# A Phylogenetic Approach to Inferring the Order in Which Mutations Arise during Cancer Progression

Yuan Gao[1*], Jeff Gaither[2], Julia Chifman[3] and Laura Kubatko[4,5,6]

*Correspondence:
gao.957@osu.edu
[1]Division of Biostatistics, The Ohio State University, 1958 Neil Ave, 43210 Columbus, US
Full list of author information is available at the end of the article

**Abstract**

Although the role of evolutionary processes in cancer progression is widely accepted, increasing attention is being given to the mechanisms by which variation in these processes across patients and across cancer types can lead to differences in clinical outcomes. Recent studies suggest that the temporal order in which somatic mutations accumulate during cancer progression is important. Single-cell sequencing provides a unique opportunity to examine the effect that mutation order has on cancer progression and treatment effect. However, the error rates associated with single-cell sequencing complicate this task. We propose a new method for inferring the order in which somatic mutations arise within a tumor using noisy single-cell sequencing data that incorporates the errors that arise from the data collection process. Through analyses of simulations and of empirical data from cancer patients, we show that our method outperforms existing methods for identifying mutation order and leads to new insights about the evolutionary trajectories of cancer.

**Keywords:** Cancer evolution; Single-cell sequencing; Cancer phylogenetics; Mutation order; Bayesian inference

## Background

Cancer progression is a dynamic evolutionary process that occurs among the individual cells within each patient's tumor. Cancer develops from a single cell in normal tissue whose genetic alterations endow a growth advantage over the surrounding cells, allowing that cell to replicate and to expand, resulting in the formation of a clonal population of identical cells. Cells within this clonal population, or clone, may then undergo their own somatic mutations, followed by replication and formation of subclones. During this complex process, many competitive and genetically diverse subpopulations may be formed, resulting in which is known as intratumoral heterogeneity (ITH) [1–5]. ITH is common in human cancers and brings great challenges to the diagnosis and treatment of cancer [6, 7]. For example, Ortmann et al. (2015) showed that the type of malignancy and the response to treatment for myeloproliferative neoplasm patients were affected by the order in which somatic mutations arose within the patients' tumors. Though this study was specific to one type of cancer, the timing and organization of somatic mutations may be crucial to clinical outcomes for other cancers as well. Studying cancer genome diversity, understanding fitness landscapes of neoplastic cells, and determining the number and order of mutations required for tumor progression are thus critical, especially in the field of targeted therapeutics.

Fig 1a is a pictorial representation of the evolution of a tumor from the first initiating somatic mutation. In the process of cancer-associated somatic mutations accumulating and clones evolving, selective pressure may favor one or more of these clones, ultimately resulting in ITH and cancer. Fig 1b is a representation of the tissue at the time of sampling, in which clones with different genotypes residing at different locations within the tumor are present in the sample of cells. Knowledge of the evolutionary history of the tumor, especially the order in which the somatic mutations arose, would allow for understanding of a tumor's ITH and perhaps aid in treatment decisions. However, this information cannot be observed directly, since genomic data is most often collected from one snapshot in time. Consequently, use of computational methods that infer the order of mutations from DNA sequence data is the approach of choice.

Most studies on cancer phylogenetics utilize bulk high-throughput sequencing data, but signals obtained from bulk sequencing only reflect the overall characteris-

tics of a population of sequenced cells, rather than the characteristics of individual cells within the tumor. Variation in the mutational signatures among different cells in a tumor is thus difficult to evaluate when bulk sequencing data is used. Single-cell sequencing (SCS) data is promising because it enables sequencing of individual cells, thus provides the highest possible resolution in data available on the mutational history of cancer. However, the high error rates associated with SCS data complicate the development of methods for inference of the mutational history. This is because the whole-genome amplification (WGA) process, a crucial step in producing SCS data, results in a variety of errors. Some well-known sources of error include allelic dropout (ADO) errors, false positive errors (FPs), non-uniform-coverage distribution, and low-coverage regions [9]. ADO is the main source of error in SCS data, contributing a large number of false negative errors (FNs) in inferring the presence of point mutations.

Earlier studies have proposed various mathematical methods for the reconstruction of tumor phylogenies (Fig 1c - 1e) from data arising from single-cell somatic mutations. Of particular interest are the methods of Jahn et al. (2016) and Zafar et al. (2017), called SCITE and SiFit, respectively. SCITE uses an MCMC algorithm to infer a mutation tree from the imperfect SCS data. SiFit introduces a component-wise MCMC algorithm to infer a tumor phylogenetic tree with the imperfect SCS data as a first step. Based on the inferred tumor phylogenetic tree, SiFit then estimates the mutation order by inferring the mutation status of the internal nodes.

In this paper, we propose a novel method for inferring the order in which mutations arise within an individual tumor given SCS data from the tumor at a single time point. Our approach utilizes models for both the mutational process within the tumor and the errors that arise during SCS data collection in a Bayesian framework, thus allowing us to quantify the uncertainty in the inferred mutation orders.

## Results and discussion

### Overview of mutation order inference from single-cell sequencing data

We assume that we are given a phylogenetic tree that displays the evolutionary relationships among a sample of $J$ cells within a tumor. To infer the locations (branches) on which a set of somatic mutations are acquired in the tree, we need

to model the evolutionary process of these somatic mutations and quantify the technical errors that arise from the SCS data collection process. We assume that during the evolutionary process, somatic mutations evolve independently across sites, and each mutation evolves independently on different branches in the tree. We also assume that each somatic mutation occurs once along the phylogeny and that no back mutation occurs, so that all descendant cells linked by the mutation branch will harbor the corresponding mutation. When quantifying the effect of error, we assume that SCS technical errors for mutations are independent of one another.

### *Notation and terminology*

Consider somatic mutations of interest at $I$ loci across the genome for a sample of $J$ single cells. The $J$ single cells are sampled from different spatial locations (clones) within the tumor. The mutation data can be either binary or ternary. For binary data, 0 denotes the absence of mutation and 1 means that mutation is present, while for ternary data, 0, 1 and 2 represent the homozygous reference (normal), heterozygous (mutation present) and homozygous non-reference (mutation present) genotypes, respectively.

The $I$ somatic mutations evolve along the tumor phylogenetic tree $\mathcal{T}$. Each tip in $\mathcal{T}$ represents one single cell $C_j$, where $j = 1, \ldots, J$. Let $C = \{C_1, \ldots, C_J\}$ be the set of the $J$ single cells under comparison. $\mathcal{T} = (T, \mathbf{t})$ includes two parts: the tree topology $T$ and a vector of branch lengths $\mathbf{t}$. The tree topology $T = (V, E)$ is a connected graph without cycles and is composed of nodes and branches, where $V$ is the set of nodes and $E$ is the set of branches. Here we assume that trees are rooted, and the root $r$ represents the common ancestor (a normal cell without mutations) for all of the single cells under comparison. In the context of this paper, we will focus on rooted bifurcating trees and all of the definitions in the following sections will apply to rooted bifurcating trees. There are $2J-2$ branches in a rooted bifurcating tree with $J$ tips, i.e., $E = \{e_1, e_2, \ldots, e_{2J-2}\}$. Let $v$ and $w$ be two nodes in the node set $V$ that are connected by the branch $x$ in the branch set $E$ (i.e., $x = \{v, w\}$: $v$ is the immediate ancestor node of $w$, and $x$ connects $v$ and $w$). Then the set $U^x(w)$, which includes the node $w$ and all nodes descended from $w$ in $\mathcal{T}$, is called the *clade induced by* $w$. The branch $x$ connects the ancestor node $v$ and the clade induced

by $w$, and we define branch $x$ as the *ancestor branch of clade* $U^x(w)$. $E^x(w)$ is a subset of $E$ that includes branches connecting nodes in $U^x(w)$, and $C^x(w)$ are the tips in $U^x(w)$.

Let $G_{ij}$ denote the true genotype for the $i^{th}$ genomic site of cell $C_j$. The $i^{th}$ genomic site will then have a vector $\mathbf{G}_i \in \{0,1\}^J$ (for binary data) or $\{0,1,2\}^J$ (for ternary data) representing its true genotype for all of the $J$ cells represented by the tips in the tree, where $i = 1, \ldots, I$. Let $S_{ij}$ denote the observed genotype for the $i^{th}$ genomic site of cell $C_j$. Due to the technical errors associated with SCS sequencing, the true genotype $G_{ij}$ might be different from the observed data $S_{ij}$, i.e., $S_{ij}$ does not always equal to $G_{ij}$. For both binary and ternary data, the observed state $S_{ij}$ might be flipped with respect to the true mutation $G_{ij}$ due to false positive errors or false negative errors. Missing states ("-") or low-quality states ("?") may be present for some genomic sites, as well. Fig 2 shows an example of true and observed binary genotype data for the mutations in Fig 1. In Fig 2, the observed state is marked with red color if it is not consistent with the true genotype. The red numbers are those mutations with flipped observed mutation states relative to the true mutation state. The red dash ("-") indicates a missing value and the red question mark ("?") represents a low-quality value.

Mathematically, we represent the observed mutation states of the $J$ single cells at $I$ different genomic sites by a $I \times J$ mutation matrix $\mathbf{S}$ for convenience,

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \ldots \\ \mathbf{S}_I \end{pmatrix} = \begin{pmatrix} S_{11} & \ldots & S_{1J} \\ S_{21} & \ldots & S_{2J} \\ \ldots & \ldots & \ldots \\ S_{I1} & \ldots & S_{IJ} \end{pmatrix}. \tag{1}$$

Each entry $(i, j)$ denotes the state observed for mutation $i$ in cell $C_j$, so $\mathbf{S}_i$ gives the observed data for genomic site $i$ as a vector with $J$ values corresponding to the $J$ single cells. Column $j$ represents the mutations of interest for cell $C_j$. In phylogeny $\mathcal{T}$, let $\mathcal{B}$ be the vector of locations (branches) on which the $I$ mutations occur, i.e., $\mathcal{B} = \{B_1, \ldots, B_I\}$, where $B_i$ is the location (branch) on which mutation $i$ is acquired. Note that $B_i$ takes values in $\{e_1, e_2, \ldots, e_{2J-2}\}$.

### *Somatic mutation process*

To model the somatic mutation process, we consider continuous-time Markov processes, which we specify by assigning a rate to each possible transition between states [12]. The mutation processes along each branch of $\mathcal{T}$ are assumed to be independent of each other. We consider point mutations that are not affected by deletion or loss of heterozygosity (LOH). Once a mutation $i$ is acquired on a branch $x \in E$, all the branches in the set $E^x(w)$ will harbor mutation $i$ but those branches in the set $E \backslash (x \cup E^x(w))$ will not carry this mutation. As an example, Fig 3 depicts the observed and true binary genotype for mutation $i = 1$ shown in Fig 2. The set of branches is $E = \{e_1, \ldots, e_8\}$ and the corresponding set of branch lengths is $\mathbf{t} = \{t_1, \ldots, t_8\}$. If mutation $i$ is acquired on branch $e_1$, the cell descending along branch $e_8$ will not carry the mutation, while those descending from the blue branches would carry this mutation.

