SNA. SNA resides in only one chromosomal copy. (*B*) CNA and SNA are on two separate branches. SNA is unaffected by CNA. (*C*) SNA precedes CNA. Scenario where major copies contain the SNA is shown. SNA4 from Fig. 1 is unaffected by CNA and is not shown.

observe the major and minor allelic ratio per region per sample. By overlapping CNAs, we are referring to distinct CNA events occurring in separate samples that affect the same genomic region, more specifically, overlapping across samples (e.g., CNA event E_1 and E_2 in *SI Appendix*, Fig. S3*B*); by nested CNAs, on the other hand, we are referring to CNAs that may occur in different samples or within the same sample (e.g., CNA event E_2 and E_3 in *SI Appendix*, Fig. S3*B*). Canopy can resolve CNA events, overlapping or nested, that have representation in the data. For such events, manual inspection of the segmentation input is sometimes helpful to identify nested CNAs within the same sample and to verify the type (gain, loss, or copy-neutral LOH) of each event.

As a concrete example, *SI Appendix*, Fig. S13*B* shows the allelic ratio input for CNA events from chr12 covering the *KRAS* locus. One knows by looking at the cross-sectional data that there are two separate CNA events, one whole-chromosomal–level duplication and another p-arm LOH event. These two events have different breakpoints, attain different allele-specific copy numbers, and are thus treated as two separate events. To distinguish these two events, the user has to manually inspect the breakpoints.

We use Bayesian information criterion (BIC) as a model selection method to determine the number of subclones K and design a Metropolis Hastings algorithm to sample the posterior distribution of the unknowns and enumerate all plausible histories in the tree space:

i) randomly switch a CNA or SNA to another branch on the tree;
ii) randomly select at least two clones and change their clonal frequencies;
iii) randomly select a neighborhood for local rearrangement to generate a new tree topology (*SI Appendix*, Fig. S4);
iv) randomly select a CNA and sample its major and minor copy number from {0,1,2,3} and update $\tilde{C}^M_{tk}$ and $\tilde{C}^m_{tk}$, $1 \le t \le T$, $1 \le k \le K$;
v) for SNA that resides in a CNA ($Q_s = 1$), randomly sample whether the major or the minor allele contains the SNA after the copy number change, that is, randomly sample the indicator random variable H_s $(1 \le s \le S)$.

For each run, we start with multiple chains from different start points and evaluate convergence by likelihood and acceptance rate. Posterior distribution is marginalized after combining different chains, burn-in, and thinning. When multiple posterior "modes" exist, Canopy attempts to return all phylogenies that the data support and computes the relative confidence interval in each clonal history. Quantities that can be marginalized from the posterior distributions are obtained from subtree space with trees having the same clonal and mutational compositions.

Simulation Studies. As a simple illustration, we first show how Canopy successfully identifies the subclones and recovers the phylogeny for the scenario shown in Fig. 1 (*SI Appendix*, Fig. S5), which is a simple configuration that is as typical as any other given the level of complexity. We also use this example to demonstrate the differences between Canopy and two related methods, PhyloWGS and Clomial. To generate suitable input for PhyloWGS, we converted the CNA events to pseudo-SNA events, because in this toy example we have at our disposal the true clonal proportions as well as the true SNA–CNA phasing (*SI Appendix*, Table S2), and thus simply used these true values as if it were known. Refer to *Methods* for details on simulation setup. Canopy, starting with raw CNA estimates and assuming unknown phase, returns a tree highly concordant with the ground truth; whereas PhyloWGS, even using the true phase and clonal proportions for CNAs, returns a linear tree with incorrectly inferred cellular frequencies (*SI Appendix*, Fig. S5*C*). We further introduce scenarios where CNAs overlap (*SI Appendix*, Figs. S1 and S2) and show that Canopy can successfully handle a fair amount of complexity (*SI Appendix*, Fig. S6). As a comparison, Clomial (25), which ignores the existence of CNAs, fails to correctly estimate the clonal frequencies and infers incorrect tumor purities (*SI Appendix*, Figs. S5*B* and S6*B*). *SI Appendix*, Fig. S5*D* also explores the effect of CNA estimation noise on deconvolution accuracy.

We then performed simulation studies to explore the effects of various parameters on estimation accuracy as well as computation time, and evaluate performance benchmarked against existing methods. We use the percentage of wrongly labeled Z elements (Fig. 3) and the root-mean-square error (RMSE) of the P matrix (*SI Appendix*, Fig. S7) as a measure of the deconvolution accuracy and compare Canopy's results with those returned by Clomial (25). We use clustering purity as a measure of clustering quality and compare the preclustering results of Canopy with those of SciClone (16). We sampled systematically from a comprehensive set of possible phylogenies and P matrices. More details on simulation setup are in *Methods*.

We ran simulations with varying number of mutations from two different sequencing pipelines: WGS with $d = 30$ (Fig. 3*A*) and targeted sequencing with $d = 500$ (Fig. 3*B*), where d is the mean sequencing depth. The results, compiled in Fig. 3 and *SI Appendix*, Figs. S7–S9, give a sense of how estimation accuracy depends on the number of informative mutations, the number of genotypically distinct samples, the sequencing depth, and the number of underlying clones. As expected, estimation accuracy increases with the number of genotypically distinct samples, the number of informative mutations, and the sequencing depth. Increasing the number of subclones makes the estimation problem harder, although this can be compensated for by a larger number of mutations. Also, in *SI Appendix*, Fig. S8, we show that the larger the difference in clonal proportions between the samples, the easier the estimation problem. Under all simulated scenarios, Canopy is as good as or better than Clomial (25) and SciClone (16) in terms of deconvolution and clustering

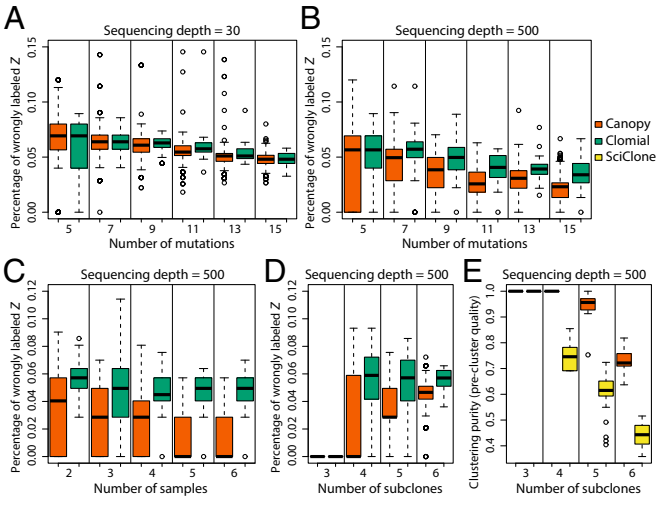


Fig. 3. Deconvolution accuracy and clustering quality via simulation studies. Various parameters show effects on deconvolution accuracy (measured by the percentage of wrongly labeled Z elements) and preclustering quality (measured by the clustering purity; *Methods*). Corresponding RMSE of the P matrix is shown in *SI Appendix*, Fig. S7. Canopy is compared against Clomial and SciClone and is shown to have better performance. (*A* and *B*) WGS compensates its low sequencing depth with more profiled mutations. (*C*) Increasing sample size helps solve reconstruction ambiguity. (*D* and *E*) Number of subclones is negatively correlated with deconvolution accuracy and preclustering quality.

accuracy (Fig. 3 and *SI Appendix*, Fig. S7). An interesting observation from the simulation studies is that, although increasing the number of samples drives the estimation error of Z to zero, the benefit of including more mutations diminishes when there is a small number of underlying subclones. In this case, only a small high confidence set of informative mutations or mutation clusters is sufficient for recovering the underlying tree (*SI Appendix*, Table S3); when the number of underlying subclones is large, more mutations are needed (*SI Appendix*, Fig. S9).

We also performed simulations to investigate the effect of the proposed binomial mixture clustering method with varying number of mutations and clones. This preclustering procedure serves as an initialization step in the MCMC sampling, where mutations are first moved along tree branches in clusters and then fine-tuned individually in the later rounds. We show that this initialization method significantly reduces computation time, offers a way to clean up data by including a uniform noise component, and has similar or better deconvolution accuracy (*SI Appendix*, Table S3 and Fig. S10).

Application to Transplantable Metastasis Model Derived from MDA-MB-231. Canopy is applied to a transplantable metastasis model system derived from a heterogeneous human breast cancer cell line MDA-MB-231. Cancer cells from the parental line MDA-MB-231 were engrafted into mouse hosts leading to organ-specific metastasis. Single-cell populations (SCPs) or mixed-cell populations (MCPs) were in vivo selected from either bone or lung metastasis and grew into phenotypically stable and metastatically competent cancer cell lines (Fig. 4*A* and *SI Appendix*, Table S5).

This transplantable model system has been widely used for understanding metastatic progression (36–38). Minn et al. (36) identified a "poor-prognosis" gene expression signature for distinct metastatic potential by studying patterns of transcriptomic profile. Recently, Jacob et al. (37) performed WES on a metastasis model derived from the same parental line MDA-MB-231 and found that in vivo selected highly metastatic cell populations showed little genetic divergence from the corresponding parental population. Their results suggest that (*i*) genetic variations (including the genes $BRAF^{G464V}$ and $KRAS^{G13D}$, validated by Sanger sequencing) preexist in the parental line and are enriched with increased metastatic capability; (*ii*) metastatic competence during tumorigenesis can emerge with selection of preexisting oncogenic alleles without a need of new mutations (37).

Here, we build a transplantable model from MDA-MB-231, where the parental line as well as the SCP and MCP samples are whole-exome sequenced and are used to investigate clonal evolution associated with metastatic progression on the DNA level. We only use the parental line and the MCP samples to infer metastatic phylogeny, whereas the SCP samples are included as a validation dataset to compare and contrast. Because SCP samples are homogeneous cell populations, their integer absolute copy numbers can be inferred by a hidden Markov model. Because we do not have a normal control for MDA-MB-231, the integer absolute copy numbers for the SCP samples are used as controls to infer copy number ratios in the MCP samples (*SI Appendix*, *SI Methods*, Figs. S11 and S12, and Table S6). SNAs and indels are called by the UnifiedGenotyper in the Genome Analysis Toolkit (GATK) (30) and are further annotated by ANNOVAR (39).

In addition to the oncogenic point mutations in *BRAF* and *KRAS* reported by Jacob et al. (37), our analysis pipeline identified two nonsynonymous mutations in the genes ALPK2 and RYR1 that are deleterious by functional annotation (Dataset S1). VAFs of the four mutations vary between bone and lung metastasis samples: *BRAF* and *KRAS* mutations are enriched in the bone samples; *ALPK2* mutation is enriched in the lung samples; *RYR1* mutation is additionally detectable in the bone samples but not in the parental line (Fig. 4*B*). These four mutations also overlap with six CNAs events, with regions in chromosome (chr) 7q and 12 being double "hit" by two nonidentical overlapping CNA events in separate samples (*SI Appendix*, Fig. S13).

The a posteriori most likely phylogenetic tree inferred by Canopy using the parental line and the MCP samples only has four subclones guided by BIC (Fig. 4*C*) and is shown in Fig. 4*D*. As expected, our results show that the bone and lung metastatic sublines contain clones that were either nonexistent or extremely rare in the parental line, which proliferated to dominate the organ-specific metastasis. All samples, except MCP2287, are almost 100% composed of cells from a single clone. Clone 2 is unique to the lung subline, and clone 3 is unique to the brain subline (Fig. 4*D*). MCP2287 partially retains the parental line and is a mixture of two subclones, which, upon detailed visual inspection, is supported by the raw SNA and CNA input (Fig. 4*B* and *SI Appendix*, Fig. S13 *A* and *B*). For CNAs, Canopy successfully resolves overlapping CNAs with correctly inferred copy number states (*SI Appendix*, Fig. S13); among the SNAs, *BRAF*, *KRAS*, and *ALPK2* each undergo a duplication event that amplifies the mutant allele, with *BRAF* and *KRAS* further losing the reference allele via a second LOH event (Fig. 4*D*) that occurs later in the evolutionary process. All sublines share chr 12 duplication, whereas the bone and lung sublines gain additional mutations that mark and/or drive their divergence (Fig. 4*D*).

Canopy's inferred phylogeny is confirmed by the SCP samples, which we use as validation. The two SCP samples derived from the lung metastasis are 100% identical to clone 2, and the two derived from the bone metastasis are 100% identical to clone 3

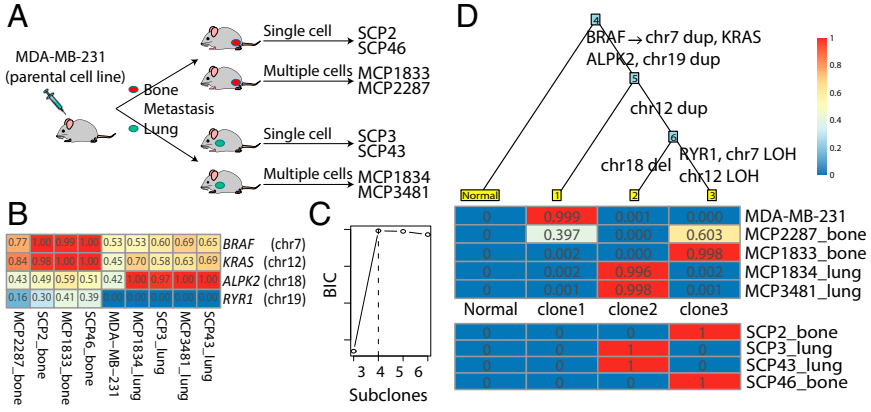


Fig. 4. Clonal history of transplantable metastasis model MDA-MB-231 with validation by SCP samples. (*A*) Transplantable model system of MDA-MB-231. Parental line is injected into mouse models and induces organ-specific metastasis. Sublines are derived from single or multiple cell(s) from different metastatic sites. (*B*) Observed VAFs of somatic SNAs, which reside in nested CNAs. CNA input is shown in *SI Appendix*, Fig. S13. Canopy takes both SNA and CNA input. (*C*) BIC as a model selection method to determine the number of subclones. (*D*) Clonal tree reconstructed by Canopy. Organ-specific subclones (clone 2 and 3) acquire additional mutations from the parental clone (clone 1) and dominate the metastasis. SCP samples successfully validate the subclones and confirm Canopy's inferred phylogeny.

(Fig. 4*D*). Similar to Jacob et al. (37), Canopy's inferred phylogeny shows that amplification of oncogenic signals preexisting in the parent cell line (*KRAS*, *BRAF*, and *ALPK2*) leads to higher tumor-initiating fitness. Nevertheless, in contradiction to the proposed model by Jacob et al. (37) where no new mutations are needed, here we report additionally acquired/detectable SNA and CNAs as DNA signatures that mark and/or drive the divergence between the lung and bone sublines. These mutation signatures—chr 18q deletion, *RYR1* point mutation, and chr 7q and 12 LOH—can indicate breast cancer metastatic potentials and serve as prognostic markers for the development of distant metastasis.

Application to Breast Cancer Patient Xenografts. We further applied Canopy to a deep-genome sequencing dataset of breast cancer patient xenografts from Eirew et al. (10). Xenografts of a patient line were generated by serially transplanting breast cancer tissue organoid suspensions into immunodeficient mice (10). WGS was performed on the initial engraftment (SA494T) and its subsequent propagation of metastatic xenograft (SA494X4). Targeted-amplicon deep sequencing was performed to validate somatic SNAs; TITAN (17) was applied to infer CNAs and LOH (10). We adopt bivariate clustering and stringent quality control procedures to remove experimental noise (Fig. 5*A*). Canopy takes as input SNAs from four clusters that are CNA-free, three SNAs that overlap with CNAs, and four CNAs (chr 1p, 3p, and 19p deletion and chr5q duplication) to reconstruct phylogeny (Dataset S2).