### *Observation of SCS errors*

In real data, the data in the observed mutation matrix is subject to error. To account for false positives and false negatives in the SCS data, our method applies the error model from [10, 11, 13]. Following their notation, we let $\alpha_{ij}$ be the probability of a false positive error and $\beta_{ij}$ be the probability of a false negative error for genomic site $i$ of cell $C_j$. For binary data, if the true genotype is 0, we may observe a 1, which is a FP. If the true genotype is 1, we may observe a 0, which is a FN. The conditional probabilities of the observed data given the true genotype at genomic site $i$ of cell $C_j$ can be stored in an error probability matrix $\mathbf{N}^{ij}$. Please see the Methods section for details.

### *Inferring the location of each mutation in $\mathcal{T}$*

Once the observed status matrix $\mathbf{S} = [\mathbf{S}_1 \ldots \mathbf{S}_I]^T$ of the $I$ mutations has been collected, the next step is to infer the branch on which mutation $i$ takes place, conditioning on $\mathbf{S}$. Given the observed data matrix $\mathbf{S}$, the tumor phylogenetic tree $\mathcal{T}$, the error probability matrix $\mathbf{N} = \{\mathbf{N}^{ij} | 1 \leq i \leq I, 1 \leq j \leq J\}$, and the mutation process $\mathcal{M}_\lambda$, we can assign a posterior probability distribution $\mathbf{P}(B_i | \mathbf{S}, \mathcal{T}, \mathbf{N}, \mathcal{M}_\lambda)$ to the location of mutation $i$ using Bayes' Theorem, i.e, we can compute the probability that mutation $i$ occurs on each of the $2J - 2$ branches. To obtain a point estimate of the branch on which mutation $i$ occurs, we pick the branch that max-

imizes this posterior probability, i.e., the maximum a posteriori (MAP) estimate. The MAP estimator for the location of mutation $i$ is given by

$$\hat{B}_{i_{MAP}} = \underset{B_i \in \{e_1,\ldots,e_{2J-2}\}}{\operatorname{argmax}} \mathbf{P}(B_i|\mathbf{S},\mathcal{T},\mathbf{N},\mathcal{M}_\lambda). \tag{2}$$

### *Inferring the mutation order in $\mathcal{T}$*

We now consider the posterior probability distribution of the locations for the $I$ mutations in the sample of $J$ single cells, which is a distribution on a set of cardinality $(2J-2)^I$. Based on the assumption of independence among the $I$ mutations being considered, the posterior distribution for $\mathcal{B}$ is given by

$$\mathbf{P}(\mathcal{B}|\mathbf{S},\mathcal{T},\mathbf{N},\mathcal{M}_\lambda) = \prod_{i=1}^{I} \mathbf{P}(B_i|\mathbf{S},\mathcal{T},\mathbf{N},\mathcal{M}_\lambda), \tag{3}$$

From this distribution, we can extract information on the ordering of mutations of interest. Please see the Methods section for details.

## Simulation Study

### *Design of the simulation study*

To evaluate the ability of our method, which we call MO (Mutation Order), to correctly identify the locations of a set of mutations and the order of these mutations under different conditions, we conduct a simulation study with 192 different settings. The goal of our simulation study is to assess the effect of data quality (complete or incomplete, high or low error rates), branch lengths in the tumor phylogeny, and type of genotype data on the performance of our method. The simulation consists of the following steps:

1. Simulate 1000 random bifurcating tumor phylogenetic trees with $J$ cells in each tree. Each tree is generated by the recursive random splitting algorithm in the R package ape [14].
2. Repeat steps (2.a)-(2.d) on each of the 1000 simulated trees from step (1).
   (a) Simulate one mutation on each branch of a fixed tree from step 1, based on the three possible mutation generation procedures in the Methods

section. Record the perfect true ternary genotype data for the $J$ cells for each mutation.

(b) Convert the simulated true ternary genotype data (1 and 2) from step (2.a) into binary genotype data (1). Record the complete true binary genotype data.

(c) Add noise to the true genotype data with error rates $\alpha$ and $\beta$, and record the observed data for each mutation. The noise is added to each cell independently.

(d) Randomly select tips in the tumor tree with a prespecified missing data percentage and delete the observed data for these tips. Record the incomplete observed ternary and binary data.

3 Repeat step (2) with different choices of branch length distributions, error rates ($\alpha$ and $\beta$) and missing data percentages.

Given the above simulation procedure, there are several options at each step. In step (1), we need to specify the branch length distribution in the bifurcating tumor phylogenetic tree and the number of tips in each tree. We let each tree have 10 tips and consider two scenarios (labeled scenario 1 and scenario 2) in which branch lengths follow the exponential distribution with mean 0.2 and mean 0.1, respectively. The lengths of branches sampled from exponential distribution with mean 0.2 are longer, corresponding to a longer evolutionary time within the tumor, while those sampled from exponential distribution with mean 0.1 are shorter, indicating shorter evolutionary times. Step (1) returns 1000 simulated trees in each of scenario 1 and scenario 2.

We repeat step (2) for each scenario. In step (2.a), we specify the evolutionary parameters for these simulated cancer-related sites, and consider $\lambda_1 = 10^{-7}$ and $\lambda_2 = 10^{-2}$ from [12]. For a mutation on each branch, we draw a sample from three possible conditions ($0 \to 1$, $0 \to 2$ or $0 \to 1$ then $1 \to 2$) based on their corresponding marginal probabilities. Once the generating mechanisms for all mutations are determined, we have the true genotype data for all mutations. In step (2.c), we add noise to each tip with error rates $\alpha$ and $\beta$ independently based on Equation 17. In step (2.d), we choose either 10% or 20% of the tips and delete their data. These selected tips have missing values at the selected loci.

We repeat step (3) 16 times, each with a unique choice of ($\alpha$,$\beta$), where $\alpha \in \{0.05, 0.1, 0.2, 0.4\}$ and $\beta \in \{0.05, 0.1, 0.2, 0.4\}$. In total, we have 192 settings with different branch length distribution (exponential distribution with mean 0.2 or 0.1), error rates ($\alpha \in \{0.05, 0.1, 0.2, 0.4\}$ and $\beta \in \{0.05, 0.1, 0.2, 0.4\}$), and missing data percentage (no missing data, 10% or 20% of data missing) and type of genotype data (binary and ternary).

Within each tree in a simulation setting, we apply MO to estimate the location for each mutation. We evaluate the accuracy of the MAP estimate across the 1000 trees within each setting in three aspects: the accuracy of the inferred branch on which the mutation occurs, the accuracy of the inferred order for pairs of mutations that occurs on adjacent branches in the tree, and the accuracy of the inferred order for any pairs of mutations in the tree. In addition to the accuracy of the MAP estimate, we also evaluate the accuracy of the credible set for the inferred branch from MO.

### *Accuracy of MAP estimate*

We assessed the accuracy of the MAP estimate in several ways, including whether the mutation was inferred to occurr on the correct branch ("location accuracy"), whether a pair of mutations that occur on adjacent branches were inferred to occur in the correct order ("adjacent order accuracy"), and whether any pair of mutations were inferred to occur in the correct order ("order accuracy"). In all cases, the accuracy of MO is very high except when error rates are high. Tables S1 and S2 in Additional file 1 show location accuracy with each cell entry corresponding to a unique setting of $\alpha$ and $\beta$, type of genotype and missing data percentage. All cells in a row have the same $\alpha$ and $\beta$, and all cells in a column have the same type of genotype and missing data percentage. Within one table, the accuracy of MAP estimation with ternary genotype data tends to be higher than that of binary genotype data when fixing other parameters. With the same type of genotype (binary or ternary) in each row (same error rate setting), the accuracy decreases as the percentage of missing values increases. When $\alpha$ (or $\beta$) is fixed, accuracy tends to decrease as $\beta$ (or $\alpha$) increases. With the same type of genotype in each row (same error rate setting), the accuracy of the inference decreases as the percentage of missing values increases. Comparing Tables S1 and S2, the accuracy of MAP

estimation in Table S2 (shorter branch lengths) tends to be higher than that of Table S1 (longer branch lengths) within the same setting with a few exceptions for large $\alpha$ and $\beta$, although the difference is very small.

The results for adjacent order accuracy (Tables S3 and S4 in Additional file 1) and order accuracy (Tables S5 and S6 in Additional file 1) are similar. In addition to the same overall trends due to data type, percentage of missing data and error rates, the magnitude of the order accuracies in Tables S5 and S6 are higher than the corresponding adjacent order accuracies in Tables S3 and S4. Please see the Methods section for details.

### *Credible set accuracy*

In addition to the MAP point estimation, the credible set accuracy of the inferred mutation branch is assessed, as well. The credible set describes and summarizes the uncertainty related to the mutation order. If the true mutation branch is among the credible set, we count this as correct; otherise, it is incorrect. We use 95% credible set for computation and the results are presented in Tables S7 and S8 in Additional file 1. The effects of the error rates, missing data percentages and genotype are similar to point estimation. The magnitude of credible set accuracy is much higher than the corresponding MAP point estimation, especially for those with large error rates and higher missing data percentages. Please see the Methods section for details.

### *Comparison with competing approaches*

To further assess the performance of MO, we compare the performance of the methods SCITE [10] and SiFit [11] on the simulation data. SCITE computes the order of mutation with either binary or ternary genotype data. SiFit can use either binary or ternary genotype data when inferring the phylogenetic tree in the first step, but it only uses binary genotype data when inferring mutation order.

The three methods were compared with respect to the order accuracy and adjacent order accuracy under different simulation settings. Fig 4 and Fig 5 plot the adjacent order accuracy of the three methods in scenario 1 and scenario 2, respectively. Fig 6 and Fig 7 plot the order accuracy of the three methods in scenario 1 and scenario 2, respectively. In all four figures, each panel plots results from the specific type of genotype data and missing data percentage. The top row is for binary data and the bottom row is for ternary data. The three columns are complete data, incomplete

data with 10% missing values, and incomplete data with 20% missing values, from left to right. In each panel, red color corresponds to MO, blue color corresponds to SCITE and green color corresponds to SiFit. The x-axis is FNR $\beta$, and different point shapes represent different FPR $\alpha$. The y-axis is either order accuracy or adjacent order accuracy.

In both scenarios, we observe that the adjacent order accuracies and order accuracies decrease as the data quality becomes worse (error rates increase or missing data percentage increases) for all three methods. MO is superior to SCITE in all settings in terms of adjacent order accuracy and order accuracy estimated from ternary genotype data. For the results estimated from the binary genotype, the performance of MO is better than SCITE when data are complete and error rates are small. When $\alpha$ is fixed and $\beta$ is increased from 0.1 to 0.2, the performance of MO is better than or similar to SCITE. Only when $\beta$ is fairly large (e.g., $\beta = 0.4$) does SCITE have slightly higher accuracy rates than MO. As the percentage of missing values increases, MO still performs better than SCITE in most cases. In all of the settings, SiFit has the worst performance since it only computes the orders of some of the input mutations, and loses too much information. Although the output of partial mutation order from SiFit are mostly correct, the accuracy is low due to the small number of inferred paired mutation orders. MO thus dominates SiFit when assessing the performance using order accuracy and adjacent order accuracy.

## Empirical examples

We apply MO to two experimental single-cell DNA sequencing datasets, one study that contains data for two prostate cancer patients [15] and another that contains data for two metastatic colon cancer patients [16]. For the prostate cancer patients, we retrieve a publicly available sequencing dataset from the single-cell study of [15], which includes 20 single-cell genomes. For the colon cancer patients, we use the somatic single nucleotide variants (SNVs) after variant calling provided in the original study (16 SNVs for patient 1 and 36 SNVs for patient 2) [16]. The details of tumor tree construction are in the Methods section. We describe the analysis for each study below.

### *Prostate cancer patients*

The data for the first prostate cancer patient, labeled P1, includes one normal cell and nine tumor cells, and the data for the second patient, labeled P2, includes two normal cells and eight tumor cells. 80 common tumor suppressor genes (TSGs) and oncogenes are selected for both P1 and P2. Fig 8 and Fig 9 show the tumor phylogenies estimated for P1 and P2, respectively. For P1, 8 tumors cells were extracted from location $X3$ of the tissue, and 1 tumor cell was from another location labeled $X4$. For P2, 2 tumor cells were obtained from the right side ($R3$) of the tissue and 6 were from the left side (these locations are labeled $L3$ and $L4$) of the tissue. Among the 8 cancer cells for patient P2, cancer cells $R3.C1$ and $R3.C2$ from the right side are clustered in a subclone, and the rest of the cells from the left side are clustered in another subclone that diverges from that lineage. More importantly, $L3.C2$ and $L4.C3$ are grouped together, indicating the close evolutionary relationship between these two cells, although they are from different locations on the left side. Even though $L3.C2$ and $L4.C3$ are grouped together, the precise locations of all cells are unknown to us and $L3.C2$ might be closer to $L4.C3$ spatially than other cells from $L3$. It is also possible that the location measurement of single cells is not precise due to limitations of the current technology. In both tumor phylogenies, the trunk connects the tumor clone to the normal clone. We annotate the gene mutations of interest on the branches of tumor trees in Fig 8 and Fig 9. The uncertainty of the mutation locations are highlighted with colors. Those genes with strong signal (defined to be a probability larger than 0.7 that the mutation occurred on a single branch) are marked as red, while those genes with moderate signal (defined to a probability larger than 0.7 on two or three branches) are marked as blue.

We also compare the estimated posterior probability distributions for each mutation for patients P1 and P2, which are used to construct credible sets and measure the uncertainty of the inferred mutation order. For P1, using binary and ternary data results in an estimate of 43 and 36 gene mutations, respectively, which are inferred to occur on the trunk of the tumor tree. Of these, 34 genes are estimated to occur on the trunk of tree regardless of which data type is used. For P2, using binary and ternary data results in an estimate of 52 and 48 gene mutations, respectively, which are inferred to occur on the trunk of the tumor tree. Of these, 44 genes are estimated to mutate on the trunk of the tumor tree regardless of which

data type is used. Figures S1 and S2 in Additional file 1 are heatmaps for the posterior probability distribution of each mutation for patients P1 and P2, where the $x$-axis corresponds to the branch number and the $y$-axis corresponds to the gene. The color in each block of the heatmap represents the conditional probability of the corresponding mutation on that branch, where red indicates the largest probability and blue indicates the smallest.

In addition to the common cancer-associated genes across different cancers, we map the mutations of prostate cancer specific genes [17, 18]. Fig 12 - Fig 15 are heatmaps of the estimated posterior probabilities for these genes. Gene *ZFHX3* has a high probability of having mutated on the trunk of the tree in both patients, which might have important implications for future study. In addition, the data for patient P1 shows a very strong signal that *FOXP1* mutate on the trunk of the tumor tree, while *BRCA2* has high probability of having mutated on the trunk of the tumor tree for patient P2. Comparing the heatmap of common cancer-associated genes with the prostate cancer specific genes, mutations inferred to have occurred on the trunk of the tree tend to be those that are common across cancer types, while mutations known to have high frequency within prostate cancer are generally found closer to the tips of the tree in both patients.

### *Metastatic colorectal cancer patients*

For the colorectal cancer study of [16], data are available for two patients, labeled CRC1 and CRC2, consisting of binary genotypes that represent the presence or absence of mutation at each of the SNV sites. The inferred tumor tree and mutation order are depicted in Fig 10 and Fig 11. The uncertainty of locations of mutations are highlighted with colors. For patient CRC1, *APC*, *LINGO2* and *IL21R* are inferred to have mutated on the branch linking the subclone mainly consisting of MA (metastatic tumor cells) and PA (primary tumor cells) with high confidence (probability close to 1). For patient CRC2, mutations of *APC*, *KRAS* and *TP53* are inferred to have been acquired on the trunk of the tumor phylogeny with posterior probability close to 1, and we have very high confidence that these genes occurred in the initial stages of the tumor, while *ZNF5211*, *TRRAP*, *EYS*, *RBFOX1*, *GATA1* and *MYH9* are inferred to have mutated on the branches linking to a subclone of the metastatic cells, suggesting a particular temporal ordering of these genes. The

mutations in the *ZNF5211*, *TRRAP*, *EYS*, *RBFOX1*, *GATA1* and *MYH9* genes might indicate the emergence of tumor metastasis. Our results are consistent with past studies. In particular, *APC* has been verified to be an extremely important gene in initializing colorectal cancer. Several studies have shown that the mutation order of these genes appears to be fixed, starting with *APC*, followed by *RAS* and then *TP53* [19–21], providing further support for our findings.

## Conclusions

Development of computational tools based on a phylogenetic framework for use in studying cancer evolution has the potential to provide tremendous insight into the mechanisms that lead to ITH, especially the role of the temporal order of mutations in subsequent cancer progression. For example, Ortmann et al. (2015) have shown differences in the rate of cancer progression and the response to treatment for patients with different mutation order, indicating that the inference of the order in which mutations arise within an individual's cancer may have direct implications in clinical oncology, both for diagnostic applications in measuring the extent of ITH and for improving targeted therapy. SCS data provides an unprecedented opportunity to estimate mutation order at the highest resolution. However, such data is subject to extensive technical errors that arise during whole-genome amplification. In this paper, we introduce MO, a new Bayesian approach for reconstructing the ordering of mutational events from imperfect mutation profiles of single cells. MO can infer the mutation order that best fits single-cell data sets that are subject to the extensive technical noise that is common for SCS data, including ADO, false positive errors, low-quality data and missing data. MO can be easily extended to account for the accelerated mutation rates common in late-stage cancers, and can include position-specific error rates. MO also provides flexibility in choosing the evolutionary model, for which we develop a model that accounts for the effects of point mutations in single-cell data sets. MO is robust to variation in error rates and performs accurately even with a small number of single cells. Our simulation results indicate that as the number of informative single cells increases, MO provides more accurate estimates.

MO performs accurately, as evident from a comprehensive set of simulation studies that take into account different aspects of modern SCS data sets by examining

a wide range of error rates, different fractions of missing data, and different branch lengths. The simulation studies also demonstrate that MO substantially outperforms the state-of-the-art methods and is robust to the technical errors that arise during whole-genome amplification. We also use MO to reconstruct the mutation order for data from two prostate cancer patients and two colorectal cancer patients, and MO is able to accurately reconstruct the mutation order in both cases. The results of MO not only provide insight into the order of mutations, but also the level of certainty in the inferred orders.

MO can be improved by incorporating copy number alterations, as well as by considering mutations that affect the same allele more than once. As SCS data become more advanced, enabling hundreds of cells to be analyzed in parallel at reduced cost and increased throughput, MO is poised to analyze the resulting large-scale data sets to make meaningful inferences of the mutation order during tumor progression for individual patients. MO thus represents a major step forward in understanding the role of mutation order in cancer evolution and as such will have important translational applications for improving cancer diagnosis, treatment, and personalized therapy. Although the current study focuses on cancer, MO can potentially also be applied to single-cell mutation profiles from a wide variety of fields. These applications are expected to provide new insights into our understanding of cancer and other human diseases.

## Methods

Recall that the mutation data can be either binary or ternary. For binary data, we define 0 and 1 as the absence of mutation and presence of mutation, while for ternary data, 0, 1 and 2 represent the homozygous reference (normal), heterozygous (mutation present) and homozygous non-reference (mutation present) genotypes, respectively.

### Model for the somatic mutation process

***Binary data*** For binary genotype data, the mutation process can be modeled by the $2 \times 2$ instantaneous rate matrix

$$\mathcal{M}_\lambda = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} -\lambda & \lambda \\ 0 & 0 \end{pmatrix} \end{array}, \tag{4}$$

where $\lambda$ denotes the instantaneous mutation rate per genomic site. The transition probability matrix $\mathbf{P}(t)$ along a branch of length $t$ is then computed by matrix exponentiation of the product of $\mathcal{M}_\lambda$ and the branch length $t$, which gives

$$\mathbf{P}(t) = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} P_{00}(t) & P_{01}(t) \\ P_{10}(t) & P_{11}(t) \end{pmatrix} \end{array} = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} e^{-\lambda t} & 1 - e^{-\lambda t} \\ 0 & 1 \end{pmatrix} \end{array}. \tag{5}$$

Note that $P_{01}(t)$ is the probability that mutation $i$ is acquired along a branch of length $t$. Under this model and recalling that each mutation evolves independently along different branches in $\mathcal{T}$, the marginal probability that mutation $i$ is acquired on branch $x \in E$, denoted by $\mathbf{P}(B_i = x)$, is thus given by

$$\mathbf{P}(B_i = x | \mathcal{T}, \mathcal{M}_\lambda) = \frac{\left[\prod_{B \in [E \setminus (x \cup E^x(w))]} P_{00}(t_B)\right] \cdot P_{01}(t_x) \cdot \left[\prod_{B \in E^x(w)} P_{11}(t_B)\right]}{\sum_{z \in E} \left( \left[\prod_{B \in [E \setminus (z \cup E^z(h))]} P_{00}(t_B)\right] \cdot P_{01}(t_z) \cdot \left[\prod_{B \in E^z(h)} P_{11}(t_B)\right] \right)}, \tag{6}$$

where $t_B$ is length of branch $B$. In the numerator, the first term is a product of probabilities over all branches without the mutation, the second term is the probability that the mutation is acquired on branch $x$, and the third term is a product of probabilities over all branches with the mutation, i.e., all branches in $E^x(w)$. The denominator is needed to create a valid probability distribution over all possible branches, and is obtained by summing the numerator over all valid branches $z \in E$. The $\mathbf{P}(B_i = x|\mathcal{T}, \mathcal{M}_\lambda)$ term is normalized by the denominator because we exclude two possible conditions: a mutation is not acquired on any branch in $\mathcal{T}$, or a mutation is acquired more than once on different branches in $\mathcal{T}$.

As an example, Fig 3 depicts the observed and true binary genotype for mutation $i = 1$ shown in Fig 2. The set of branches is $E = \{e_1, \ldots, e_8\}$ and the corresponding set of branch lengths would be $\mathbf{t} = \{t_1, \ldots, t_8\}$. If mutation $i$ is acquired on branch $e_1$, the cell descending along branch $e_8$ will not carry the mutation, while those descending from the blue branches would carry this mutation. The marginal probability that mutation $i$ is acquired on branch $e_1$ would be proportional to its numerator, i.e., $\mathbf{P}(B_i = e_1|\mathcal{T}, \mathcal{M}_\lambda) \propto P_{00}(t_8)P_{01}(t_1)[P_{11}(t_2)P_{11}(t_3)P_{11}(t_4)P_{11}(t_5)P_{11}(t_6)P_{11}(t_7)]$.