The number of subclones is chosen at four by BIC as a criterion for model selection (Fig. 5*B*). The most likely tree returned by Canopy is shown in Fig. 5*C*. Clone 2 and clone 3 (2% and 1% of the starting population, SA494T) undergo a one-copy loss event and additionally acquire SNAs in cluster 3, indicating extreme selective engraftment of minor clones (Fig. 5*C*). These two clones are further separated by SNAs in cluster 4 and become dominant in the subsequent metastatic xenograft SA494X4 with high prevalence (77% and 23% shown in Fig. 5*C*). For SNAs that overlap with CNAs, only SNA2 precedes its affecting CNA2 (one-copy loss) and has a higher mutational multiplicity after losing the healthy allele. Both SNA1 and SNA3 arise after one-copy loss events, resides in clone 1 only, and thus are present in sample SA494T but not in SA494X4.

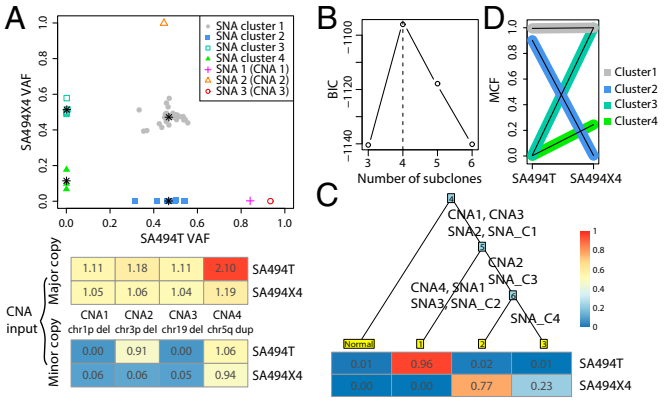


Fig. 5. Clonal architecture of breast cancer initial engraftment and passage xenograftment. Tumor sample SA494T and its subsequent xenograft SA494X4 are whole-genome sequenced with SNAs validated by deep amplicon resequencing and CNAs inferred by TITAN. (*A*) SNA and CNA input of Canopy. VAFs of four SNA clusters and three CNA-affected SNAs are shown in the top panel. Heat map of observed major and minor copy numbers are shown in the bottom panel. (*B*) BIC as a model selection metric to determine the number of subclones. (*C*) The most likely tree returned by Canopy based on the mutational profiling. Extreme selection of minor clones is imposed on engraftment. SA494T and SA494X4 bear two mutually exclusive sets of mutations in addition to shared ancestral mutations. (*D*) Mutation clusters inferred by the Pyclone model.

We compared our analysis result to the SNA clustering result achieved by Pyclone (15). SNA clusters 1–4 correspond to the four clusters inferred by Pyclone shown in Fig. 5*C*, which is expected because the SNAs within these clusters are CNA-free and these cell lines are expected to have no normal cell contamination (CCF = MCF = 2 × VAF). Although Pyclone outputs the clustering of these MCFs, Canopy also infers the evolutionary relationship between the clones represented by these clusters. Thus, from this analysis we can be quite confident that the mutations in cluster 3 are ancestral to the mutations cluster 4, that is, cells that carry the mutations in cluster 4 must also carry the mutations in cluster 3. Also, Pyclone uses CNA-corrected VAFs of SNAs as input, whereas Canopy uses both SNAs and CNAs simultaneously to infer tumor phylogeny. This allows us to infer the temporal order of the CNA events in relation to the SNA events. For example, we are quite confident that CNAs 1 and 3 are clonal events, whereas CNA 2 and CNA 4 came later affecting separate subclones.

Canopy's results are confirmed by single-cell sequencing carried out by Eirew et al. (10)—two mutually exclusive sets of mutant alleles from SA494 tumor and passage 4 xenograft, respectively, were identified in addition to a set of shared alleles (10).

Discussion

Intratumor heterogeneity contributes to drug resistance and failures of targeted therapies (40). To gain a comprehensive understanding of the evolutionary dynamics of tumors, it is important not only to determine which alterations drive the progression of a tumor but also to understand their relative temporal and spatial order during tumor evolution. Here, we propose a method, Canopy, to assess intratumor heterogeneity and infer clonal evolutionary history. The distinguishing features of Canopy compared with existing methods are as follows: (*i*) SNAs and CNAs are jointly modeled and overlapping events are phased and temporally ordered; (*ii*) the SNA input can be taken directly from the UnifiedGenotyper in GATK (30) or MuTect (31) and the CNA input are continuous-valued allele-specific copy number ratios, which can be directly obtained from allele-specific copy number estimation methods (28); (*iii*) a preclustering initialization step for SNAs improves robustness to noise and significantly reduces computation time; (*iv*) CNA events are allowed to be subclonal (19, 20); (*v*) overlapping and nested CNA events with different breakpoints affecting the same region are treated as separate evolutionary events, as illustrated by our analysis of MDA-MB-231; (*vi*) the Bayesian framework reconstructs the phylogeny together with posterior confidence assessment, which is useful when the data support multiple configurations.

Despite the fact that Canopy starts with a preclustering, an input that contains too many false detections can still lead to unreliable phylogeny inference. Most current CNA and SNA detection algorithms still have a high false-positive rate, and thus we suggest rigorous quality checking of input before a Canopy analysis. As we showed in our simulations, Canopy does not require a large set of variant loci to attain precise phylogeny inference; that is, the payoff for including multiple variants derived from the same clone quickly diminishes. A Canopy analysis should start with manual inspection and visualization of the input data, followed by removing short CNAs that may be unreliably called, and using the preclustering procedure with a multivariate uniform component on SNAs, as illustrated in our analysis of the data from Ding et al. (7) and Eirew et al. (10).

Canopy has been demonstrated on four cancer sequencing datasets of varying study design, as well as on extensive simulation data. On a whole-exome study of breast cancer cell line MDA-MB-231, Canopy successfully deconvolved the mixed cell sublines, identifying subclones that were validated by comparing

to single-cell sublines as ground truth. On a WGS dataset of the breast cancer tumor and its subsequent metastatic xenograft, Canopy's inferred clonal phylogeny is concordant with genomic markers of major clonal genotype and is confirmed by single-cell sequencing. On a WGS dataset of the primary tumor and relapse genome of a leukemia patient, and on a spatially sampled targeted sequencing study of ovarian cancer, Canopy predicted phylogenetic histories in concordance with existing knowledge (details in *SI Appendix*, *SI Results*). Finally, through simulations, we explored the effects of various parameters on deconvolution accuracy, and evaluate performance with comparison against existing methods. Collectively, Canopy provides a rigorous foundation for statistical inference on repeated sequencing data from evolving populations.

Many factors determine the accuracy of Canopy's results: higher sequencing depth allows for higher sensitivity for detection of rare subclones; more samples and more difference between samples in their clonal composition allow for higher accuracy in estimating the phylogeny. In particular, the maximum number of subclones that can be reliably inferred depends on all of these factors. As the number of subclones increase, the proportion of cells attributable to at least some of the subclones would necessarily decrease, and higher coverage would be needed to detect mutations present in those smaller subclones. A survey of recent multiregion and multi-time point cancer genome-sequencing studies shows that, even in scenarios where up to 11 bulk samples were analyzed from the same patient, the number of subclones identified was typically less than 8 (summarized in *SI Appendix*, Table S1). A similar range for the number of subclones was found by single-cell sequencing. To increase resolution for rare subclones, deeper sequencing or sequencing of a larger number of single cells is needed.

Most current cancer-sequencing studies sequence only one sample from each patient, from which it is difficult to deconvolve clonal mixtures. The recent advances in single-cell sequencing technologies make possible a different approach to study tissue heterogeneity at higher resolution. Nevertheless, reliable simultaneous profiling of copy number and single-nucleotide mutations by single-cell sequencing is still at infancy. Here, we show that traditional bulk sequencing can lead to accurate subclone identification and phylogenetic inference, if only the researcher is willing to sequence multiple slices of the tissue. Thus, bulk tissue sequencing can play an important part in our understanding of tumor heterogeneity, and in the coming years experimental designs that combine bulk tissue sampling and single-cell analysis needs to be better explored.

Methods

Allele-Specific Copy Number. For the tth ($1 \le t \le T$) CNA, we let N_t be the number of germline heterozygous loci within its segment [segmentation carried out by FALCON (28) or FALCON-X]. From FALCON's segmentation and phasing outputs, we can get for each tumor-normal pair the read counts of major and minor allele in the jth tumor slice, M_{ij} and m_{ij}, and in the matched normal sample, M_{i0} and m_{i0}, where $1 \le i \le N_t$ is the germline SNP index and $1 \le j \le N$ is the sample index.

For CNA events that are nonoverlapping (*SI Appendix*, Fig. S3*A*), we use the germline heterozygous loci within each CNA segment to compute major and minor copy number input across all samples:

$$W^M_{tj} = \frac{1}{N_t}\sum_{i=1}^{N_t}(M_{ij}/M_{i0}), \quad \left(\varepsilon^M_{tj}\right)^2 = \frac{\sum_{i=1}^{N_t}(M_{ij}/M_{i0})^2 - N_t\left(W^M_{tj}\right)^2}{N_t(N_t-1)};$$

$$W^m_{tj} = \frac{1}{N_t}\sum_{i=1}^{N_t}(m_{ij}/m_{i0}), \quad \left(\varepsilon^m_{tj}\right)^2 = \frac{\sum_{i=1}^{N_t}(m_{ij}/m_{i0})^2 - N_t\left(W^m_{tj}\right)^2}{N_t(N_t-1)}.$$

In the above, W^M_{tj}, W^m_{tj} are the estimates of the major and minor copy numbers, respectively, and ε^M_{tj}, ε^m_{tj} can be considered as their standard errors.

For CNA events that are overlapping or nested (*SI Appendix*, Fig. S3*B*), we propose an algorithm that automates the preprocessing of allele-specific copy number for input to Canopy (refer to *SI Appendix*, *SI Methods* for details). If external ploidy information is available, this can be added as a fixed CNA event (e.g., a genome doubling event for tetraploidy).

Generalization of VAF and MCF Relationship for All Three Cases. Here, we derive a general formula for the numerator encompassing all three cases. We denote $H \in \mathbb{R}^S$ as a vector of indicator of whether an SNA is from the major or the minor copy of the CNA that affects it and occurs after it. We further define Q as a vector indicating whether an SNA precedes the CNA it resides in, which can be directly obtained from the tree τ_K. Let $\tilde{H} = [H', H', \ldots, H']_K \in \mathbb{R}^{S\times K}$ and $\tilde{Q} = [Q', Q', \ldots, Q']_K \in \mathbb{R}^{S\times K}$. Then, the numerator for all three cases shown in Fig. 2 can be generalized and division of the numerator by the denominator gives us the VAF matrix:

$$\mathrm{VAF} = \frac{\left\{Z\cdot\left(Y\times\begin{bmatrix}\mathbb{1}\\ \cdots\\ \tilde{C}^M\end{bmatrix}\right)^{\tilde{Q}}\cdot\tilde{H} + Z\cdot\left(Y\times\begin{bmatrix}\mathbb{1}\\ \cdots\\ \tilde{C}^m\end{bmatrix}\right)^{\tilde{Q}}\cdot(1-\tilde{H})\right\}\times P}{\left(Y\times\begin{bmatrix}\mathbb{1}\\ \cdots\\ \tilde{C}^M\end{bmatrix} + Y\times\begin{bmatrix}\mathbb{1}\\ \cdots\\ \tilde{C}^m\end{bmatrix}\right)\times P} \in \mathbb{R}^{S\times N}.$$

Note that the exponentiation and division are carried out in an elementwise fashion and that 0^0 is defined to be equal to 1. This generalized matrix representation form to get VAFs of SNAs only apply to SNAs that are CNA-free or those that are affected by a single CNA event. For SNAs that are affected by more than one CNA event, VAFs are obtained iteratively for each SNA with adjustment of the affecting CNA events that are overlapping or nested.

Simulation Setup. We first generate input data from the true underlying tree as is illustrated in Fig. 1 (with nonoverlapping CNAs) and *SI Appendix*, Figs. S1 and S2 (with overlapping CNAs) and apply Canopy to reconstruct the phylogeny. For SNAs, the total read depth matrix X has each of its column sampled from a multinomial distribution:

$$X_{\cdot j} \sim \mathrm{Multinomial}\left(d\times S, \frac{1}{S}, \ldots, \frac{1}{S}\right),$$

where d is the mean sequencing depth and $1 \le j \le N$. The mutant read depth matrix R is sampled from a binomial distribution indexed at X with success probabilities VAF derived in SNA–CNA phase and combined likelihood section (numerator divided by denominator). For CNAs, the input matrix W^M and W^m are sampled from a normal distribution with mean $\tilde{C}^M \times P$ and $\tilde{C}^m \times P$ and SD ε^M and ε^m ranging from 0.001 to 0.64 (*SI Appendix*, Fig. S5*D*). The matrices X, R, W^M, W^m, ε^M, and ε^m are then used as input for Canopy to infer phylogeny with output shown in *SI Appendix*, Figs. S5*A* and S6*A*. For Clomial (25), we keep its assumptions and use X and R as input to infer phylogeny with result shown in *SI Appendix*, Fig. S5*B* and S6*B*.

We then separately investigate the effects of the number of mutations, the sequencing depth, the number of samples, the number of subclones, and the preclustering procedure as an initialization step on deconvolution and preclustering accuracy and computation time. Without loss of generality, we focus on using SNAs to reconstruct phylogeny and compare against two existing methods, Clomial (25) and SciClone (16). For each investigation, we control for confounding parameters, run 30 simulations in parallel, and integrate results from each run. Within each simulation, we run 10 Markov chains with random starts and correspondingly choose $binomTryNum = 10$ for Clomial (25), a parameter specifying the number of random starts for the EM algorithm. The true clonal frequency matrix P is prefixed but varies between different runs with a perturbation added to each of its element from a Gaussian distribution with mean of 0 and SD of 0.01. The generated matrix is then scaled so that each element is nonnegative and that the columns sum up to 1. We calculate the percentage of wrongly labeled elements in Z (Fig. 3) and the RMSE of the inferred P matrix (*SI Appendix*, Fig. S7) across all simulation runs.

Number of mutations and sequencing depth. We start with constructing a true underlying tree with a fixed number of subclones. Various numbers of mutations are placed on branches of the tree (except for the leftmost one) with equal probabilities, and as a result we can get a true genotyping matrix Z. The clonal frequency matrix P is fixed so that we can control for the number of subclones, the number of samples, and the clonal compositions. Here, we mimic two different sequencing pipelines—WGS with $d = 30$ and targeted sequencing with $d = 500$. The input matrix X is sampled from the

multinomial distribution and the mutant read depth matrix R is then sampled from a binomial distribution:

$$R \sim \text{Binomial}\left(X, \frac{1}{2}Z \times P\right).$$

Number of samples. We evaluate the effect of the number of samples by running parallel simulations with fixed number of subclones ($K = 5$) and mutation clusters ($S = 7$) but varied number of samples, which correspond to columns of the clonal frequency matrix P. Because adding a same sample does not guarantee adding additional information for phylogeny reconstruction, we choose and fix the elements of the P matrix so that the additive summation result is the most distinct in the unit space and that different combinations of subclones are present across different samples (*SI Appendix*, Table S4*A*). We further measure the deconvolution difficulty quantitatively from the P matrix itself. Specifically, we define $q \in \mathbb{R}^{(2K-3)\times N}$ as the summation of the offspring subclonal frequencies at each of the $(2K - 3)$ internal edges across all samples,

$$q_{ij} = \sum_{\{s:s \text{ is descendant of edge } i\}} P_{sj} (1 \le i \le 2K-3,\ 1 \le j \le N).$$

The statistic that we use to measure the deconvolution difficulty of the P matrix is as follows:

$$q_{min} = \min_{\{i \ne i'\}} \|q_i - q_{i'}\|^2,$$

where $q_i = (q_{i1}, q_{i2}, \ldots, q_{iN})$ (*SI Appendix*, Fig. S8).