***Ternary genotype data*** The mutation model for ternary data is more complex and includes three possible ways that mutation $i$ originates on a branch $x$ in $\mathcal{T}$:

I The status of mutation $i$ transitions from $0 \to 1$ on a branch $x$ and there is no further mutation in $\mathcal{T}$.

II The status of mutation $i$ transitions directly from $0 \to 2$ on a branch $x$ in $\mathcal{T}$.

III The status of mutation $i$ transitions from $0 \to 1$ on a branch $x$ and then transitions from $1 \to 2$ on a branch $y \in E^x(w)$ in $\mathcal{T}$.

Letting $B_i$ be the location at which mutation $i$ originates, $B_i^{0\to1}$ would be the branch on which mutation status transitions from 0 to 1, $B_i^{0\to2}$ is the branch on which mutation status transitions from 0 to 2, and $B_i^{1\to2}$ is the branch on which mutation status transitions from 1 to 2. If the mutation $i$ occurs on branch $x$, all cells belonging to $C^x(w)$ will carry 1 or 2 mutations. In other words, $G_{ij} = 1$ or 2 for all $C_j \in C^x(w)$ and $G_{ij} = 0$ for all $C_j \in C \backslash C^x(w)$.

We again consider a continuous-time Markov process to model the above mutation process [12]. We define the instantaneous rate matrix $\mathcal{M}_\lambda$ as

$$\mathcal{M}_\lambda = \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} -(\lambda_1 + \lambda_1\lambda_2) & \lambda_1 & \lambda_1\lambda_2 \\ 0 & -\lambda_2 & \lambda_2 \\ 0 & 0 & 0 \end{pmatrix} \end{array}, \tag{7}$$

where $\lambda_1$ and $\lambda_2$ denote the instantaneous mutation rate per genomic site of the transitions $0 \to 1$ and $1 \to 2$, respectively. Studies have provided evidence that direct mutation of $0 \to 2$ at rate $\lambda_1\lambda_2$ is possible in principle, although it is extremely rare [12]. The transition probability matrix $\mathbf{P}(t) = \exp(\mathcal{M}_\lambda t)$ is then given by

$$\begin{aligned} \mathbf{P}(t) &= \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} P_{00}(t) & P_{01}(t) & P_{02}(t) \\ P_{10}(t) & P_{11}(t) & P_{12}(t) \\ P_{20}(t) & P_{21}(t) & P_{22}(t) \end{pmatrix} \end{array} \\ &= \begin{array}{c} \\ 0 \\ 1 \\ 2 \end{array} \begin{array}{c} \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{pmatrix} e^{-(\lambda_1+\lambda_1\lambda_2)t} & \frac{\lambda_1(e^{-(\lambda_1+\lambda_1\lambda_2)t} - e^{-\lambda_2 t})}{\lambda_2 - (\lambda_1+\lambda_1\lambda_2)} & \frac{(\lambda_1\lambda_2 - \lambda_2)e^{-(\lambda_1+\lambda_1\lambda_2 t)} + \lambda_1 e^{-\lambda_2 t}}{\lambda_2 - (\lambda_1+\lambda_1\lambda_2)} + 1 \\ 0 & e^{-\lambda_2 t} & 1 - e^{-\lambda_2 t} \\ 0 & 0 & 1 \end{pmatrix} \end{array}. \end{aligned} \tag{8}$$

Under this model and recalling that each mutation evolves independently along different branches in $\mathcal{T}$, the marginal probability that mutation $i$ originates on branch $x \in E$ for these three possible conditions is thus given by

$$\mathbf{P}(B_i^{0\to1} = x|\mathcal{T}, \mathcal{M}_\lambda) = \frac{\mathbf{Q}(B_i^{0\to1} = x)}{\sum_{z_1\in E}[\mathbf{Q}(B_i^{0\to1} = z_1) + \mathbf{Q}(B_i^{0\to2} = z_1) + \sum_{z_2}\mathbf{Q}(B_i^{0\to1} = z_1, B_i^{0\to2} = z_2)]}, \tag{9}$$

$$\mathbf{P}(B_i^{0\to2} = x|\mathcal{T}, \mathcal{M}_\lambda) = \frac{\mathbf{Q}(B_i^{0\to2} = x)}{\sum_{z_1\in E}[\mathbf{Q}(B_i^{0\to1} = z_1) + \mathbf{Q}(B_i^{0\to2} = z_1) + \sum_{z_2}\mathbf{Q}(B_i^{0\to1} = z_1, B_i^{0\to2} = z_2)]}, \tag{10}$$

$$\mathbf{P}(B_i^{0\to1} = x, B_i^{1\to2} = y|\mathcal{T}, \mathcal{M}_\lambda) = \frac{\mathbf{Q}(B_i^{0\to1} = x, B_i^{1\to2} = y)}{\sum_{z_1\in E}[\mathbf{Q}(B_i^{0\to1} = z_1) + \mathbf{Q}(B_i^{0\to2} = z_2) + \sum_{z_2}\mathbf{Q}(B_i^{0\to1} = z_1, B_i^{0\to2} = z_2)]}, \tag{11}$$

where

$$\mathbf{Q}(B_i^{0\to1} = x) = \left[\prod_{B\in[E\setminus(x\cup E^x(w))]} P_{00}(t_B)\right] \cdot P_{01}(t_x) \cdot \left[\prod_{B\in E^x(w)} P_{11}(t_B)\right], \tag{12}$$

$$\mathbf{Q}(B_i^{0\to2} = x) = \left[\prod_{B\in[E\setminus(x\cup E^x(w))]} P_{00}(t_B)\right] \cdot P_{02}(t_x) \cdot \left[\prod_{B\in E^x(w)} P_{22}(t_B)\right], \tag{13}$$

$$\mathbf{Q}(B_i^{0\to1}=x, B_i^{1\to2}=y) = \left[\prod_{B\in[E\setminus(x\cup E^x(w))]} P_{00}(t_B))\right]\cdot(P_{01}(t_x))\cdot \left[\prod_{B\in[E^x(w)\setminus(y\cup E^y(b))])} P_{11}(t_B))\right]\cdot (P_{12}(t_y))\cdot\left[\prod_{B\in E^y(b)} P_{22}(t_B)\right]. \tag{14}$$

As in the case of binary data, the reason we normalize the marginal probabilities for the three conditions is that we do not consider scenarios in which those mutations are acquired more than once or those mutations are not acquired in $\mathcal{T}$. As an example, Figure S4 in Additional file 1 depicts the same mutation as in Fig 3, but considers ternary data, leading to the following:

I The marginal probability that mutation $i$ transitions from $0\to1$ on branch $e_1$ is $\mathbf{P}(B_i^{0\to1}=e_1|\mathcal{T},\mathcal{M}_\lambda)\propto P_{00}(t_8)P_{01}(t_1)[P_{11}(t_2)P_{11}(t_3)P_{11}(t_4)P_{11}(t_5)P_{11}(t_6)P_{11}(t_7)]$.

II The marginal probability that mutation $i$ transitions from $0\to2$ on branch $e_1$ is $\mathbf{P}(B_i^{0\to2}=e_1|\mathcal{T},\mathcal{M}_\lambda)\propto P_{00}(t_8)P_{02}(t_1)[P_{22}(t_2)P_{22}(t_3)P_{22}(t_4)P_{22}(t_5)P_{22}(t_6)P_{22}(t_7)]$.

III The marginal probability that mutation $i$ transitions from $0\to1$ on $e_1$, and from $1\to2$ on $e_3$ is $\mathbf{P}(B_i^{0\to1}=e_1, B_i^{1\to2}=e_3|\mathcal{T},\mathcal{M}_\lambda)\propto P_{00}(t_8)P_{01}(t_1)P_{11}(t_2)P_{12}(t_3)P_{22}(t_4)P_{22}(t_5)P_{22}(t_6)P_{22}(t_7)$.

$\mathbf{P}(B_i^{0\to1}=e_1, B_i^{1\to2}=e_3|\mathcal{T},\mathcal{M}_\lambda)$ is the marginal probability that two mutations of the same locus along the genome mutate on two branches $e_1$ and $e_3$, respectively. After the first mutation occurs in branch $e_1$, the second mutation can occur on any branch except $e_1$ and $e_8$.

## Quantification of SCS errors

***Quantification of SCS errors for binary data*** For binary data, the conditional probabilities of the observed data given the true genotype at genomic site $i$ of cell $C_j$ are

$$\mathbf{N}^{ij} = \begin{array}{c} \\ G_{ij}=0 \\ G_{ij}=1 \end{array} \begin{array}{c} \begin{array}{cc} S_{ij}=0 & S_{ij}=1 \end{array} \\ \begin{pmatrix} 1-\alpha_{ij} & \alpha_{ij} \\ \beta_{ij} & 1-\beta ij \end{pmatrix} \end{array}, \tag{15}$$

where $\mathbf{N}^{ij}_{01} = P(S_{ij}=1|G_{ij}=0) = \alpha_{ij}$, and $\mathbf{N}^{ij}_{00}$, $\mathbf{N}^{ij}_{10}$, and $\mathbf{N}^{ij}_{11}$ are defined similarly. Under the assumption that sequencing errors are independent, if mutation $i$ is acquired on branch $x$, we can precisely quantify the effect of SCS technical errors for mutation $i$ as

$$\mathbf{P}(\mathbf{S}_i|B_i = x, \mathcal{T}, \mathbf{N}^i) = \prod_{j=1}^{J} \mathbf{P}(S_{ij}|G_{ij}), \tag{16}$$

where $\mathbf{N}^i = \{\mathbf{N}^{i1}, \ldots, \mathbf{N}^{iJ}\}$. Using the example in Fig 3, the error probability of the observed genotype conditioning on the mutation $i = 1$ occurring on branch $e_1$ would be $\mathbf{P}(\mathbf{S}_1|B_1 = e_1, \mathcal{T}, \mathbf{N}^1) = \mathbf{N}^{11}_{11}\mathbf{N}^{12}_{10}\mathbf{N}^{13}_{11}\mathbf{N}^{14}_{10}\mathbf{N}^{15}_{00}$, where $\mathbf{N}^1 = \{\mathbf{N}^{11}, \ldots, \mathbf{N}^{15}\}$ for this binary data example.