Number of subclones. We study the effect of the number of subclones by keeping the P matrix the same as is shown in *SI Appendix*, Table S4*B* with varied number of rows ($3 \le K \le 10$). The number of samples is fixed at 3, among which there is the greatest distinction of clonal compositions (*SI Appendix*, Table S4*B*); the number of mutations is set at $K + 2$. In addition to measure the accuracy of the inferred Z and P matrix, we also compare Canopy's preclustering result against that of SciClone's (16). We use clustering purity as a measure of clustering quality. To compute clustering purity, each cluster is assigned to the class that is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned mutations and dividing by the total number of mutations. We further carry out simulations to examine a larger subclone space and to investigate the trade-off between the number of subclones, the sequencing depth, as well as the number of mutations. Running time and estimation errors of the Z and the P matrix are recorded (*SI Appendix*, Fig. S9).

Binomial mixture clustering. We investigate the effect of the binomial mixture clustering on computation time and deconvolution accuracy. The binomial mixture clustering is carried out as an initialization step to guide the MCMC sampling procedure—we first move the mutational clusters along the tree branches and then fine-tune every mutation within each cluster. Simulation is carried out with varying number of mutations $N \in \{25,\ 50,\ 100,\ 200\}$ along trees with different number of clones $K \in \{3,\ 4,\ 5,\ 6\}$ from three samples. The true underlying clonal frequency matrix P is the same as is in *SI Appendix*, Table S4*B*. Convergence is measured by both the log-likelihood and the acceptance rate (*SI Appendix*, Fig. S10), with running time recorded and estimation errors measured (*SI Appendix*, Table S3).

WES of Transplantable Metastasis Model Derived from MDA-MB-231. The parental cell line MDA-MB-231 was obtained from the American Type Tissue Collection. Its derivative cell lines (both SCPs and MCPs) were described previously (36, 41, 42). Cells were grown in high-glucose DMEM with 10% (vol/vol) FBS. Genomic DNA was harvested with Purelink genomic DNA kit (Invitrogen). Exome libraries were prepared with SureSelect Human All Exon kit (Agilent) and were sequenced on an Illumina HiSeq-2000 sequencer. The WES data have been deposited in the BioProject database with accession number PRJNA315318.

ACKNOWLEDGMENTS. This work was supported by NIH Grant R01-HG006137 (to N.R.Z.).

1. Nowell PC (1976) The clonal evolution of tumor cell populations. *Science* 194(4260): 23–28.
2. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: The next generation. *Cell* 144(5):646–674.
3. Vogelstein B, Kinzler KW (1993) The multistep nature of cancer. *Trends Genet* 9(4): 138–141.
4. Weinstein JN, et al.; Cancer Genome Atlas Research Network (2013) The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat Genet* 45(10):1113–1120.
5. Hudson TJ, et al.; International Cancer Genome Consortium (2010) International network of cancer genome projects. *Nature* 464(7291):993–998.
6. Navin N, et al. (2011) Tumour evolution inferred by single-cell sequencing. *Nature* 472(7341):90–94.
7. Ding L, et al. (2012) Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 481(7382):506–510.
8. Bashashati A, et al. (2013) Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *J Pathol* 231(1):21–34.
9. Gerlinger M, et al. (2014) Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat Genet* 46(3):225–233.
10. Eirew P, et al. (2015) Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* 518(7539):422–426.
11. Sottoriva A, et al. (2015) A Big Bang model of human colorectal tumor growth. *Nat Genet* 47(3):209–216.
12. Boutros PC, et al. (2015) Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat Genet* 47(7):736–745.
13. Carter SL, et al. (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 30(5):413–421.
14. Andor N, Harness JV, Müller S, Mewes HW, Petritsch C (2014) EXPANDS: Expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics* 30(1):50–60.
15. Roth A, et al. (2014) PyClone: Statistical inference of clonal population structure in cancer. *Nat Methods* 11(4):396–398.
16. Miller CA, et al. (2014) SciClone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol* 10(8):e1003665.
17. Ha G, et al. (2014) TITAN: Inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res* 24(11):1881–1893.
18. Oesper L, Mahmoody A, Raphael BJ (2013) THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol* 14(7):R80.
19. Li B, Li JZ (2014) A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biol* 15(9):473.
20. Deshwar AG, et al. (2015) PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol* 16:35.
21. Wang Y, et al. (2014) Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* 512(7513):155–160.
22. Hou Y, et al. (2012) Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* 148(5):873–885.
23. Popic V, et al. (2015) Fast and scalable inference of multi-sample cancer lineages. *Genome Biol* 16(1):91.
24. Niknafs N, Beleva-Guthrie V, Naiman DQ, Karchin R (2015) Subclonal hierarchy inference from somatic mutations: Automatic reconstruction of cancer evolutionary trees from multi-region next generation sequencing. *PLoS Comput Biol* 11(10):e1004416.
25. Zare H, et al. (2014) Inferring clonal composition from multiple sections of a breast cancer. *PLoS Comput Biol* 10(7):e1003703.
26. Yuan K, Sakoparnig T, Markowetz F, Beerenwinkel N (2015) BitPhylogeny: A probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biol* 16:36.
27. El-Kebir M, Satas G, Oesper L, Raphael BJ (2016) Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. *Cell Systems* 3(1):43–53.
28. Chen H, Bell JM, Zavala NA, Ji HP, Zhang NR (2015) Allele-specific copy number profiling by next-generation DNA sequencing. *Nucleic Acids Res* 43(4):e23.
29. Favero F, et al. (2015) Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* 26(1):64–70.
30. Van der Auwera GA, et al. (2013) From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 11(1110):11.10.11–11.10.33.
31. Cibulskis K, et al. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 31(3):213–219.
32. Lönnstedt IM, et al. (2014) Deciphering clonality in aneuploid breast tumors using SNP array and sequencing data. *Genome Biol* 15(9):470.
33. 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.
34. Gusfield D (1991) Efficient algorithms for inferring evolutionary trees. *Networks* 21(1):19–28.
35. Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61(4):893–903.
36. Minn AJ, et al. (2005) Distinct organ-specific metastatic potential of individual breast cancer cells and primary tumors. *J Clin Invest* 115(1):44–55.
37. Jacob LS, et al. (2015) Metastatic competence can emerge with selection of preexisting oncogenic alleles without a need of new mutations. *Cancer Res* 75(18):3713–3719.
38. Minn AJ, et al. (2007) Lung metastasis genes couple breast tumor size and metastatic spread. *Proc Natl Acad Sci USA* 104(16):6740–6745.
39. Wang K, Li M, Hakonarson H (2010) ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38(16):e164.
40. Wagle N, et al. (2011) Dissecting therapeutic resistance to RAF inhibition in melanoma by tumor genomic profiling. *J Clin Oncol* 29(22):3085–3096.
41. Kang Y, et al. (2003) A multigenic program mediating breast cancer metastasis to bone. *Cancer Cell* 3(6):537–549.
42. Minn AJ, et al. (2005) Genes that mediate breast cancer metastasis to lung. *Nature* 436(7050):518–524.

Supplementary Material

Assessing intra-tumor heterogeneity and tracking longitudinal and spatial clonal evolution by next-generation sequencing.

Yuchao Jiang[1,2]
Email: yuchaoj@mail.med.upenn.edu

Yu Qiu[3,4,5,6]
Email: yuqiu@upenn.edu

Andy J Minn[3,4,5,6]
Email: andyminn@exchange.upenn.edu

Nancy R Zhang[2,*]
Email: nzh@wharton.upenn.edu

[1] Genomics and Computational Biology Graduate Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
[2] Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA
[3] Abramson Family Cancer Research Institute, Perelman School of Medicine, University of Pennsylvania, PA 19104, USA
[4] Department of Radiation Oncology, Perelman School of Medicine, University of Pennsylvania, PA 19104, USA
[5] Abramson Cancer Center, Perelman School of Medicine, University of Pennsylvania, PA 19104, USA
[6] Institute of Immunology, Perelman School of Medicine, University of Pennsylvania, PA 19104, USA

* To whom correspondence should be addressed. Tel: (+1) 215-898-8222; Fax: (+1) 215-898-1280; Email: nzh@wharton.upenn.edu.

Supplementary Methods

Allele-specific copy number

For the tth ($1 \le t \le T$) CNA, we let N_t be the number of germline heterozygous loci within its segment (segmentation carried out by FALCON (1) or FALCON-X). From FALCON's segmentation and phasing outputs, we can get for each tumor-normal pair the read counts of major and minor allele in the jth tumor slice, M_{ij} and m_{ij}, and in the matched normal sample, M_{i0} and m_{i0}, where $1 \le i \le N_t$ is the SNP index and $1 \le j \le N$ is the sample index.

For CNA events that are non-overlapping (Supplementary Figure S3a), we use the germline heterozygous loci within each CNA segment to compute major and minor copy number input across all samples:

$$W_{tj}^{M} = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(M_{ij}/M_{i0}\right), \left(\varepsilon_{tj}^{M}\right)^2 = \frac{\sum_{i=1}^{N_t}\left(M_{ij}/M_{i0}\right)^2 - N_t\left(W_{tj}^{M}\right)^2}{N_t(N_t-1)};$$

$$W_{tj}^{m} = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(m_{ij}/m_{i0}\right), \left(\varepsilon_{tj}^{m}\right)^2 = \frac{\sum_{i=1}^{N_t}\left(m_{ij}/m_{i0}\right)^2 - N_t\left(W_{tj}^{m}\right)^2}{N_t(N_t-1)}.$$

In the above, W_{tj}^{M}, W_{tj}^{m} are the estimates of the major and minor copy numbers, respectively, and $\varepsilon_{tj}^{M}, \varepsilon_{tj}^{m}$ can be considered their standard errors.

For CNA events that are overlapping or nested (Supplementary Figure S3b), we propose a 4-step prioritization algorithm to get the major and minor copy numbers for each event, briefly summarized as follows: (i) Merge CNA events where both endpoints are close, e.g. within 1 kb of each other; (ii) Identify nested CNA events, e.g., a homozygous deletion residing in a one-copy deletion region; (iii) Rank overlapping and nested CNA events by a Chi-square score, details below; (iv) Get major and minor copy number estimates through a recursive procedure. Now we expand on the details, with an illustrative example shown in Supplementary Figure S3. Let $E_1, E_2, \ldots, E_T$ be the CNA events collected across all samples after the merging step (i), which may contain nested or overlapping events; let $\pi_t^{(j)}$ ($1 \le \pi_t^{(j)} \le T$) be the ranking of event t ($1 \le t \le T$) in sample j ($1 \le j \le N$) (Supplementary Figure S3c) based on its Chi-squared statistic,

$$Q_{tj} = \left(\frac{W_{tj}^{M}-1}{\varepsilon_{tj}^{M}}\right)^2 + \left(\frac{W_{tj}^{m}-1}{\varepsilon_{tj}^{m}}\right)^2 \sim \chi_2^2,$$

with larger Chi-square ranked higher (i.e. smaller $\pi_t^{(j)}$ value), but with an important caveat that nested events always takes precedence over the event that it resides in regardless of their Chi-square values, e.g., homozygous deletion event E_3 always has a higher ranking than heterozygous deletion E_2 (Supplementary Figure S3c). Another important detail is that, at this point, the input values W^M, W^m, ε^M, and ε^m used to compute Q_{tj} are estimated from segments with shared breakpoints across all samples due to the preceding merging step. As a

result, in some samples certain segments may have a mixture of more than one copy number state if it overlaps with a different CNA from another sample, e.g., in Supplementary Figure S3b sample 1 has three copy number states in the segment that corresponds to event E_1. These segments won't have the highest Chi-squared values so they should be ranked low, as desired. To get an accurate estimate of major and minor copy numbers for overlapping and nested CNAs we adopt the algorithm outlined below, the result of which on the illustrative example is also shown in Supplementary Figure S3c.

For each sample j,

(1) Start with event t with the highest ranking: $\pi_t^{(j)} = 1$, get W_{tj}^M, W_{tj}^m, ε_{tj}^M, and ε_{tj}^m by taking the mean and standard error across all heterozygous loci that reside within this event;

(2) For event t: $\pi_t^{(j)} > 1$, in computing the major and minor copy number input, use segment E_t excluding all segments of lower rank, that is,

$$E_t \setminus \bigcup_{\pi_{t'}^j < \pi_t^j} E_{t'}.$$

Binomial mixture clustering

Let $R = \{R_s : 1 \le s \le S\}$ be the number of reads supporting variant and $X = \{X_s : 1 \le s \le S\}$ be the total number of reads. We assume that $p(R, X)$ is defined as a finite Binomial mixture model with $(K + 1)$ components. The incomplete-data log-likelihood expression is given by:

$$\log\big(L(\Theta|R,X)\big) = \log \prod_{s=1}^{S} p(R_s, X_s|\Theta)$$
$$\propto \sum_{s=1}^{S} \log\left\{\sum_{k=1}^{K} [\varphi_k {\mu_k}^{R_s}(1-\mu_k)^{X_s-R_s}] + \varphi_{K+1}\int_0^1 p^{R_s}(1-p)^{X_s-R_s}dp\right\},$$

where $\sum_{k=1}^{K+1} \varphi_k = 1$ are the mixture weights with the kth component parameterized by μ_k $(1 \le k \le K)$. To gain robustness against random mutational calls that are false positives, we add a $(K + 1)$th mixture component, with a small weight φ_{K+1}, that is bivariate uniform on the unit interval. We adopt an expectation-maximization (EM) algorithm to estimate the Binomial mixtures and use BIC to determine the number of clusters. Our method takes into account varying read depth of each mutation and is thus robust against outliers. The estimated mutation frequencies of each mutation wave are given in Supplementary Figure S14a and are used as input to infer longitudinal evolutionary trajectories. This pre-clustering procedure serves as an

initialization step in our MCMC samplings, where we firstly move mutations along tree branches in clusters and then fine tune individual mutations within each cluster.

CNA profiling of MDA-MB-231 and *in vivo* derived sublines

We firstly apply a HMM to segment the SCP samples. The input are the alternative / B allele frequencies (BAFs) computed by the Genome Analysis Toolkit best practices pipeline (2) using hg19 as reference. After variant recalibration and quality control procedures (mutations have only one alternative allele and lie within exonic baits), 10,242 mutational loci are called across all samples and the BAFs are used as the known states of the HMM.

Since there are no confounding effects of cell mixtures in the SCP samples, the allele specific copy numbers should be integer values and they correspond to the hidden states of the HMM—*LOH* (0+1, 0+2, 0+3, 0+4, 0+5, 0+6), *Neutral* (1+1, 2+2, 3+3), *OneTwo* (1+2, 2+4), *OneThree* (1+3), *OneFour* (1+4), *TwoThree* (2+3—this will be rare since both the reference and alternative alleles get amplified). These states include all combinations of allelic copy numbers whose total copy numbers are less than or equal to 6 (except for 1+5, which is hard to be distinguished from and is thus grouped with 1+4). The emission probabilities are taken from a Gaussian mixture, with each component centered at the expected allelic frequency given the copy number state, $\{0, C^m/(C^M + C^m), C^M/(C^M + C^m), 1\}$ (Supplementary Figure S11). The transition probabilities favor incremental changes that move between adjacent states (Supplementary Table S6).

After HMM's segmentation, we further compute the log ratios of exonic coverages between two SCP samples (black dots in Supplementary Figure S12a) and compare that against the log ratios of the called total copy numbers (purple line in Supplementary Figure S12a). This can control for false discoveries and solve ambiguities of allele specific copy numbers since HMM cannot differentiate 0+1 from 0+2 (both LOH) or 1+1 from 2+2 (both copy neutral) using BAFs as input. We then use the SCP samples as controls to infer the CNA states in the MCP samples (Supplementary Figure S12b).