***Quantification of SCS errors for ternary data*** We apply the error model from [10, 11, 13] for ternary data, and the conditional probabilities of the observed data given the true genotype are given by

$$\mathbf{N}^{ij} = \begin{array}{c} \\ G_{ij}=0 \\ G_{ij}=1 \\ G_{ij}=2 \end{array} \begin{array}{c} \begin{array}{ccc} S_{ij}=0 & S_{ij}=1 & S_{ij}=2 \end{array} \\ \begin{pmatrix} 1-\alpha_{ij}-\alpha_{ij}\beta_{ij}/2 & \alpha_{ij} & \alpha_{ij}\beta_{ij}/2 \\ \beta_{ij}/2 & 1-\beta_{ij} & \beta_{ij}/2 \\ 0 & 0 & 1 \end{pmatrix} \end{array}, \tag{17}$$

where $\mathbf{N}_{01}^{ij} = P(S_{ij} = 1|G_{ij} = 0) = \alpha_{ij}$, and $\mathbf{N}_{00}^{ij}$, $\mathbf{N}_{02}^{ij}$, $\mathbf{N}_{10}^{ij}$, $\mathbf{N}_{11}^{ij}$, $\mathbf{N}_{12}^{ij}$, $\mathbf{N}_{20}^{ij}$, $\mathbf{N}_{21}^{ij}$, and $\mathbf{N}_{22}^{ij}$ are defined similarly. Under the same assumptions as for binary genotype data, if mutation $i$ is acquired on branch $x$, we can precisely quantify the effect of SCS technical errors as

$$\mathbf{P}(\mathbf{S}_i|B_i = x, \mathcal{T}, \mathbf{N}^i) = \prod_{j=1}^{J} \mathbf{P}(S_{ij}|G_{ij}). \tag{18}$$

Using the example in Figure S4 in Additional file 1, the error probabilities for the three possible ways that mutation $i$ may arise on branch $e_1$ are

I The error probability under the condition that the true mutation transitions from $0 \to 1$ on branch $e_1$ is $\mathbf{P}(\mathbf{S}_1|B_i^{0\to1} = e_1, \mathcal{T}, \mathbf{N}^1) = \mathbf{N}_{12}^{11}\mathbf{N}_{10}^{12}\mathbf{N}_{11}^{13}\mathbf{N}_{10}^{14}\mathbf{N}_{00}^{15}$.

II The error probability under the condition that the true mutation transitions from $0 \to 2$ on branch $e_1$ is $\mathbf{P}(\mathbf{S}_1|B_i^{0\to2} = e_1, \mathcal{T}, \mathbf{N}^1) = \mathbf{N}_{22}^{11}\mathbf{N}_{20}^{12}\mathbf{N}_{21}^{13}\mathbf{N}_{20}^{14}\mathbf{N}_{00}^{15}$.

III The error probability under the condition that the true mutation transitions from $0 \to 1$ on branch $e_1$, and then transitions from $1 \to 2$ on branch $e_3$ is $\mathbf{P}(\mathbf{S}_1|B_i^{0\to1} = e_1, B_i^{1\to2} = e_3, \mathcal{T}, \mathbf{N}^1) = \mathbf{N}_{12}^{11}\mathbf{N}_{20}^{12}\mathbf{N}_{21}^{13}\mathbf{N}_{20}^{14}\mathbf{N}_{00}^{15}$.

And $\mathbf{N}^1 = \{\mathbf{N}^{11}, \ldots, \mathbf{N}^{15}\}$ for this ternary data example. $\mathbf{P}(\mathbf{S}_1|B_i^{0\to1} = e_1, B_i^{1\to2} = e_3, \mathcal{T}, \mathbf{N}^1)$ is the error probability that two mutations occur on two branches $e_1$ and $e_3$, respectively. The second mutation can occur on any branch except $e_1$ and $e_8$.

## Missing and low-quality data

In real data, missing and low-quality states are observed due to non-uniform or low coverage, and must be taken into account. For each mutation $i$, we exclude those cells with missing states, and a subtree $\mathcal{T}_i$ from the fixed phylogenetic tumor tree $\mathcal{T}$ is extracted. The number of tips $J_i$ in the subtree $\mathcal{T}_i$ is less than or equal to $J$. Let the set of branches $E_i = \{\text{branches in the subtree } \mathcal{T}_i\}$. The probability that mutation $i$ occurs on branch $x$ is then given by

$$\mathbf{P}(B_i = x|\mathcal{T}, \mathcal{M}_\lambda) = \mathbf{P}(B_i = x|\mathcal{T}_i, \mathcal{M}_\lambda), \tag{19}$$

where $\mathbf{P}(B_i = x|\mathcal{T}_i, \mathcal{M}_\lambda)$ is computed based on branches in the subtree $\mathcal{T}_i$, and $\mathbf{P}(B_i = x|\mathcal{T}_i, \mathcal{M}_\lambda)$ is 0 for those branches $x \in E \backslash E_i$. We quantify the effect of the SCS technical errors as

$$\begin{aligned}\mathbf{P}(\mathbf{S}_i|B_i = x, \mathcal{T}, \mathbf{N}^i) &= \mathbf{P}(\mathbf{S}_i|B_i = x, \mathcal{T}_i, \mathbf{N}^i) \\ &= \prod_{j=1}^{J_i}(\sum_{S_{ij}} w_{ij}\mathbf{P}(S_{ij}|G_{ij})),\end{aligned} \tag{20}$$

where $w_{ij}$ is the weight for each possible genotype state at a mutation site. For a site with an observed state that is not missing or ambiguous, $w_{ij}$ is 1 for the observed state and 0 for all other states. For a site with unknown mutation state (ambiguous site), we can assign equal weight for each possible state of the ambiguous site, or we can assign weight based on the sequencing results (using, e.g., the observed frequencies of that state) or other biological characteristics.

### Inferring the location of a mutation in $\mathcal{T}$

Given the observed data matrix $\mathbf{S}$, the tumor phylogenetic tree $\mathcal{T}$, the error rate matrix $\mathbf{N}$, and the mutation process $\mathcal{M}_\lambda$, we can assign a posterior probability distribution to the location of mutation $i$ using Bayes' Theorem,

$$\begin{aligned}\mathbf{P}(B_i = x|\mathbf{S}, \mathcal{T}, \mathbf{N}, \mathcal{M}_\lambda) &= \mathbf{P}(B_i = x|\mathbf{S}_i, \mathcal{T}_i, \mathbf{N}^i, \mathcal{M}_\lambda) \\ &= \frac{\mathbf{P}(\mathbf{S}_i|B_i = x, \mathcal{T}_i, \mathbf{N}^i)\mathbf{P}(B_i = x|\mathcal{T}_i, \mathcal{M}_\lambda)}{\mathbf{P}(\mathbf{S}_i|\mathcal{T}_i, \mathbf{N}^i, \mathcal{M}_\lambda)}.\end{aligned} \tag{21}$$

For mutation $i$, $\mathbf{P}(B_i = x|\mathbf{S}_i, \mathcal{T}_i, \mathbf{N}^i, \mathcal{M}_\lambda) \propto \mathbf{P}(\mathbf{S}_i|B_i = x, \mathcal{T}_i, \mathbf{N}^i)\mathbf{P}(B_i = x|\mathcal{T}_i, \mathcal{M}_\lambda)$ and $\mathbf{P}(B_i = x|\mathbf{S}_i, \mathcal{T}_i, \mathbf{N}^i, \mathcal{M}_\lambda)$ is computed for all $x$ in set $E_i$. For example, there are 8 branches in the phylogeny in Fig 3, so the branch on which mutation $i = 1$ occurs, $B_1$, can be any of the 8 branches. For the binary example, the posterior probability that mutation $i = 1$ occurrs on $e_1$ is $\mathbf{P}(B_1 = e_1|\mathbf{S}_1, \mathcal{T}_1, \mathbf{N}^1, \mathcal{M}_\lambda) \propto P_{00}(t_8)P_{01}(t_1)[P_{11}(t_2)P_{11}(t_3)P_{11}(t_4)P_{11}(t_5)P_{11}(t_6)P_{11}(t_7)] \cdot \mathbf{N}^{11}_{11}\mathbf{N}^{12}_{10}\mathbf{N}^{13}_{11}\mathbf{N}^{14}_{10}\mathbf{N}^{15}_{00}$. In this way, the posterior probability that the mutation occurs on each of the 8 branches can be computed, giving the probability distribution for the location of mutation $i = 1$, i.e. $\mathbf{P}(B_1 = x|\mathbf{S}_1, \mathcal{T}_1, \mathbf{N}^1, \mathcal{M}_\lambda)$ for $x \in \{e_1, \ldots, e_8\}$.

To summarize this probability distribution, we construct a $(1-\theta)\times 100\%$ credible set for the location of mutation $i$ as follows and $\theta$ is between 0 and 1. First, the branches are ranked by their posterior probabilities, and then branches are added to the credible set in the order of decreasing posterior probability until the sum of their probabilities reaches $(1-\theta)$. The size of the credible set is informative about the level of certainty associated with the inferred location for the mutation.

To obtain a point estimate of the branch on which mutation $i$ occurs, we pick the branch that maximizes this posterior probability, i.e., the maximum a posteriori (MAP) estimate. The MAP estimator for the location of mutation $i$ is given by

$$\hat{B}_{i_{MAP}} = \underset{B_i \in \{e_1,\ldots,e_{2J-2}\}}{\operatorname{argmax}} \mathbf{P}(B_i|\mathbf{S}_i, \mathcal{T}, \mathbf{N}^i, \mathcal{M}_\lambda). \tag{22}$$

For the example in Fig 3, the branch with the largest posterior probability is $\hat{B}_{1_{MAP}}$ for mutation $i = 1$.

### Inferring the mutation order in $\mathcal{T}$

In previous section, we assign a probability distribution to the location of each mutation $i$. We now consider the posterior probability distribution of the locations for the $I$ mutations in the sample of $J$ single cells, which is a distribution on a set of cardinality $(2J-2)^I$. Based on the assumption of independence among the $I$ mutations being considered, the posterior distribution for $\mathcal{B}$ is given by

$$\mathbf{P}(\mathcal{B}|\mathbf{S}, \mathcal{T}, \mathbf{N}, \mathcal{M}_\lambda) = \prod_{i=1}^{I} \mathbf{P}(B_i|\mathbf{S}_i, \mathcal{T}, \mathbf{N}^i, \mathcal{M}_\lambda), \tag{23}$$

where $\mathbf{N}^i = \{\mathbf{N}^{i1}, \ldots, \mathbf{N}^{iJ}\}$. From this distribution, we can extract information on the ordering of mutations of interest. For example, if we are interested in the order of mutation $i = 1$ and mutation $i = 2$ in Fig 3, the joint posterior probability distribution that mutation 1 occurs on branch $x \in E$ and mutation 2 occurs on branch $y \in E$ can be used to find the probability that mutation 1 occurs earlier in the tree than mutation $y$. Note that $\mathbf{P}^{B_1=x,B_2=y} = \mathbf{P}(B_1 = x, B_2 = y|\mathbf{S}, \mathcal{T}, \mathbf{N}, \mathcal{M}_\lambda) =$

$\mathbf{P}(B_1 = x|\mathbf{S}, \mathcal{T}, \mathbf{N}, \mathcal{M}_\lambda) \cdot \mathbf{P}(B_2 = y|\mathbf{S}, \mathcal{T}, \mathbf{N}, \mathcal{M}_\lambda)$. This joint distribution can be represented in a matrix given by

$$\begin{array}{c} \\ B_2 = e_1 \\ B_2 = e_2 \\ \vdots \\ B_2 = e_8 \end{array} \begin{array}{c} \begin{array}{cccc} B_1 = e_1 & B_1 = e_2 & \dots & B_1 = e_8 \end{array} \\ \begin{pmatrix} \mathbf{P}^{B_1=e_1,B_2=e_1} & \mathbf{P}^{B_1=e_2,B_2=e_1} & \dots & \mathbf{P}^{B_1=e_8,B_2=e_1} \\ \mathbf{P}^{B_1=e_1,B_2=e_2} & \mathbf{P}^{B_1=e_2,B_2=e_2} & \dots & \mathbf{P}^{B_1=e_8,B_2=e_2} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}^{B_1=e_1,B_2=e_8} & \mathbf{P}^{B_1=e_2,B_2=e_8} & \dots & \mathbf{P}^{B_1=e_8,B_2=e_8} \end{pmatrix} \end{array}.$$

Adding entries of the matrix for which branch $e_1$ is earlier in the tree than branch $e_2$ thus gives the probability that mutation 1 occurs before mutation 2.

To summarize the joint distribution of the locations of mutations with respect to their ordering, we rank all possible mutation orders by their posterior probabilities, and construct a $(1-\theta) \times 100\%$ credible set by adding orders with decreasing probability until their sum exceeds $1-\theta$. To obtain a point estimate of the order, we pick the order that maximizes $\mathbf{P}(\mathcal{B}|\mathbf{S}, \mathcal{T}, \boldsymbol{\theta}, \mathcal{M}_\lambda)$. The MAP estimator for the order of $I$ mutations is thus given by

$$\hat{\mathcal{B}}_{MAP} = \underset{\mathcal{B} \in \{e_1, \dots, e_{2J-2}\}^I}{\operatorname{argmax}} \mathbf{P}(\mathcal{B}|\mathbf{S}, \mathcal{T}, \mathbf{N}, \mathcal{M}_\lambda). \tag{24}$$

One can see that the MAP estimate is actually the order that maximizes the individual probabilities in Equation 21.

## Incorporating uncertainty of the paramaters

Both the binary and ternary model involve specification of mutation rates. While some information about mutation rates is available in the literature, we incorporate uncertainty in these rates by assigning prior distributions. Specifically, we assume gamma priors to describe the distribution of mutation rate for each genomic site,

$$\lambda|\mathbf{S}_i \sim Gamma(\lambda_1, \lambda_2), \tag{25}$$

$$\lambda_1 | \mathbf{S}_i \sim Gamma(\lambda_{11}, \lambda_{12}), \tag{26}$$

$$\lambda_2 | \mathbf{S}_i \sim Gamma(\lambda_{21}, \lambda_{22}), \tag{27}$$

We assume beta priors for the false positive probability $\alpha_{ij}$ and the false negative probability $\beta_{ij}$ for each genomic site in each cell,

$$\alpha_{ij} | \mathbf{S}_i \sim Beta(\alpha_{ij}^1, \alpha_{ij}^2), \tag{28}$$

$$\beta_{ij} | \mathbf{S}_i \sim Beta(\beta_{ij}^1, \beta_{ij}^2). \tag{29}$$

The parameters in these prior distributions are specified using information obtained during sequencing in the case of $\alpha_{ij}$ and $\beta_{ij}$, and using information from the literature for the mutation rate parameters ($\lambda$, $\lambda_1$, and $\lambda_2$). To carry out inference, we use Monte Carlo integration to integrate over the distributions of these parameters.

## Assessment of accuracy of point estimates in simulation study

***Mutation location accuracy*** We first evaluate how accurately MO can estimate the branch on which the mutations occurs. In each setting, we quantify the mutation location accuracy by

$$\frac{\text{total number of mutations which have correct inferred branch}}{\text{total number of mutations}}, \tag{30}$$

where the inferred branch is correct if it is the same as the true mutation branch, and both the numerator and denominator are evaluated among the 1000 trees in

each setting. The results are given in Table S1 (scenario 1) and Table S2 (scenario 2) in Additional file 1, respectively.

***Adjacent order accuracy*** The second way to evaluate MO is to compare the adjacent order accuracy. In each simulated tree, if one pair of mutations are acquired on two adjacent branches, we compare if they would have the same inferred mutation order on any two adjacent branches. In each setting, the adjacent order accuracy is measured by

$$\frac{\text{number of all adjacent paired mutations with correct inferred order}}{\text{number of all true adjacent paired mutations}}, \tag{31}$$

where both the numerator and denominator are evaluated among the 1000 trees, and the results are in Table S3 (scenario 1) and Table S4 (scenario 2) in Additional file 1, respectively.

***Order accuracy*** The third way to evaluate MO is to compare the order accuracy of all pairs of mutations among the 1000 trees without considering if the branches are adjacent, and is measured by

$$\frac{\text{number of all paired mutations with correct inferred order}}{\text{number of all true paired mutations}}, \tag{32}$$

where both the numerator and denominator are evaluated among the 1000 trees, and the results are in Table S5 (scenario 1) and Table S6 (scenario 2) in Additional file 1, respectively.

## Analysis of empirical data

### *Prostate cancer data*

To infer the tumor phylogenies for P1 and P2, we use the SVDQuartets method [22] as implemented in PAUP* [23] using the aligned DNA sequences of all somatic mutations as input with the expected rank of the flattening matrix set to 4. We evaluate all possible quartets and perform a bootstrap analysis to assess confidence in the inferred phylogeny. To root the tree, we specify the normal cell sample as the outgroup. The invariable sites in the aligned data are then filtered out and the

trimmed alignment is used to estimate maximum likelihood branch lengths under the General Time Reversible (GTR) model with heterogeneity of substitution rates among sites modeled with a gamma distribution.

We also select several prostate cancer specific genes to examine. For both the binary and ternary genotype data for these genes, we use MO to compute the posterior probability of mutation on each branch of the tumor phylogeny for each of the two patients. Su et al. (2018) estimated the error probabilities to be $(\alpha, \beta) = (0.29, 0.02)$ for P1, and $(\alpha, \beta) = (0.31, 0.02)$ for P2. Althouth we assign varying error probabilities across genomic sites and cells in Methods section, here we use same priors for all sites, with the estimated error probabilities used to set the priors for $\alpha$ and $\beta$.

To examine the effect of informativeness of the prior on the resulting inference, we consider two priors for each parameter with mean equal to the estimated error probability from the empirical data and with either a large or a small variance. Specifically, for P1 we consider $\alpha|\mathbf{S}_i \sim Beta(0.29, 0.71)$ (larger variance) and $\alpha|\mathbf{S}_i \sim Beta(2.9, 7.1)$ (smaller variance). For P2, we considered $\alpha|\mathbf{S}_i \sim Beta(0.31, 0.69)$ (larger variance) and $\alpha|\mathbf{S}_i \sim Beta(3.1, 6.9)$ (smaller variance). For $\beta$ for both P1 and P2, we consider $\beta|\mathbf{S}_i \sim Beta(0.02, 0.98)$ (larger variance) and $\beta|\mathbf{S}_i \sim Beta(0.2, 9.8)$ (smaller variance).

For ternary genotype data, it is known that the mutation rate $\lambda_2$ (the rate of mutation from the heterozygous state to the homozygous mutant state) is higher than $\lambda_1$ (the rate of mutation from the homozygous normal state to the heterozygous state). According to Iwasa et al. (2004), the mutation rates for first and second type of mutation were estimated to be $\lambda_1 = 10^{-7}$ and $\lambda_2 = 10^{-2}$, respectively. We use these values to specify the priors for the mutation rates. For binary genotype data, the mutation includes both heterozygous mutation and homozygous mutation, and is thus the sum of $\lambda_1$ and $\lambda_2$. Similar to the sequencing error parameter priors, we also set two priors for each mutation rate with equal means but different variances. The distribution of the mutation rate $\lambda_1$ ($0 \to 1$ for ternary genotype) is set as $\lambda_1|\mathbf{S}_i \sim Gamma(10^{-5}, 10^{-2})$ (larger variance) and $\lambda_1|\mathbf{S}_i \sim Gamma(1, 10^{-7})$ (smaller variance). The distribution of the mutation rate $\lambda_2$ ($1 \to 2$ for ternary genotype) is set as $\lambda_2|\mathbf{S}_i \sim Gamma(10^{-1}, 10^{-1})$ (larger variance) and $\lambda_2|\mathbf{S}_i \sim Gamma(1, 10^{-2})$ (smaller variance). The estimated proba-

bilities of mutation do not vary substantially when the priors with larger or smaller variance are used for any of these parameters.

### *Colorectal cancer data*

The SVDQuartets method [22] is also applied to these data to estimate each patient's tumor phylogeny, but the expected rank of flattening matrix is set as 2 since these data are binary. The normal cells in each patient are merged into one normal sample and used as the outgroups. There are 178 and 182 cells collected from CRC1 and CRC2, respectively. We merge those cells with high similarity (similar mutations) and reduce the numbers of cells in the two tumor phylogenies to 64 and 85, respectively. We use the maximum parsimony method to compute the number of changes on each branch and scale the number of changes on each branch by the total number of changes to get the branch lengths.

The original study of Leung et al. (2016) reported 16 and 36 SNVs after variant calling. We view each gene as a unit and thus consider mutation anywhere in a gene as a single mutation, resulting in 16 and 30 mutations of interest for patients CRC1 and CRC2, respectively. Leung et al. (2016) report error rates of $(\alpha, \beta) = (0.015, 0.078)$ and $(\alpha, \beta) = (0.0174, 0.1256)$ for CRC1 and CRC2, respectively. For each patient, we use these values to specify the same priors across all sites. For CRC1, we consider $\alpha|\mathbf{S}_i \sim Beta(0.015, 0.985)$ (larger variance) and $\alpha|\mathbf{S}_i \sim Beta(0.15, 9.85)$ (smaller variance); and $\beta|\mathbf{s}_i \sim Beta(0.078, 0.922)$ (larger variance) and $\beta|\mathbf{S}_i \sim Beta(0.78, 9.22)$ (smaller variance). For CRC2, we consider $\alpha|\mathbf{S}_i \sim Beta(0.0174, 0.826)$ (larger variance) and $\alpha|\mathbf{S}_i \sim Beta(0.174, 8.26)$ (smaller variance); and $\beta|\mathbf{S}_i \sim Beta(0.1256, 0.8744)$ (larger variance) and $\beta|\mathbf{S}_i \sim Beta(1.256, 8.744)$ (smaller variance). The priors for the mutation rates for CRC1 and CRC2 are same as for P1 and P1. As found for the prostate cancer patients, the estimated probabilities do not vary substantially when we use priors with small or large variance.

**Author details**

[1]Division of Biostatistics, The Ohio State University, 1958 Neil Ave, 43210 Columbus, US. [2]Institute for Genomic Medicine, Nationwide Children's Hospital, 700 Childrens Drive, 43205 Columbus, US. [3]Department of Mathematics and Statistics, American University, 3501 Nebraska Ave NW, 20016 Washington, US. [4]Mathematical Biosciences Institute, The Ohio State University, 1735 Neil Ave, 43210 Columbus, US. [5]Department of Statistics, The Ohio State University, 1958 Neil Ave, 43210 Columbus, US. [6]Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, 318 W 12th Ave, 43210 Columbus, US.