Supplementary Results

Application to Normal, Primary Tumor, and Relapse Genome of Leukemia Patients

As proof of principle and to further illustrate our method, we apply Canopy to the longitudinal dataset from Ding *et al.* (3), where whole-genome sequencing was performed on the normal tissue, the primary tumor, and the relapse genome of leukemia patients. 1292 and 412 candidate somatic SNAs and indels were identified in sample AML43/UPN869586 and AML1/UPN933142 respectively and were confirmed by deep sequencing (3). CNAs (total copy numbers) were also predicted (3).

By the weak parsimony assumption and in a similar fashion to Pyclone (4) and SciClone (5), we first adopt a binomial mixture clustering method (details in Supplementary Methods) to cluster all the mutations into

mutational 'waves' and give an estimate of the VAF of each cluster (Supplementary Figure S14a). To gain robustness against false positives calls, we add a mixture component (shown as pink dots in Supplementary Figure S14a), with a small weight, that is uniform on the unit interval. Our clustering results show that in both patients, there is a one unique mutation cluster identified in the primary tumor, and one found at relapse. Furthermore, all mutation clusters are heterozygous and diploid, except mutation cluster 1 (mut1) in AML43, the mutations in which all reside in a copy number neutral LOH region from chr 16. The VAFs of the SNAs as well as the absolute copy number of the LOH (major copy 2, minor copy 0) are used as input for Canopy to infer phylogenetic trees.

For this dataset, Canopy returns only one plausible clonal history that can explain the observed mutation profiles, shown in Supplementary Figure S14b. We re-parameterize the model to accommodate a redistribution event between the two time-points to improve interpretability. The tree is observed twice, first at the collection of the primary tumor, and then at the stage of relapse malignancy. Through the selection bottleneck that is imposed between the two time-points, mutations can arise (e.g., mutation cluster 5 (mut5) shown in red in Supplementary Figure S14b) and clonal frequencies (shown in blue in Supplementary Figure S14b) can change—some subclones expand while others become extinct or remain dormant. Meaningful quantities can be marginalized from the posterior distribution in the tree space. For example, Supplementary Figure S14c shows the posterior distribution of the clonal frequencies through the selection bottleneck—a minor clone (clone 3 in AML43 and clone 4 in AML1) carrying the vast majority (but not all) of the primary tumor mutations survives the chemotherapy and becomes dominant at relapse by acquiring additional mutations (mut5 in both samples) while the remaining clones diminish (Supplementary Figure S14c). Normal cell fractions are also estimated with their posterior distributions shown in the first columns of Supplementary Figure S14c.

While Ding *et al.* (3) arrived at this same clonal history manually (a minor clone carrying the vast majority of the primary tumor mutations survived and expanded at relapse), we automate the analysis pipeline via Canopy and allow the inclusion of both SNAs and CNAs. Canopy's inferred phylogenies shown in Supplementary Figure S14b, as well as its estimated clonal frequencies and tumor purities shown in Supplementary Figure S14c, are concordant with the results and conclusions in Ding *et al.* (3).

Application to ten spatially separated samples of ovarian cancer

We further evaluate Canopy's performance on a data set with spatial experimental design from Bashashati *et al.* (6). 63 somatic mutations (SNAs and indels) were confirmed by deep amplicon resequencing in ten tumor samples from different dissections (4a-4e, right ovary; 4f-4i, left ovary; 4j, left fallopian tube) of a high-grade serous ovarian cancer patient (Supplementary Figure S15a-b). We keep the same assumption as in Bashashati *et al.* (6) that the 63 SNAs across all samples are heterozygous from copy number neutral regions as in the original

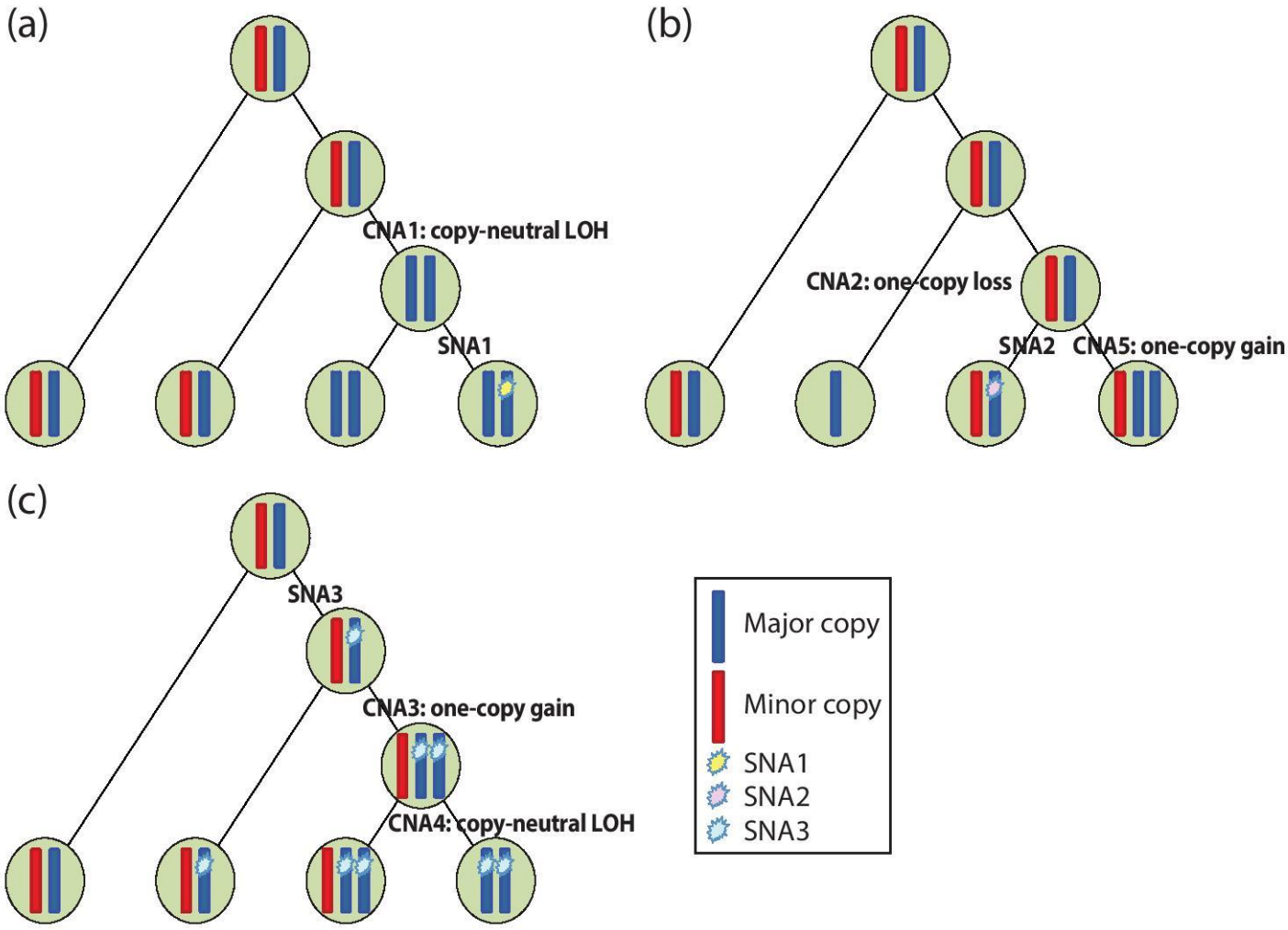
studies: (i) CNAs weren't profiled in all samples by Affymetrix SNP genotyping arrays; (ii) for the samples with CNA calls, only total copy number is available (6).

BIC for model selection is shown in Supplementary Figure S15c and the number of subclones is chosen at 5. Canopy returns posterior trees with one configuration and it is shown in Supplementary Figure S15d. Different mutations correspond to rows in the heatmap in Supplementary Figure S15a and are grouped on branches with different colors. Specifically, all ten samples share and acquire somatic mutations in *TP53* and *DHX8*, along with 13 other mutations in mutation set 2, 3, and 4 shown in light blue, green, and orange, indicating a common cell of origin. It is also observed that there is a clear separation between the samples from the right ovary and the samples from the left ovary in the clonal frequency matrix *P*. *GLDC*, *LIG1* as well as the rest mutations in mutation set 5 shown in blue drive and/or mark the divergence and thus have the potential to serve as a biomarker to indicate whether distal metastasis is formed in ovarian cancer patient. Mutation set 7 in red further distinguish case4a from 4b-be and form a unique subclone in case4a.

Collectively, our results suggest that multiple sublcones migrate from the left ovary to the right ovary and that both sample sets are mixtures of different subclones with diversified mutational profiles. These mutational profiles from spatially separated samples correlate with spatial distribution due to regional evolutionary selection and reflect different histological evolutionary trajectories within a single patient.

Notably, spatial distribution of the samples in the phylogeny is concordant with the tree configuration inferred by Bashashati *et al.* (6) (Supplementary Figure S15e). Nevertheless, the neighbor joining method with Pearson correlation distance metric doesn't account for many aspects including: (i) varying standard errors in the estimates for mutational frequencies due to varying sequencing depths; (ii) each spatial sample offers a snapshot of different combinations of subclones and therefore they cannot be treated as homogeneous samples at the tips of the tree branches; (iii) there is no placement of mutation along the tree; (iv) the inference of branch lengths assumes a constant biological clock, which doesn't hold in cancer genomes. Popic *et al.* (7) also reconstructed a clonal tree (Supplementary Figure S15f) that is highly similar to the one returned by Bashashati *et al.* (6). Somatic mutations arise from the germline (GL) sample and are placed in the phylogeny with numbers shown on tree branches. There are three subclones with distinct mutational. The proportion of the subclonal admixtures, however, remains unknown with samples at tree tips.

Supplementary Figure S1. Phases and orders of overlapping CNAs and SNAs. Five CNAs from three genomic regions affect three SNAs. Major and minor copies are in blue and red respectively; SNA mutational loci are shown as stars. **(a)** CNA precedes SNA. **2b.** SNA and CNAs are on separate branches. SNA is unaffected by CNAs. **2c.** SNA precedes CNAs. SNA can reside in major or minor copy after the CNA events. SNA4 is unaffected by CNA and is not shown.



Supplementary Figure S2. Input and matrix representation of Canopy's inference of phylogeny with overlapping CNAs. The true underlying tree structure is shown in Supplementary Figure S1. **(a)** Input of Canopy for cases where overlapping CNAs are observed. Matrix C specifies CNA and CNA region overlap. **(b)** Matrix representation of Canopy's inference for cases where overlapping CNAs are observed. SNA-CNA phase H is adapted from a vector shown in Figure 1 to a matrix since an SNA can be affected by multiple CNAs.

(a) Observed major copy (W^M)

	sample1	sample2	sample3
CNA_region1	1.59	1.46	1.49
CNA_region2	1.29	1.49	0.97
CNA_region3	1.58	1.5	1.49

(b) Integer major copy ($\tilde{C}^M$)

	clone1	clone2	clone3	clone4
CNA_region1	1	1	2	2
CNA_region2	1	1	1	2
CNA_region3	1	1	2	2

Observed minor copy (W^m)

	sample1	sample2	sample3
CNA_region1	0.43	0.52	0.54
CNA_region2	0.8	0.81	0.69
CNA_region3	0.71	0.53	1.02

Integer minor copy ($\tilde{C}^m$)

	clone1	clone2	clone3	clone4
CNA_region1	1	1	0	0
CNA_region2	1	0	1	1
CNA_region3	1	1	1	0

Observed VAF (R/X)

	sample1	sample2	sample3
SNA1	0.169	0.217	0.001
SNA2	0.158	0.001	0.319
SNA3	0.587	0.57	0.544
SNA4	0.284	0.22	0.24

SNA carrier status (Z)

	clone1	clone2	clone4	clone5
SNA1	0	0	0	1
SNA2	0	0	1	0
SNA3	0	1	1	1
SNA4	0	0	1	1

SNA-CNA_region overlap (Y^T)

	SNA1	SNA2	SNA3	SNA4
non_CNA_region	0	0	0	1
CNA_region1	1	0	0	0
CNA_region2	0	1	0	0
CNA_region3	0	0	1	0

SNA-CNA phase (H)

	CNA1	CNA2	CNA3	CNA4	CNA5
SNA1	0	0	0	0	0
SNA2	0	0	0	0	0
SNA3	0	0	1	1	0
SNA4	0	0	0	0	0

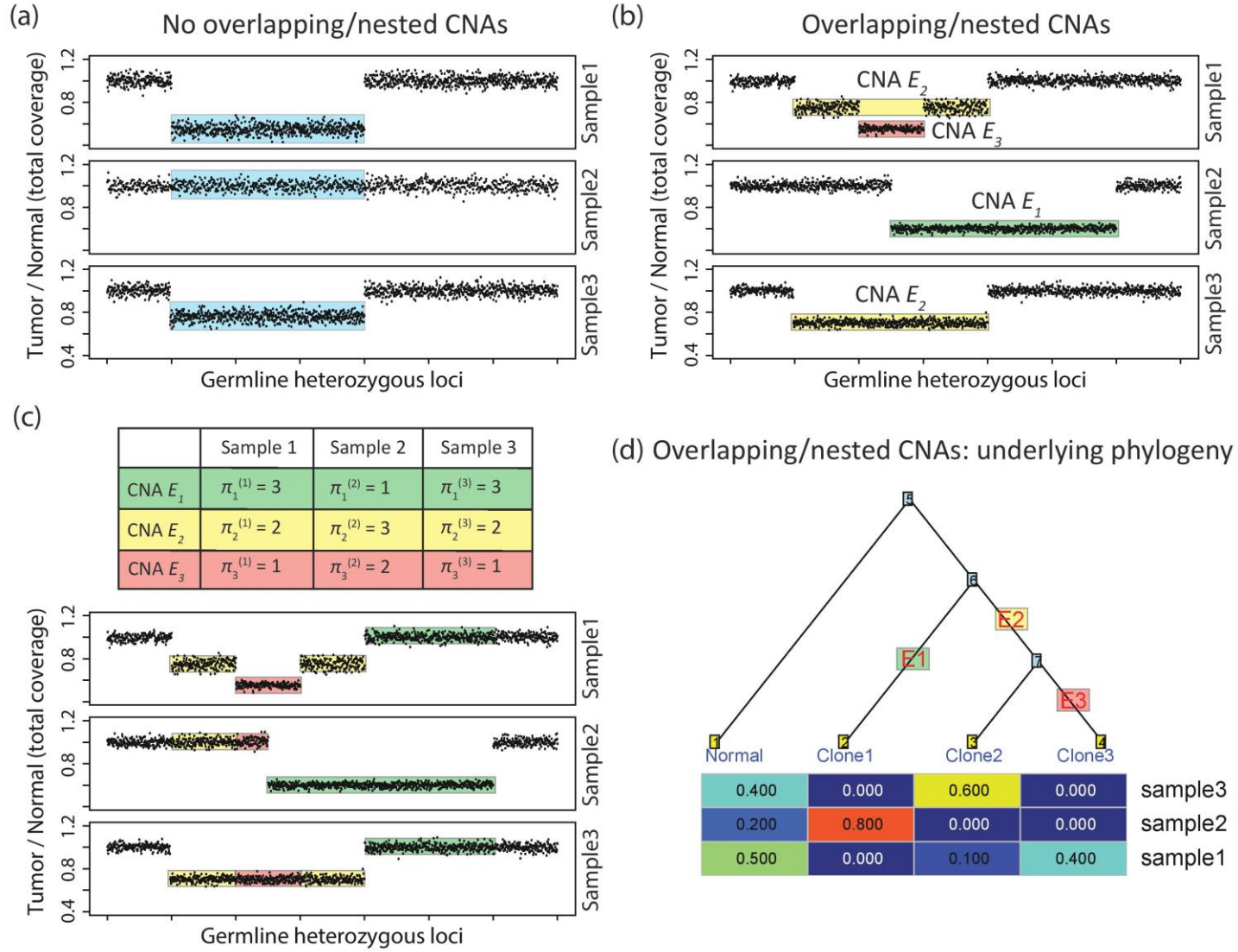
SNA-CNA_region overlap (C)

	CNA1	CNA2	CNA3	CNA4	CNA5
CNA_region1	1	0	0	0	0
CNA_region2	0	1	0	0	1
CNA_region3	0	0	1	1	0

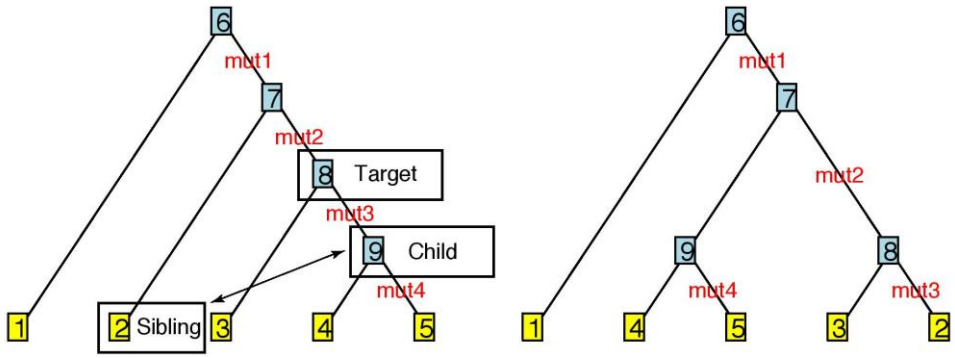
Clonal frequency (P)

	sample1	sample2	sample3
clone1	0.2	0.3	0.2
clone2	0.2	0.2	0.3
clone3	0.3	0	0.5
clone4	0.3	0.5	0

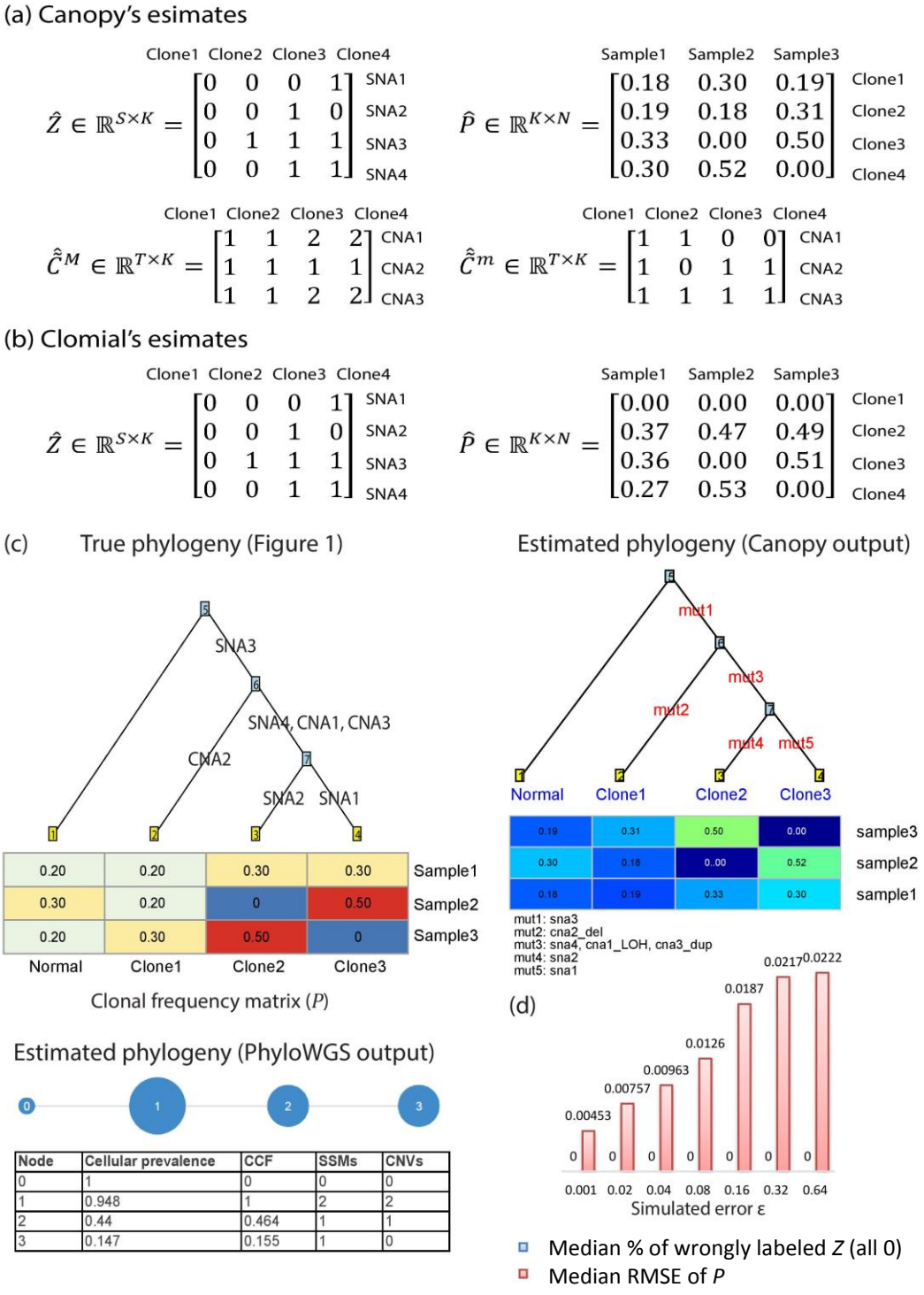
Supplementary Figure S3. Illustration on generating CNA input for Canopy. Initial segmentation is performed by FALCON-X. **(a)** For CNAs that aren't overlapping or nested, the segment mean and standard error are computed for each segment across all samples (Methods in main manuscript). **(b)** For CNA events that display overlapping/nested structure, a four-step CNA prioritization algorithm (Supplementary Methods) is adopted. **(c)** The ranking of CNA events in each sample and the segments that are used to generate allele-specific copy number calls. **(d)** The underlying tree structure for samples and CNA events shown in (b).



Supplementary Figure S4. Generating new tree topology by local rearrangement. A neighborhood—an internal node that has both a parent and two children—is selected for local rearrangement. Switch the sibling with one of the children to generate a new tree topology (8).



Supplementary Figure S5. Inferred phylogenies by Canopy, Clomial and PhyloWGS with true underlying phylogeny shown in Figure 1 as input. (a) Canopy successfully decomposes all matrices with confidence assessment. **(b)** Clomial doesn't utilize somatic CNA information and fails to estimate the clonal frequencies with zero normal cell contaminations in all three samples. The true quantities are shown in Figure 1. **(c)** True phylogeny and estimated phylogeny by Canopy and PhyloWGS. Canopy returned a tree highly concordant with the ground truth whereas PhyloWGS returned a linear tree with incorrectly inferred cellular frequencies. The input for this dataset can be found in the Canopy R-package. **(d)** Higher noise (spiked-in error term ε) doesn't seem to affect Canopy's estimation of the genotyping matrix Z but leads to higher estimation error of the clonal proportion P. The estimation error is taken as the median across ten parallel runs.



Supplementary Figure S6. Inferred phylogenies by Canopy and Clomial (9) with true underlying tree shown in Supplementary Figure S1-S2 as input. (a) Canopy successfully decomposes all matrices with confidence assessment. **(b)** Clomial doesn't utilize somatic CNA information and fails to estimate the clonal frequencies with zero normal cell contaminations in all three samples. The true quantities are shown in Supplementary Figure S2b.

(a) Canopy's esimates

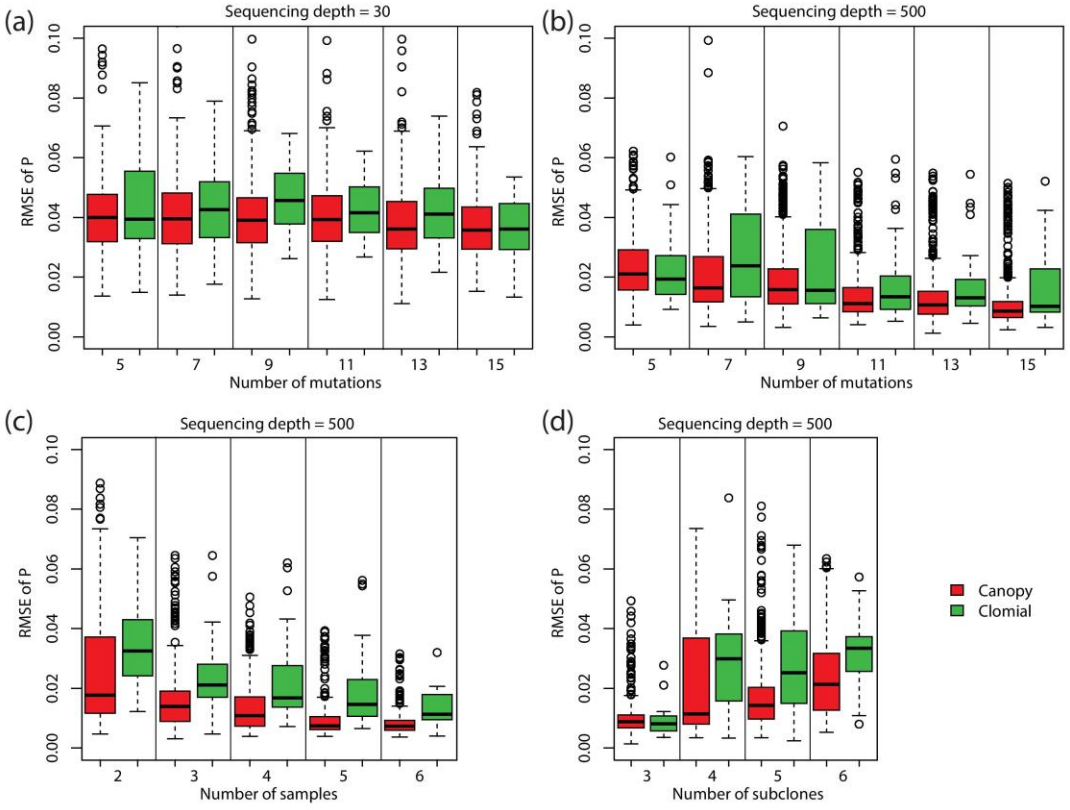
$$\hat{Z} \in \mathbb{R}^{S\times K} = \begin{matrix} \\ \text{SNA1} \\ \text{SNA2} \\ \text{SNA3} \\ \text{SNA4} \end{matrix} \begin{matrix} \begin{matrix} \text{Clone1} & \text{Clone2} & \text{Clone3} & \text{Clone4} \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{matrix} \qquad \hat{P} \in \mathbb{R}^{K\times N} = \begin{matrix} \begin{matrix} \text{Sample1} & \text{Sample2} & \text{Sample3} \end{matrix} \\ \begin{bmatrix} 0.22 & 0.34 & 0.18 \\ 0.20 & 0.18 & 0.32 \\ 0.30 & 0.00 & 0.50 \\ 0.28 & 0.48 & 0.00 \end{bmatrix} \end{matrix} \begin{matrix} \\ \text{Clone1} \\ \text{Clone2} \\ \text{Clone3} \\ \text{Clone4} \end{matrix}$$

$$\hat{\hat{C}}^M \in \mathbb{R}^{T\times K} = \begin{matrix} \\ \text{CNA1} \\ \text{CNA2} \\ \text{CNA3} \end{matrix} \begin{matrix} \begin{matrix} \text{Clone1} & \text{Clone2} & \text{Clone3} & \text{Clone4} \end{matrix} \\ \begin{bmatrix} 1 & 1 & 2 & 2 \\ 1 & 1 & 1 & 2 \\ 1 & 1 & 2 & 2 \end{bmatrix} \end{matrix} \qquad \hat{\hat{C}}^m \in \mathbb{R}^{T\times K} = \begin{matrix} \\ \text{CNA1} \\ \text{CNA2} \\ \text{CNA3} \end{matrix} \begin{matrix} \begin{matrix} \text{Clone1} & \text{Clone2} & \text{Clone3} & \text{Clone4} \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \end{matrix}$$

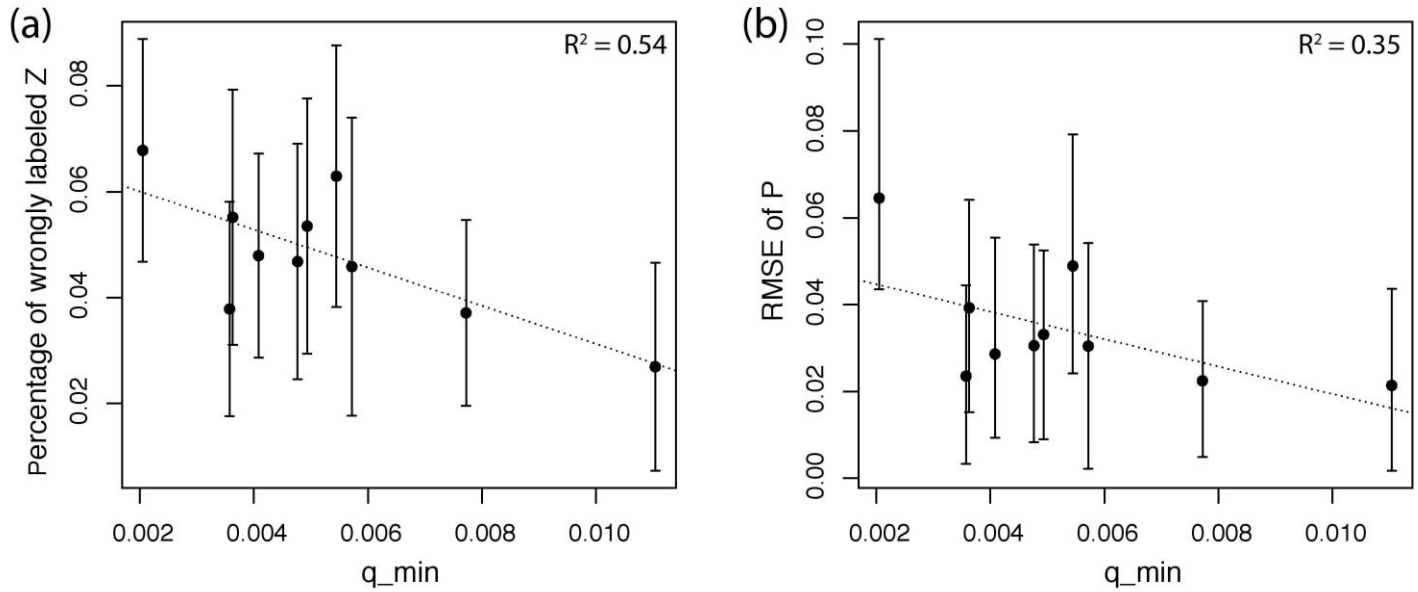
(b) Clomial's esimates

$$\hat{Z} \in \mathbb{R}^{S\times K} = \begin{matrix} \\ \text{SNA1} \\ \text{SNA2} \\ \text{SNA3} \\ \text{SNA4} \end{matrix} \begin{matrix} \begin{matrix} \text{Clone1} & \text{Clone2} & \text{Clone3} & \text{Clone4} \end{matrix} \\ \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{matrix} \qquad \hat{P} \in \mathbb{R}^{K\times N} = \begin{matrix} \begin{matrix} \text{Sample1} & \text{Sample2} & \text{Sample3} \end{matrix} \\ \begin{bmatrix} 0.00 & 0.00 & 0.00 \\ 0.36 & 0.47 & 0.45 \\ 0.35 & 0.00 & 0.55 \\ 0.29 & 0.53 & 0.00 \end{bmatrix} \end{matrix} \begin{matrix} \\ \text{Clone1} \\ \text{Clone2} \\ \text{Clone3} \\ \text{Clone4} \end{matrix}$$