**References**


1. Carter SL, Eklund AC, Kohane IS, Harris LN, Szallasi Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nature genetics. 2006;38(9):1043.
2. Holland AJ, Cleveland DW. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. Nature reviews Molecular cell biology. 2009;10(7):478.
3. Campbell PJ, Yachida S, Mudie LJ, Stephens PJ, Pleasance ED, Stebbings LA, et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. Nature. 2010;467(7319):1109.
4. de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science. 2014;346(6206):251–256.
5. Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TB, Veeriah S, et al. Tracking the evolution of non–small-cell lung cancer. New England Journal of Medicine. 2017;376(22):2109–2121.
6. Michor F, Polyak K. The origins and implications of intratumor heterogeneity. Cancer prevention research. 2010;3(11):1361–1364.
7. Alizadeh AA, Aranda V, Bardelli A, Blanpain C, Bock C, Borowski C, et al. Toward understanding and exploiting tumor heterogeneity. Nature medicine. 2015;21(8):846.
8. Ortmann CA, Kent DG, Nangalia J, Silber Y, Wedge DC, Grinfeld J, et al. Effect of mutation order on myeloproliferative neoplasms. New England Journal of Medicine. 2015;372(7):601–612.
9. Navin NE. Cancer genomics: one cell at a time. Genome biology. 2014;15(8):452.
10. Jahn K, Kuipers J, Beerenwinkel N. Tree inference for single-cell data. Genome biology. 2016;17(1):86.
11. Zafar H, Tzen A, Navin N, Chen K, Nakhleh L. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. Genome biology. 2017;18(1):178.
12. Iwasa Y, Michor F, Nowak MA. Stochastic tunnels in evolutionary dynamics. Genetics. 2004;166(3):1571–1579.
13. Kim KI, Simon R. Using single cell sequencing data to model the evolutionary history of a tumor. BMC bioinformatics. 2014;15(1):27.
14. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20(2):289–290.
15. Su F, Zhang W, Zhang D, Zhang Y, Pang C, Huang Y, et al. Spatial Intratumor Genomic Heterogeneity within Localized Prostate Cancer Revealed by Single-nucleus Sequencing. European Urology. 2018;74(5):551–559.
16. Leung ML, Wang Y, Kim C, Gao R, Jiang J, Sei E, et al. Highly multiplexed targeted DNA sequencing from single nuclei. Nature protocols. 2016;11(2):214.
17. Barbieri CE, Bangma CH, Bjartell A, Catto JW, Culig Z, Grönberg H, et al. The mutational landscape of prostate cancer. European urology. 2013;64(4):567–576.
18. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the catalogue of somatic mutations in cancer. Nucleic acids research. 2018;47(D1):D941–D947.
19. Network CGA, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330.
20. Seshagiri S, Stawiski EW, Durinck S, Modrusan Z, Storm EE, Conboy CB, et al. Recurrent R-spondin fusions in colon cancer. Nature. 2012;488(7413):660.
21. Giannakis M, Mu XJ, Shukla SA, Qian ZR, Cohen O, Nishihara R, et al. Genomic correlates of immune-cell infiltrates in colorectal carcinoma. Cell reports. 2016;15(4):857–865.
22. Chifman J, Kubatko L. Quartet inference from SNP data under the coalescent model. Bioinformatics. 2014;30(23):3317–3324.
23. Swofford D. Phylogenetic Analysis Using Parsimony, PAUP* 4.0, beta version 4.0 b2. Sinauer Assoc, Boston, Mass. 1999;.
24. Wu H, Zhang X, Hu Z, Hou Q, Zhang H, Li Y, et al. Evolution and heterogeneity of non-hereditary colorectal cancer revealed by single-cell exome sequencing. Oncogene. 2017;36(20):2857.
25. Ross EM, Markowetz F. OncoNEM: inferring tumor evolution from single-cell sequencing data. Genome biology. 2016;17(1):69.

# Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Availability of data and materials**
MO has been implemented in R and is available at https://github.com/lkubatko/MO. The published data sets used in this manuscript are available through the following accession numbers: single-cell exome sequencing data from a non-hereditary colorectal cancer (CRC1) with SRP067815 [24]; highly-multiplexed SCS data from a metastatic colorectal cancer (CRC2) with SRP058890 [16]; and whole genome single-cell sequencing data from two prostatectomy patients (P1 and P2) SRP127755 [15].

**Competing interests**
The authors declare that they have no competing interests.

**Authors' contribution**
LK designed and directly supervised this work. YG has taken the primary role in developing the ideas in the manuscript and in preparing the work for publication. JG and JC have contributed to the formation of the ideas underlying the work and to our discussions of the implications of the work for understanding evolutionary processes in cancer. All authors wrote and approved the manuscript. Correspondence to Yuan Gao.

**Acknowledgements**
The simulation experiments and data analyses were carried out using the ASC Unity Cluster at The Ohio State University, USA.

**Additional file**
Additional file 1 is a supplementary material. This file contains supplementary tables and supplementary figures.

.

# Figures

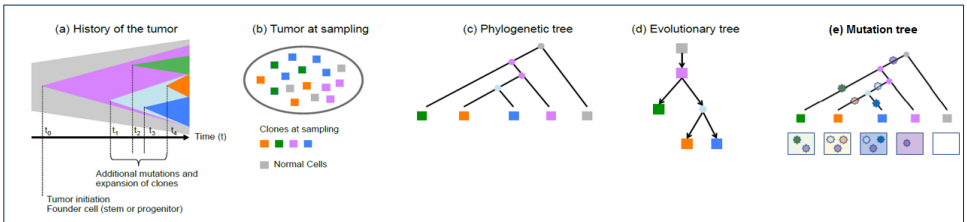


**Figure 1 Pictorial representation of tumor evolution.** (a) - (b) A pictorial representation of the evolution of a tumor from the first initiating mutation to the heterogeneous tissue at the time of sampling, which consists of four different clones and normal tissue. (c) A phylogenetic tree with single cells as the tips; SiFit estimates this type of tree [11]. (d) A clonal lineage tree inferred from sampled cells where each node represents a subclone (cluster of cells). OncoNEM infers a clonal lineage tree [25]. (e) A mutation tree inferred from sampled cells where each star represents the occurrence of one mutation. Boxes underneath each tip show which mutations are present in the cell represented by the tip.

(a) True binary genotype

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ |
|---|---|---|---|---|---|
| $G_1$ | 1 | 1 | 1 | 1 | 0 |
| $G_2$ | 1 | 0 | 0 | 0 | 0 |
| $G_3$ | 0 | 1 | 1 | 0 | 0 |
| $G_4$ | 0 | 1 | 0 | 0 | 0 |
| $G_5$ | 0 | 0 | 1 | 0 | 0 |

(b) Observed binary genotype

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ |
|---|---|---|---|---|---|
| $S_1$ | 1 | 0 | 1 | 0 | 0 |
| $S_2$ | 0 | 0 | 0 | 1 | 0 |
| $S_3$ | 0 | 0 | 1 | 0 | - |
| $S_4$ | 0 | ? | 0 | 1 | 0 |
| $S_5$ | 0 | 1 | ? | 0 | 0 |

**Figure 2 True and observed binary data.** (a) True binary mutation matrix representing the mutation status of the sequenced tumor cells in the mutation tree in Fig 1. Each row represents true genotypes for one genomic site in all cells and each column represents the true genotypes of multiple genomic sites for one single cell. (b) Observed mutation matrix with missing and ambiguous values (red), as well as genotypes that are misrecorded with respect to the true mutation matrix (red numbers; these are either false positives or false negatives). The red dash indicates a missing value since the sequencing process does not return signal at this site of this cell, and the red question mark represents an ambiguous value. Each row represents observed genotypes for one genomic site in all cells and each column represents the observed genotypes of multiple genomic sites for one single cell.

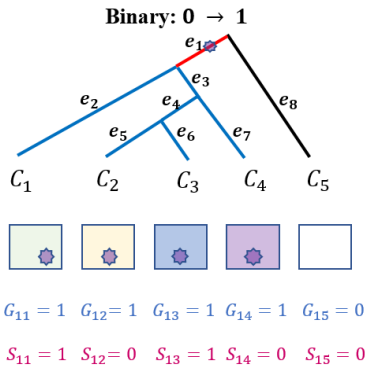


**Figure 3 Binary mutation process example.** Example in which a mutation is acquired on branch $e_1$ (marked with red color). The cell descending from branch $e_8$ (marked with black color) do not carry the mutation, while the cells descending from the blue branches do carry the mutation.

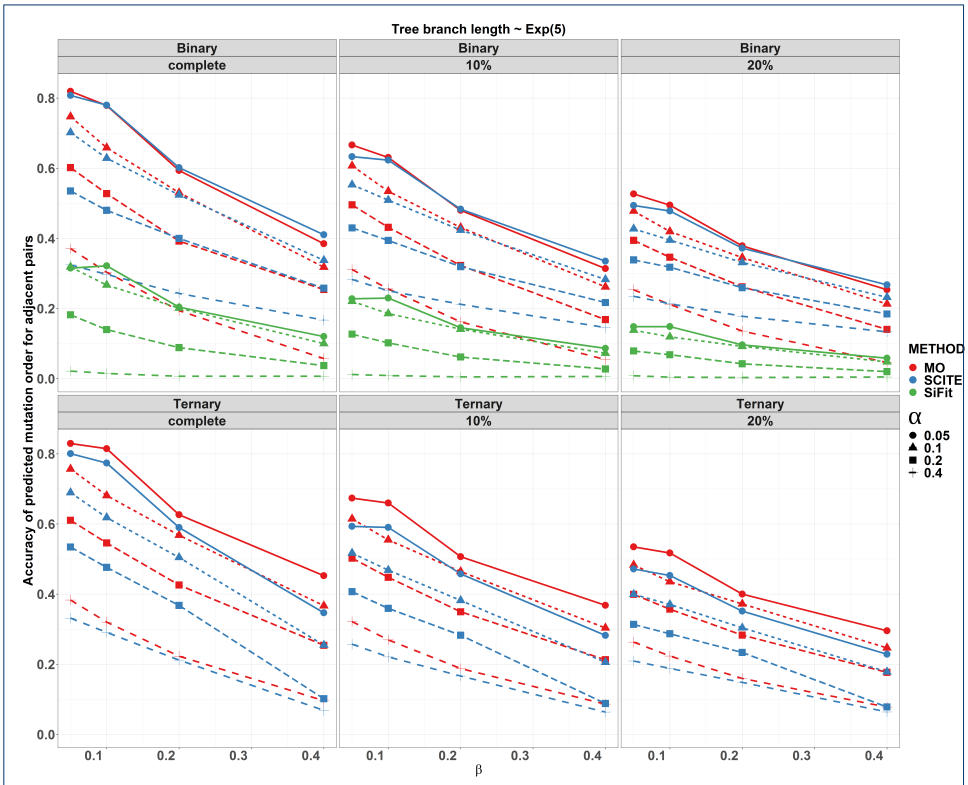


**Figure 4 Adjacent order accuracy in scenario 1.** Plot of adjacent order accuracy of MO, SCITE and SiFit in scenario 1. Branch lengths follow an exponential distribution with mean 0.2. Each panel includes the results from the specific type of genotype and missing data percentage. In each panel, red, blue and green colors correspond to MO, SCITE and SiFit, respectively. Each point shape on the line represent a different FPR $\alpha$. The x-axis is FNR $\beta$, and the y-axis is adjacent order accuracy.