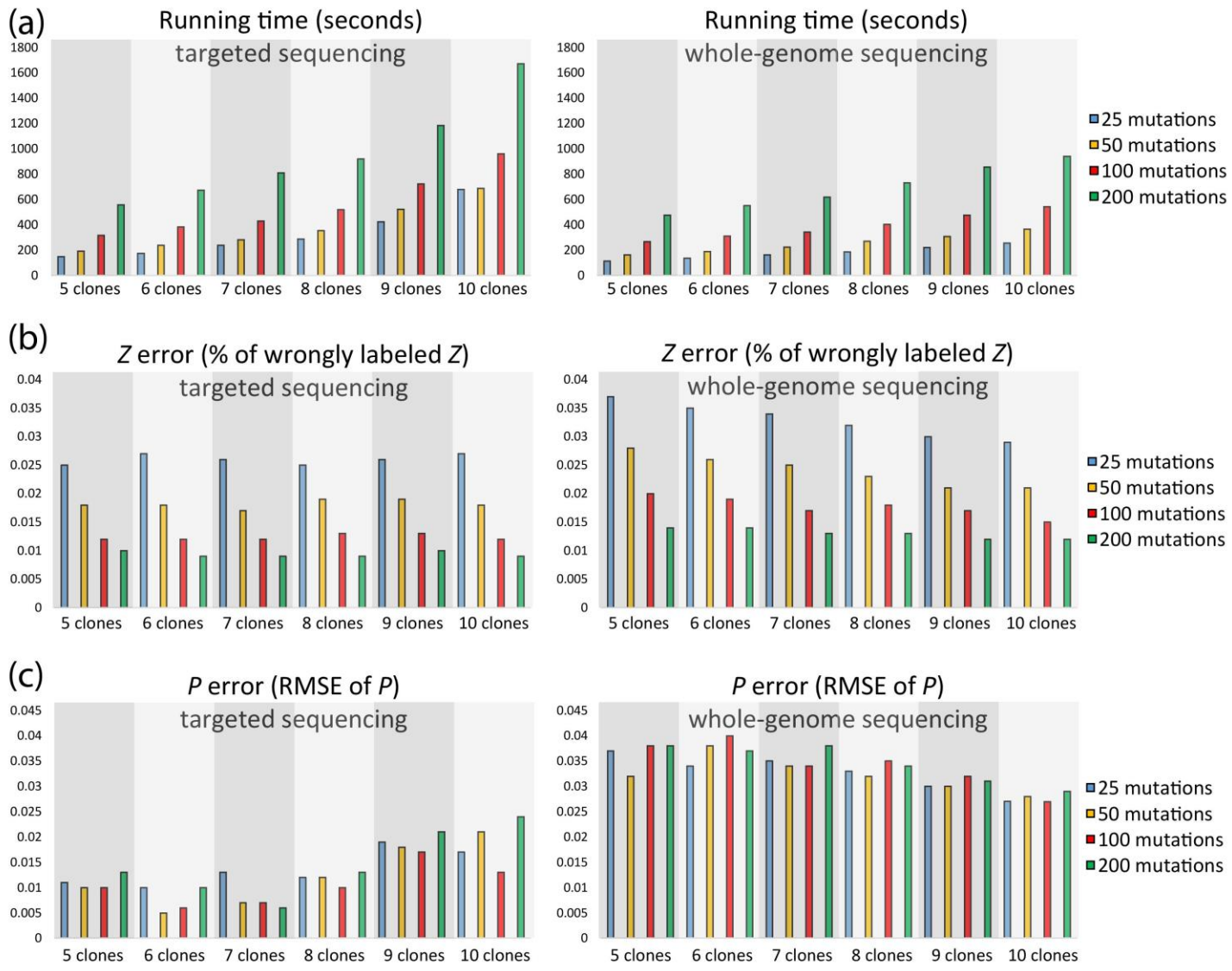
Supplementary Figure S7. Deconvolution accuracy via simulation studies. Various parameters show effects on deconvolution accuracy (measured by RMSE of the *P* matrix). The corresponding percentage of wrongly labeled *Z* elements is shown in Figure 3. **(a-b)** Whole-genome sequencing compensates the lower sequencing depth with more profiled mutations. **(c)** Large number of samples helps solve reconstruction ambiguity. **(d)** Number of subclones is negatively correlated with deconvolution accuracy. Canopy outperforms Clomial under all settings.



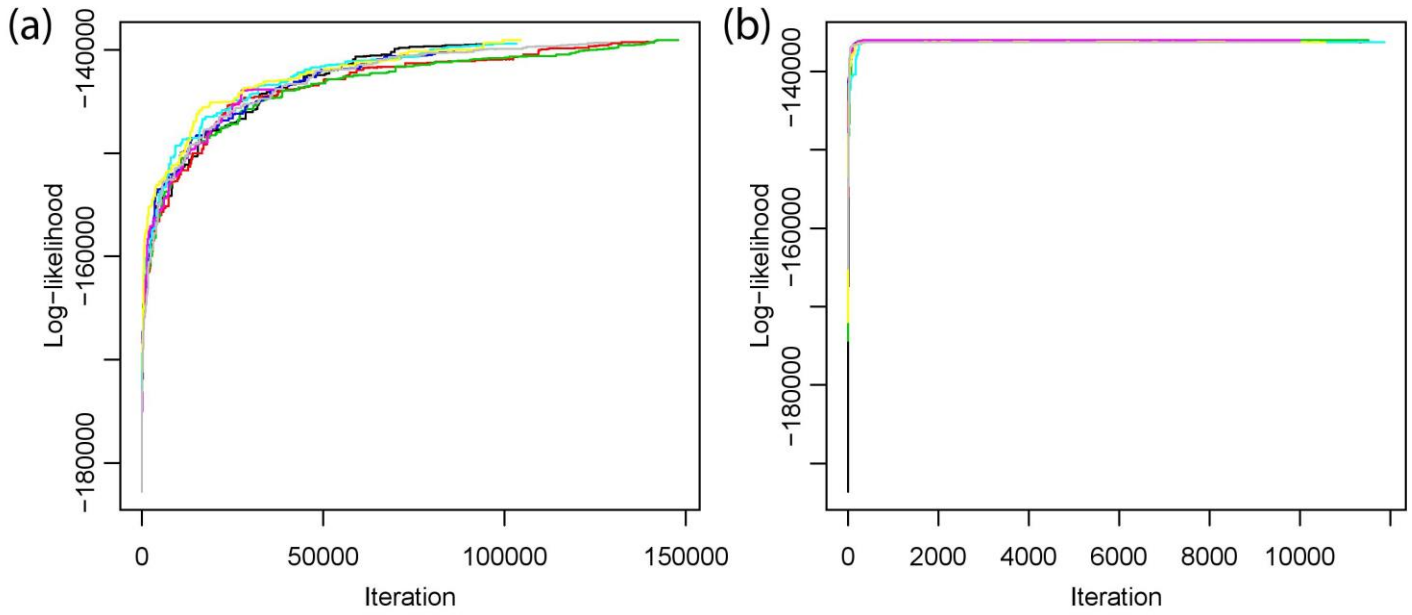
Supplementary Figure S8. q_{min} as a measure of deconvolution difficulty from the clonal frequency matrix $\boldsymbol{P}$. The larger the q_{min} is, the more distinct the clonal frequencies at the tree edges are, and thus the more difficult the deconvolution problem is.



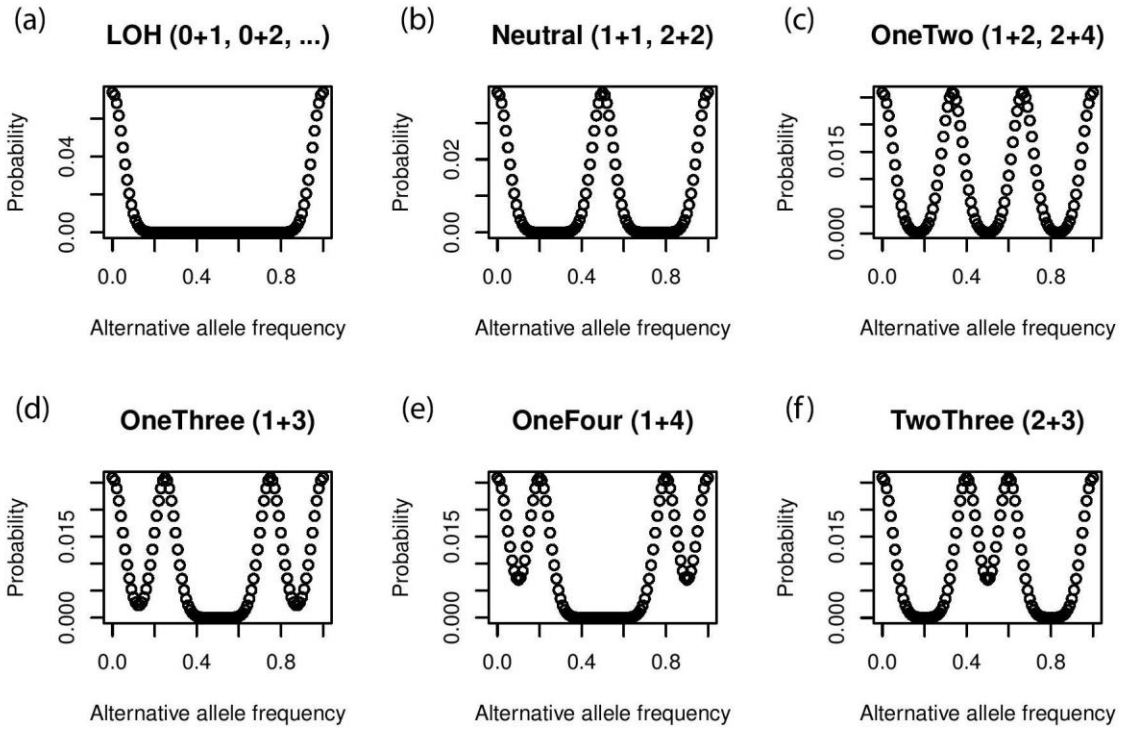
Supplementary Figure S9. Simulation results on tradeoff for large number of subclones, number of mutations, and sequencing depth. Refer to Methods section for details on the simulation setup. Whole-genome sequencing ($d = 30$) has higher estimation errors but shorted running time, compared to deep targeted sequencing ($d = 30$). Increasing number of clones doesn't seem to affect estimation under the condition that enough genotypically distinct slices of tumor have been profiled. Increasing number of mutations leads to more accurate estimate of the genotyping matrix Z but has little effect on the clonal frequency matrix P.



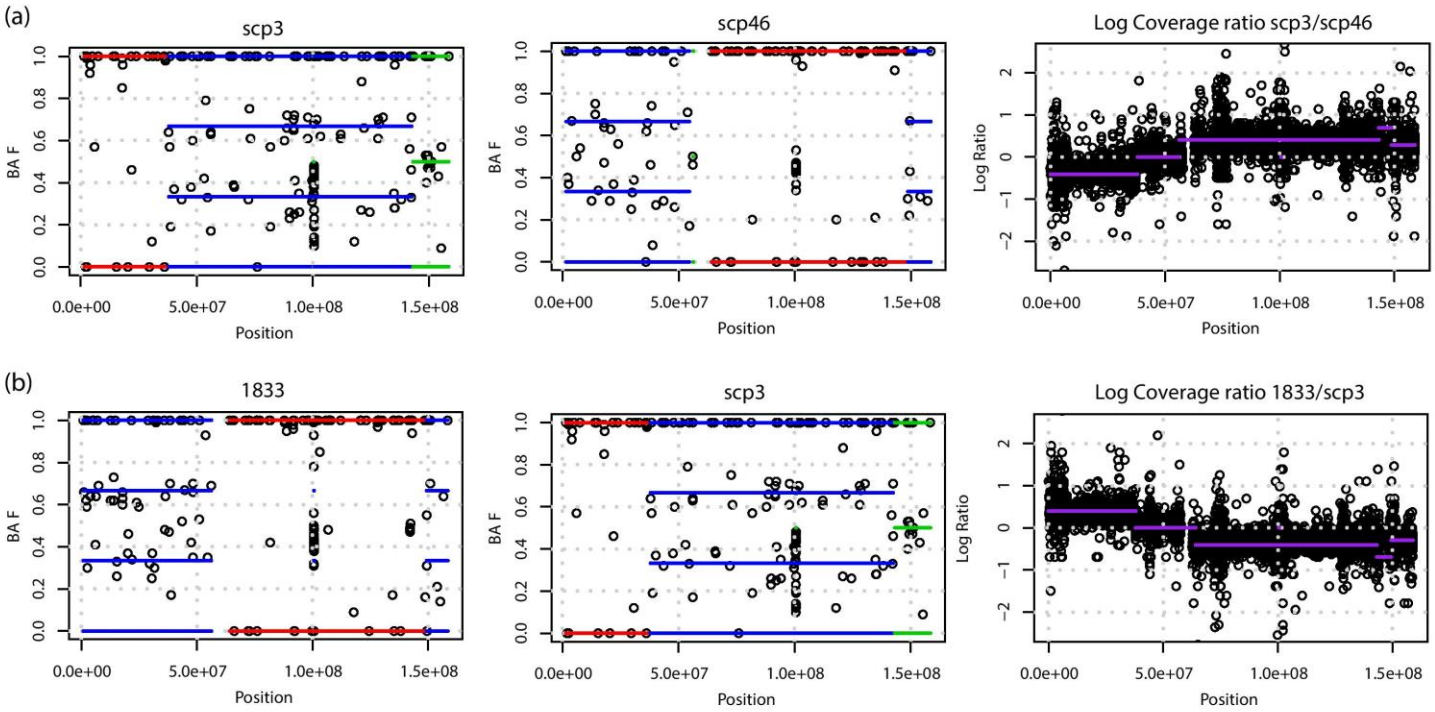
Supplementary Figure S10. Log-likelihood of MCMC sampling with and without pre-clustering step. Simulation is carried out with 200 mutations along a five-branch tree using three samples. Ten chains shown in different color are randomly started with (a) and without (b) a Binomial mixture clustering step. Convergence is measured by both the log-likelihood and the acceptance rate. Pre-clustering step significantly reduces computation time with MCMC converging faster.



Supplementary Figure S11. Emission probabilities of the HMM to profile allele specific copy number in SCP samples. Probabilities are taken from Gaussian mixtures with each component centered at the expected BAF for six hidden states **(a)** LOH, **(b)** Neutral, **(c)** OneTwo, **(d)** OneThree, **(e)** OneFour, **(f)** TwoThree.

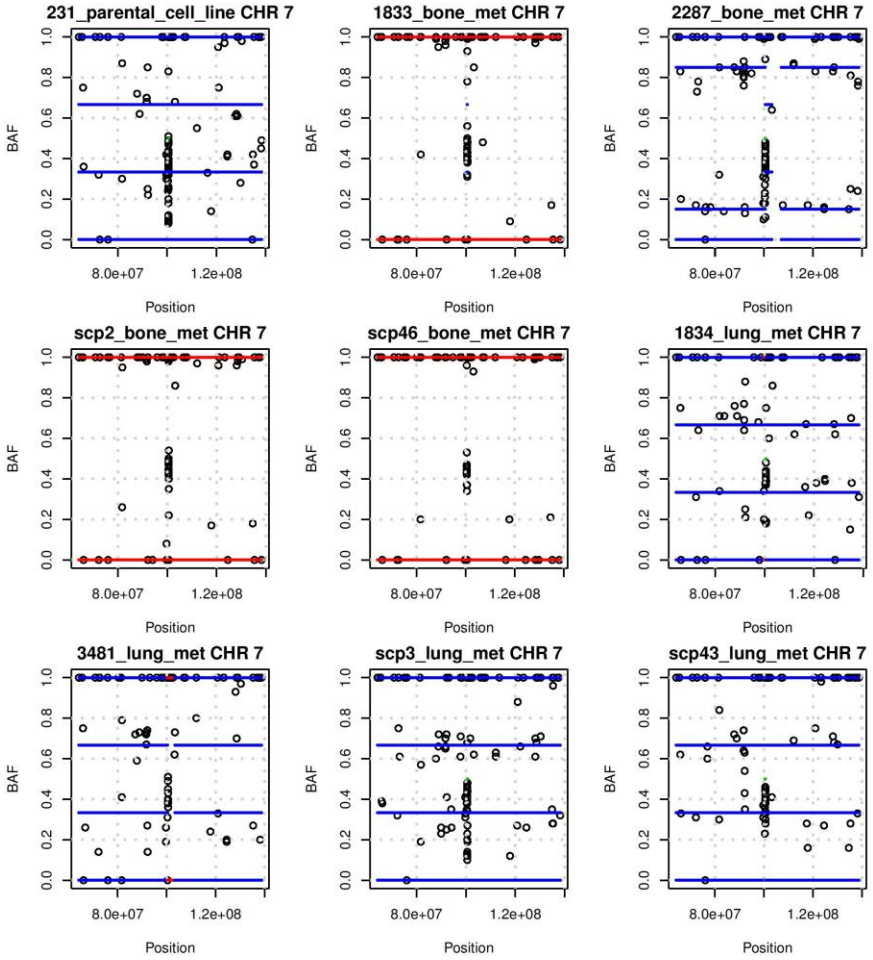


Supplementary Figure S12. CNA inference by HMM. (a) HMM is applied to segment the genome in SCP samples and manually corrected by the exonic coverage ratios between two SCPs. B allele frequencies (BAFs) are used as input. Deletion/LOH is shown in red, duplication in blue, and copy number neutral region in green. Purple line is the log ratio of the segmented total copy numbers, overlaid by the corresponding depth of coverage ratio. **(b)** Using SCP as a normal control, CNAs for the MCP sample is called.

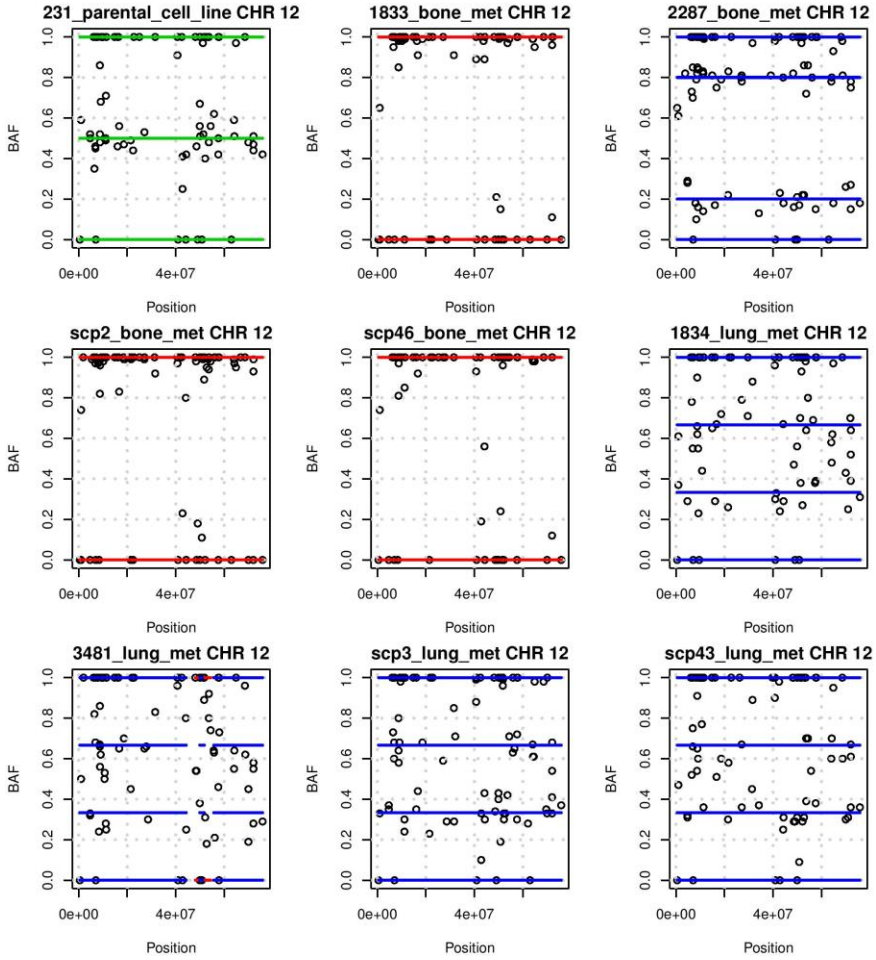


Supplementary Figure S13. Canopy's CNA input to infer phylogeny in the parental cell line and its sublines. Six somatic CNAs from four different chromosomes—**(a)** chr7, **(b)** chr12, **(c)** chr18, **(d)** chr19. Chr7 and chr12 are double 'hit' by two CNAs; chr18 and chr19 undergo one-copy loss and gain respectively. CNA subclonal events result in different allele specific copy number states across different samples. The observed B allele frequencies (BAFs), i.e., $W^M/(W^M + W^m)$ and $W^m/(W^M + W^m)$, are used as input for Canopy to infer the clonal tree.

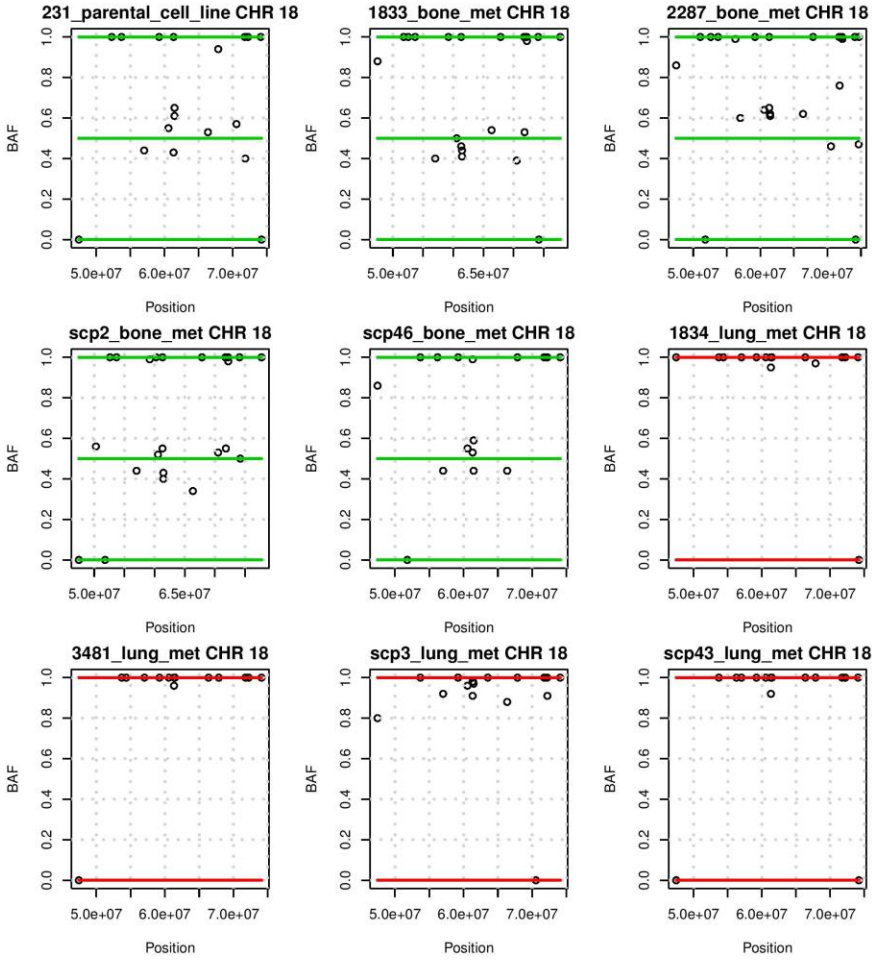
(a)



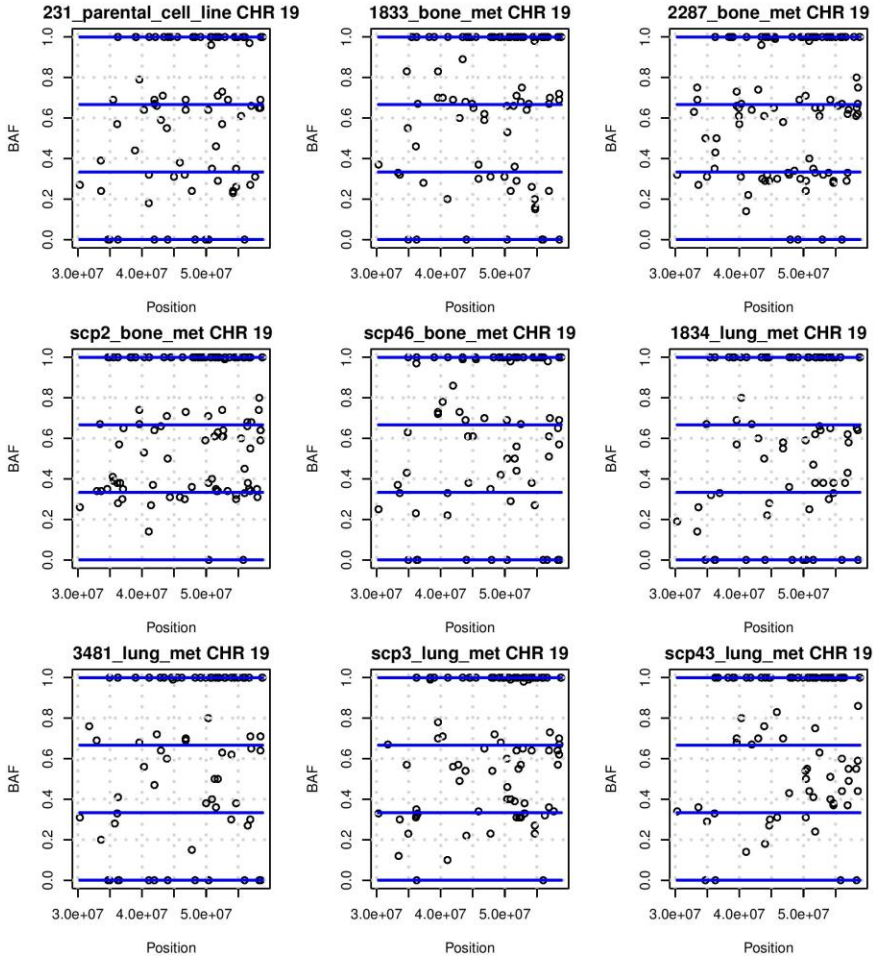
(b)



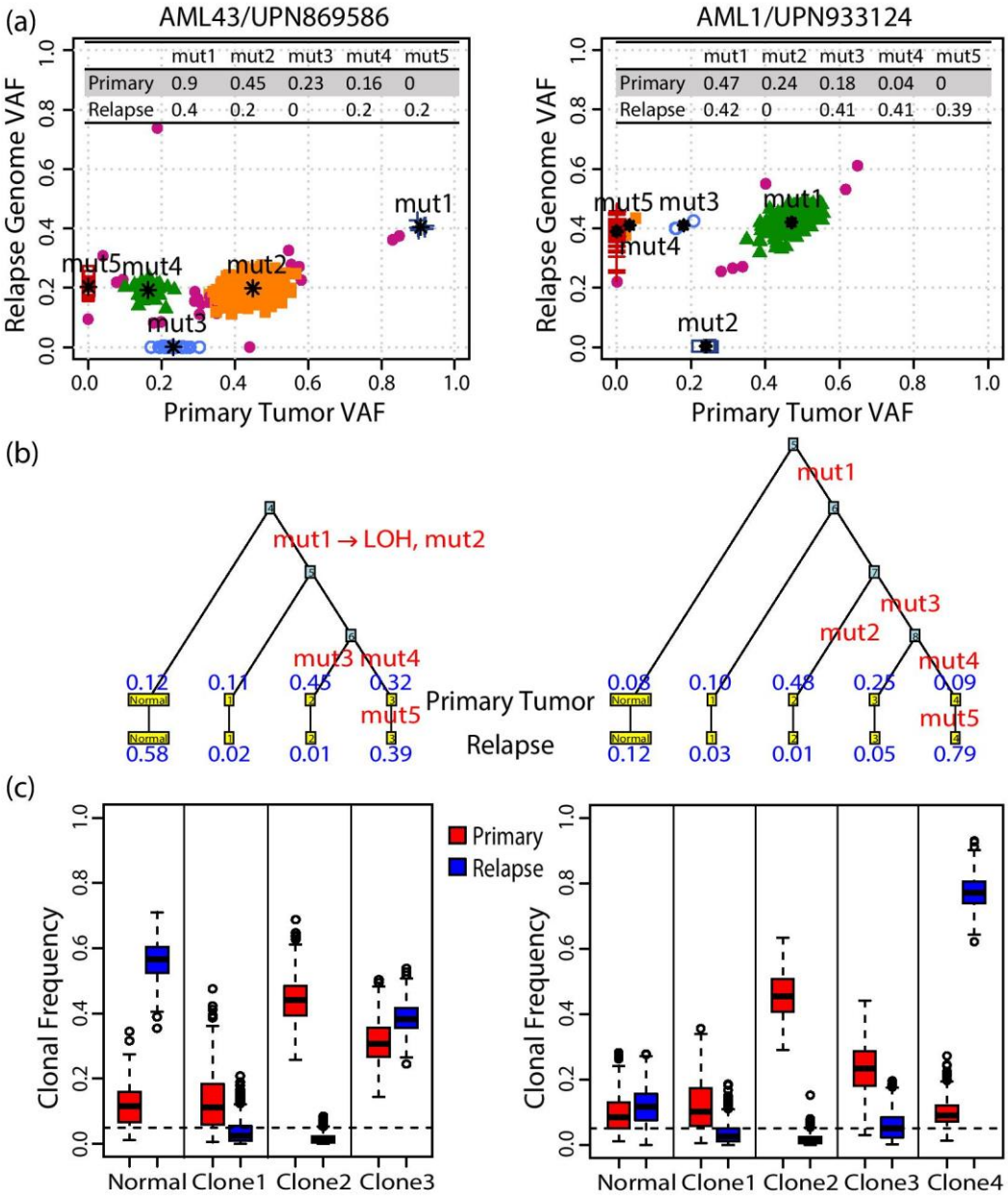
(c)


231_parental_cell_line CHR 18
1833_bone_met CHR 18
2287_bone_met CHR 18
scp2_bone_met CHR 18
scp46_bone_met CHR 18
1834_lung_met CHR 18
3481_lung_met CHR 18
scp3_lung_met CHR 18
scp43_lung_met CHR 18
BAF
Position

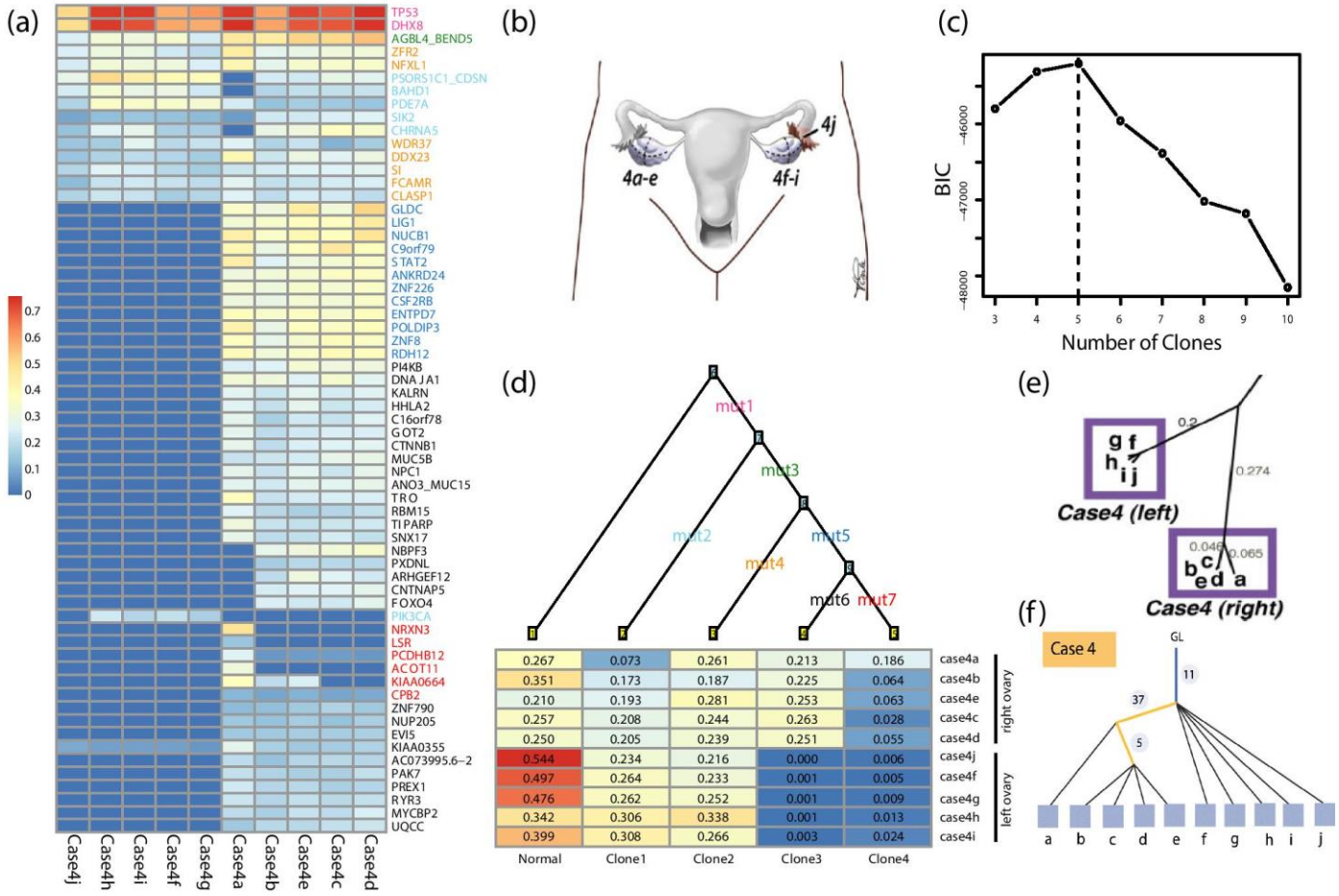
(d)


231_parental_cell_line CHR 19
1833_bone_met CHR 19
2287_bone_met CHR 19
scp2_bone_met CHR 19
scp46_bone_met CHR 19
1834_lung_met CHR 19
3481_lung_met CHR 19
scp3_lung_met CHR 19
scp43_lung_met CHR 19
BAF
Position

Supplementary Figure S14. Clonal history reconstructed from primary tumor and the relapse genome of leukemia patients. (a) VAFs of SNAs and indels of the primary tumor and the relapse genome of patient AML43 and AML1 are clustered into mutational waves shown in different colors. A mixture component with a small weight shown as pink dots is included to gain robustness against false positives. CNAs for each mutational cluster are profiled. SNAs and CNAs are used as input for Canopy. **(b)** Plausible phylogenies inferred by Canopy, observed at two time-points. Mutations and clonal proportions are shown in red and blue respectively. Both trees support the model that a subclone from the primary tumor gains additional mutations and expands at relapse. **(c)** Inference of clonal frequency from the posterior distribution. One subclone survives the chemotherapy and becomes dominant. Normal cell contaminations/tumor purities are estimated as the first columns.



Supplementary Figure S15. Clonal history reconstructed from ten spatially separated samples. Ten ovarian cancer tumor samples from different regions (4a-4e, right ovary; 4f-4i, left ovary; 4j, left fallopian tube) from case 4 in Bashashati *et al.* (6) are whole exome sequenced. 63 mutations are confirmed by deep amplicon resequencing. **(a)** Heatmap of mutational profiling across 63 genes, 10 samples. **(b)** Anatomical sites of the ten spatially separated samples. **(c)** BIC as a model selection metric to determine the number of subclones. **(d)** The most likely tree returned by Canopy based on the mutational profiling. Mutations in blue are additionally acquired by the right ovary samples from the left ovary samples and drive the divergence. Mutations in red further distinguish case4a from the rest of the samples from the right ovary. Each sample offers a snapshot of different combinations of the subclones that is correlated with their spatial distribution. **(e)** Tree reconstructed by Bashashati *et al.* (6) by a nearest neighbor method. **(f)** Tree reconstructed by Popic *et al.* (7) as an acyclic directed graph. Both methods put samples at the tree leaves as homogeneous populations.



Supplementary Table S1. Cancer genomic studies by sequencing multiple samples from the same patients. Multiple types of cancer were sequenced by different platforms by a longitudinal (multi-time point) or a spatial experimental (multi-region) design. For bulk-tissue sequencing, 5 to 12 samples were sequenced from the same individual; for single-cell sequencing, ~100 single cells were sequenced. Across all studies, less than 8 cancer clones were identified.

Literature	Cancer type	Sequencing	Number of tumor samples from the same individual	Number of clones
Navin *et al.*, 2011 (10)	Breast cancer	Single-cell	100 single cells	5
Ding *et al.*, 2012 (3)	Acute myeloid leukaemia	Whole-genome	2 bulk samples	2-5
Bashashati *et al.*, 2013 (6)	Ovarian cancer	Whole-exome + deep amplicon	10 bulk samples	NA*
Gerlinger *et al.*, 2014 (11)	Clear cell renal cell carcinoma	Whole-exome + ultra deep	5-10 bulk samples	NA*
Zare *et al.*, 2014 (9)	Breast cancer	Whole-exome + targeted	12 bulk samples	4-6
Eirew *et al.*, 2015 (12)	Breast cancer xenografts	Whole-genome + targeted + single-cell	11 bulk samples and 90 single cells (SA501)	5
Sottoriva *et al.*, 2015 (13)	Colorectal adenomas and carcinomas	Whole-exome + targeted; single-cell FISH	On average 23 tumor glands and 2 bulk samples	2-7
Boutros *et al.*, 2015 (14)	Prostate cancer	Whole-genome	4 bulk samples (CPCG0183)	<6

NA*: Bashashati *et al.* (6) and Gerlinger *et al.* (11) constructed phylogenetic tree by neighbour-joining and maximum parsimony method and put bulk-tumor samples as tree leaves.

Supplementary Table S2. SNA and CNA input for PhyloWGS in simulation. The true underlying phylogeny is shown in Figure 1a. The data input for Canopy is shown in Figure 1b and is included in the Canopy R package. **(a)** The SNA input is the same as that for Canopy. **(b)** We are able, on this toy data set, to convert the CNA events to the pseudo-SNA events required by PhyloWGS by using the true clonal proportions of the CNA events from the ground truth (this proportion is equal to a/d for the CNAs). Read depth (d) is set at 100 for the conversion to be on par with the SNAs. The estimated phylogeny for Canopy and PhyloWGS is shown in Supplementary Figure S5c.

(a) SNA input (ssm_data.txt):

id	**gene**	**a**	**d**	**mu_r**	**mu_v**
s0	sna1	81,74,101	94,102,101	0.999	0.5
s1	sna2	82,102,71	99,102,103	0.999	0.5
s2	sna3	42,47,46	98,96,97	0.999	0.5
s3	sna4	74,76,74	108,101,99	0.999	0.5

(b) CNA input (cnv_data.txt):

cnv	a	d	ssms
c0	40,50,50	100,100,100	s0,0,2
c1	80,80,70	100,100,100	s1,1,1
c2	40,50,50	100,100,100	s2,1,2

Supplementary Table S3. Running time and estimation error with and without pre-clustering step. Simulation is carried out with varying number of mutations $N \in \{25,50,100,200\}$ along trees with different number of branches $K \in \{3,4,5,6\}$ from three samples. Canopy is run with and without a Binomial clustering procedure (C for clustering and NC for non-clustering) as an initialization step for MCMC. Convergence is measured by both the log-likelihood and the acceptance rate. Run time is measured in seconds; estimation error of the genotyping matrix Z is measured as the percentage of wrongly labeled elements; RMSE is used to measure the estimation error of the clonal proportion matrix P. Pre-clustering step significantly reduces computation time for larger number of mutations and results in comparable or smaller estimation errors.

Number of mutations N	Number of branches K	Run time (C, sec)	Run time (NC, sec)	Z error (C)	Z error (NC)	P error (C)	P error (NC)
25	3	84.1	57.0	0	0	0.003	0.003
	4	124.6	87.3	0.006	0.003	0.008	0.005
	5	142.0	126.0	0.022	0.019	0.01	0.008
	6	180.1	143.2	0.025	0.023	0.012	0.011
50	3	145.6	134.8	0	0	0.003	0.003
	4	235.3	295.6	0.013	0.009	0.013	0.007
	5	191.1	360.4	0.015	0.02	0.009	0.013
	6	261.2	429.5	0.019	0.019	0.013	0.009
100	3	348.5	436.5	0.003	0.005	0.005	0.005
	4	374.7	911.2	0.012	0.012	0.016	0.009
	5	334.9	1011.7	0.011	0.016	0.009	0.011
	6	372.6	1191.3	0.012	0.016	0.008	0.011
200	3	498.2	1463.6	0.002	0.007	0.007	0.011
	4	512.6	2454.9	0.008	0.012	0.017	0.015
	5	558.0	2871.0	0.009	0.014	0.010	0.017
	6	643.1	3580.9	0.010	0.013	0.014	0.014

Supplementary Table S4. Fixed clonal frequency matrix *P* in simulation studies. True underlying clonal frequency matrix for different number of samples **(a)** and different number of subclones **(b)**. The elements are chosen so that the additive summation result is the most distinct in the unit space and that different combinations of subclones are present across different samples.

(a)

	Sample 1	**Sample 2**	**Sample 3**	**Sample 4**	**Sample 5**	**Sample 6**	**Sample 7**	**Sample 8**	**Sample 9**	**Sample 10**	**Sample 11**
Clone1	1/31	1/15	1/15	1/15	1/15	1/7	1/7	1/7	1/7	1/7	1/7
Clone2	2/31	2/15	0	2/15	2/15	0	0	0	2/7	2/7	2/7
Clone3	4/31	4/15	2/15	0	4/15	0	2/7	2/7	0	0	3/7
Clone4	8/31	8/15	4/15	4/15	0	2/7	0	4/7	0	4/7	0
Clone5	16/31	0	8/15	8/15	8/15	4/7	4/7	0	4/7	0	0

(b)

	Sample1	**Sample2**	**Sample3**
Clone 1	$1/(2^k\text{-}1)$	$1/(2^{k-1}\text{-}1)$	$1/(2^{k-1}\text{-}1)$
Clone 2	$2/(2^k\text{-}1)$	0	$2/(2^{k-1}\text{-}1)$
Clone 3	$4/(2^k\text{-}1)$	$2/(2^{k-1}\text{-}1)$	0
…	…	…	…
Clone k	$2^{k-1}/(2^k\text{-}1)$	$2^{k-2}/(2^{k-1}\text{-}1)$	$2^{k-2}/(2^{k-1}\text{-}1)$

Supplementary Table S5. Metastatic outcomes and cell population types of MDA-MB-231 and its sublines.

	Cell type	**Metastatic outcome**
MDA-MB-231	Parental line	-
1833	Mixed-cell subline (MCP)	Bone
2287	Mixed-cell subline (MCP)	Bone
SCP2	Single-cell subline (SCP)	Bone
SCP46	Single-cell subline (SCP)	Bone
1834	Mixed-cell subline (MCP)	Lung
3481	Mixed-cell subline (MCP)	Lung
SCP3	Single-cell subline (SCP)	Lung
SCP43	Single-cell subline (SCP)	Lung

Supplementary Table S6. Transition probabilities of the HMM to profile allele specific copy number in SCP samples. Let $p = 0.005$ be the transition probability between states that need one change (e.g. from LOH to Neutral) and $r = 0.001$ be the probability between states that need more than one change (e.g. from LOH to TwoThree). The values for p and r are arbitrarily set but can be optimized via the EM algorithm.

From ↓ To ⟶	LOH	Neutral	OneTwo	OneThree	OneFour	TwoThree
LOH	$1-4p-r$	p	p	p	p	r
Neutral	p	$1-3p-2r$	p	r	r	p
OneTwo	p	p	$1-3p-2r$	p	r	r
OneThree	p	r	p	$1-4p-r$	p	p
OneFour	p	r	r	p	$1-2p-3r$	r
TwoThree	r	p	r	p	r	$1-2p-3r$

Supplementary Dataset S1. Somatic point mutations identified and annotated in the MDA-MB-231 metastasis model system. SNAs are called by the Genome Analysis Toolkit's UnifiedGenotyper (2) and are annotated by ANNOVAR (15). Separately attached.

Supplementary Dataset S2. SNA and CNA input from Eirew *et al.* (12). SNA input is scaled from deep amplicon sequencing reads to avoid binomial overdispersion. Input for SNA cluster is taken as the median value of all SNAs within the same cluster. CNA input is obtained by refining and combining TITAN's allele-specific segmentation results. Separately attached.

Reference

1. Chen H, Bell JM, Zavala NA, Ji HP, & Zhang NR (2015) Allele-specific copy number profiling by next-generation DNA sequencing. *Nucleic acids research* 43(4):e23.
2. Van der Auwera GA*, et al.* (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* 11(1110):11 10 11-11 10 33.
3. Ding L*, et al.* (2012) Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 481(7382):506-510.
4. Roth A*, et al.* (2014) PyClone: statistical inference of clonal population structure in cancer. *Nature methods* 11(4):396-398.
5. Miller CA*, et al.* (2014) SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS computational biology* 10(8):e1003665.
6. Bashashati A*, et al.* (2013) Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *The Journal of pathology* 231(1):21-34.
7. Popic V*, et al.* (2015) Fast and scalable inference of multi-sample cancer lineages. *Genome biology* 16(1):91.
8. Li SY, Pearl DK, & Doss H (2000) Phylogenetic tree construction using Markov chain Monte Carlo. *J Am Stat Assoc* 95(450):493-508.
9. Zare H*, et al.* (2014) Inferring clonal composition from multiple sections of a breast cancer. *PLoS computational biology* 10(7):e1003703.
10. Navin N*, et al.* (2011) Tumour evolution inferred by single-cell sequencing. *Nature* 472(7341):90-94.
11. Gerlinger M*, et al.* (2014) Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature genetics* 46(3):225-233.
12. Eirew P*, et al.* (2015) Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* 518(7539):422-426.
13. Sottoriva A*, et al.* (2015) A Big Bang model of human colorectal tumor growth. *Nature genetics* 47(3):209-216.
14. Boutros PC*, et al.* (2015) Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nature genetics* 47(7):736-745.
15. Wang K, Li M, & Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* 38(16):e164.