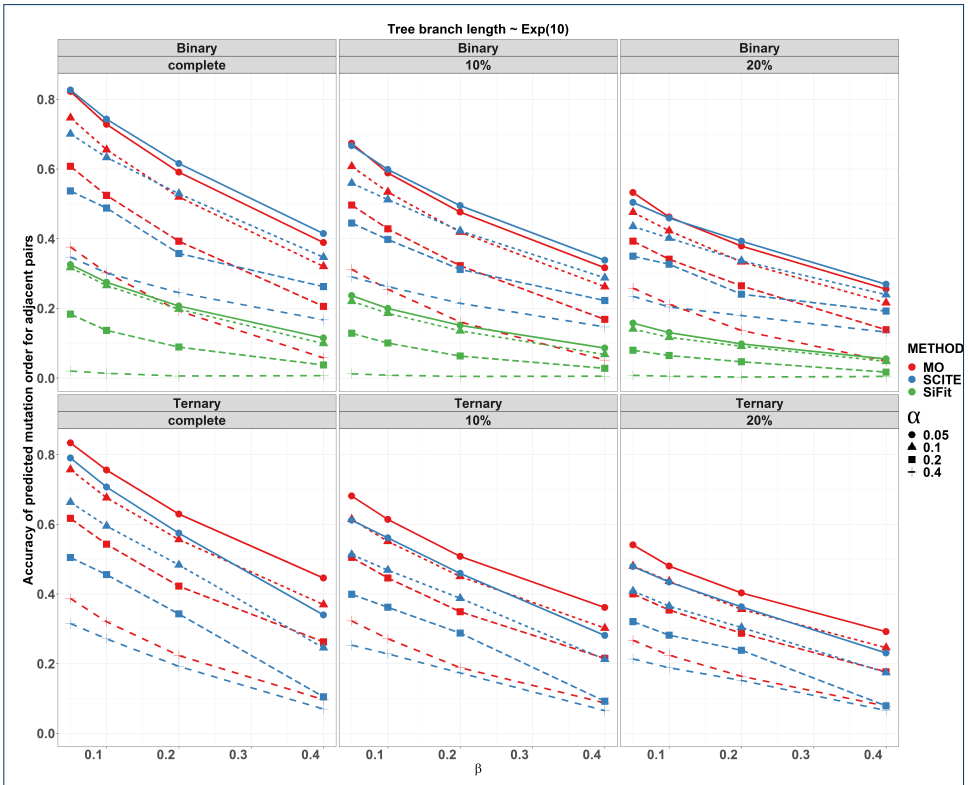


**Figure 5 Adjacent order accuracy in scenario 2.** Plot of adjacent order accuracy of MO, SCITE and SiFit in scenario 2. Branch lengths follow an exponential distribution with mean 0.1. Each panel includes the results from the specific type of genotype and missing data percentage. In each panel, red, blue and green colors correspond to MO, SCITE and SiFit, respectively. Each point shape on the line represent a different FPR $\alpha$. The x-axis is FNR $\beta$, and the y-axis is adjacent order accuracy.

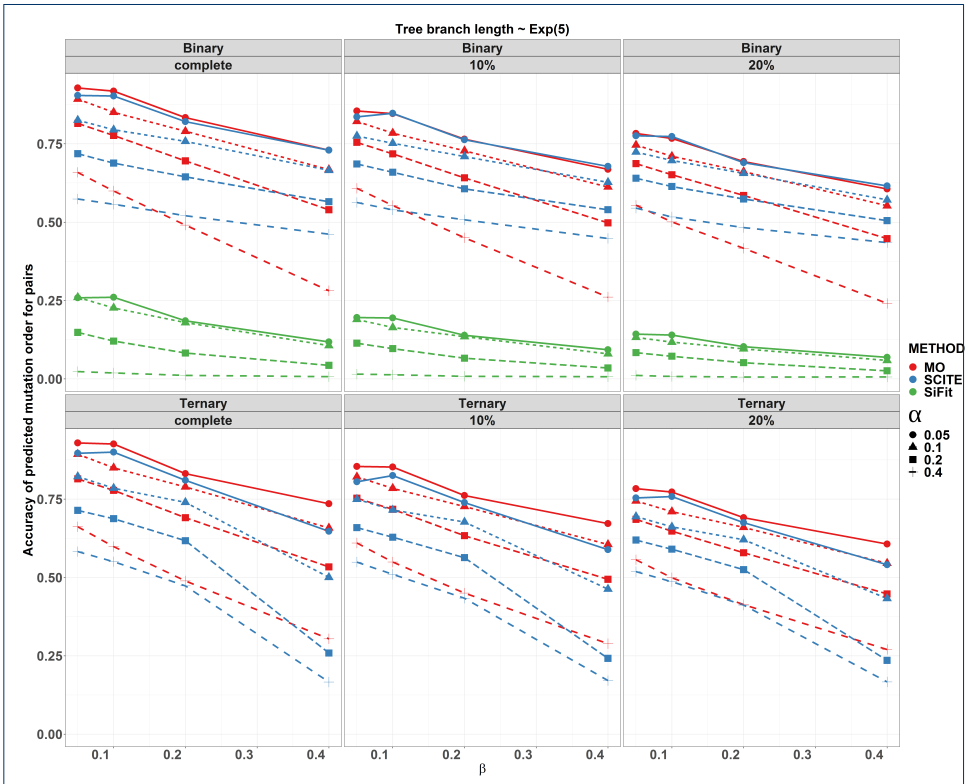


**Figure 6 Order accuracy in scenario 1.** Plot of order accuracy of MO, SCITE and SiFit in scenario 1. Branch lengths follow an exponential distribution with mean 0.2. Each panel includes the results from the specific type of genotype and missing data percentage. In each panel, red, blue and green colors correspond to MO, SCITE and SiFit, respectively. Each point shape on the line represent a different FPR $\alpha$. The x-axis is FNR $\beta$, and the y-axis is order accuracy.

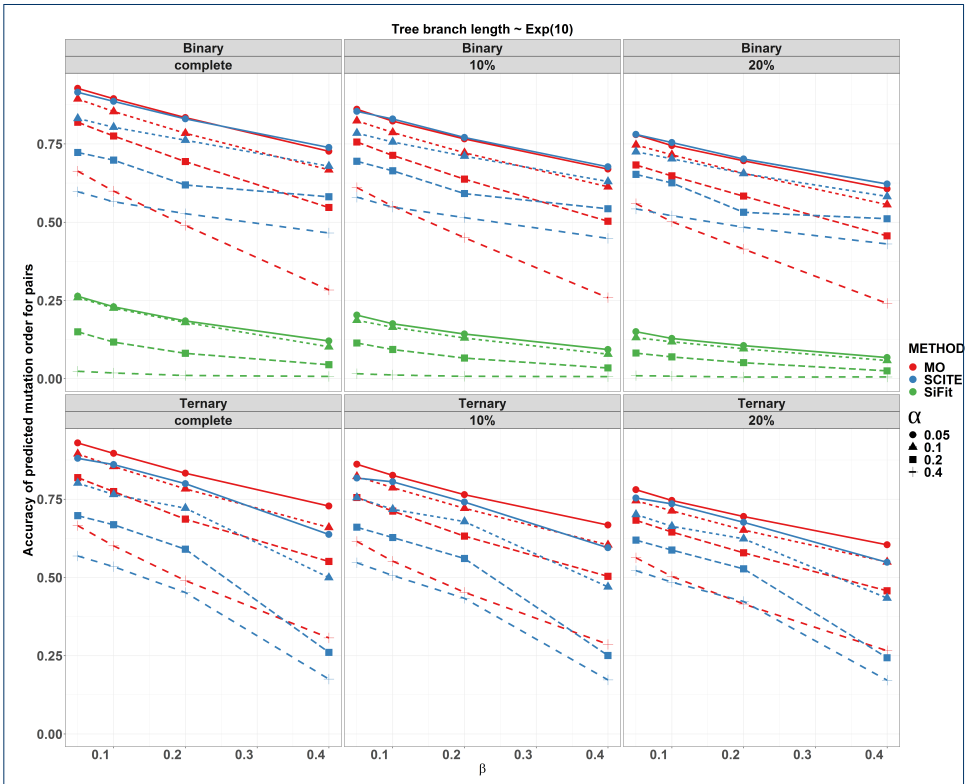


**Figure 7 Order accuracy in scenario 2.** Plot of order accuracy of MO, SCITE and SiFit in scenario 2. Branch lengths follow an exponential distribution with mean 0.1. Each panel includes the results from the specific type of genotype and missing data percentage. In each panel, red, blue and green colors correspond to MO, SCITE and SiFit, respectively. Each point shape on the line represent a different FPR $\alpha$. The x-axis is FNR $\beta$, and the y-axis is order accuracy.

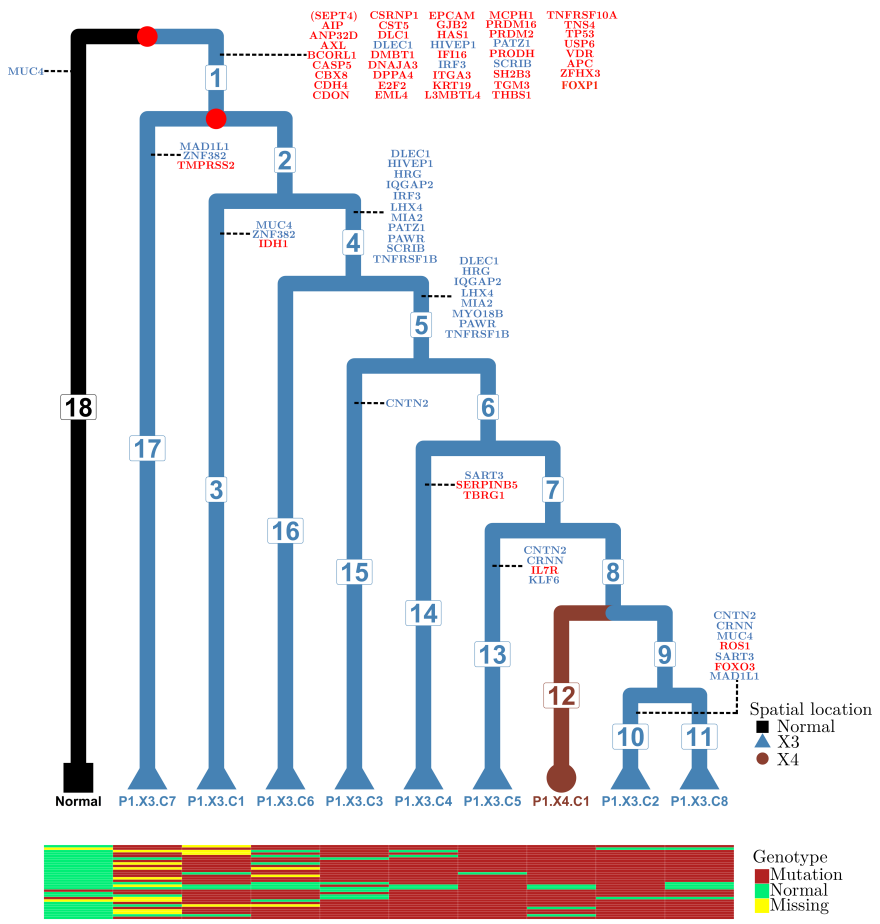


**Figure 8 Patient P1 tumor phylogeny.** P1 tumor phylogenetic tree and inferred temporal order of the mutations. The normal cell and is set as the outgroup. There are 18 branches in this tree. We do not assume the molecular clock when estimating the branch lengths, and branch lengths in this figure are not drawn to scale. The color and tip shape represents the spatial locations of the cells. The temporal order of the mutations is annotated on the branches of the tree. The uncertainty of mutation locations are highlighted with colors. Mutations with very strong signals (probabilities of occurring on one branch greater than 0.7) are marked in red, while genes with moderate signals (probabilities that sum to between 0.1 and 0.7 on more than one branch) are marked in blue.

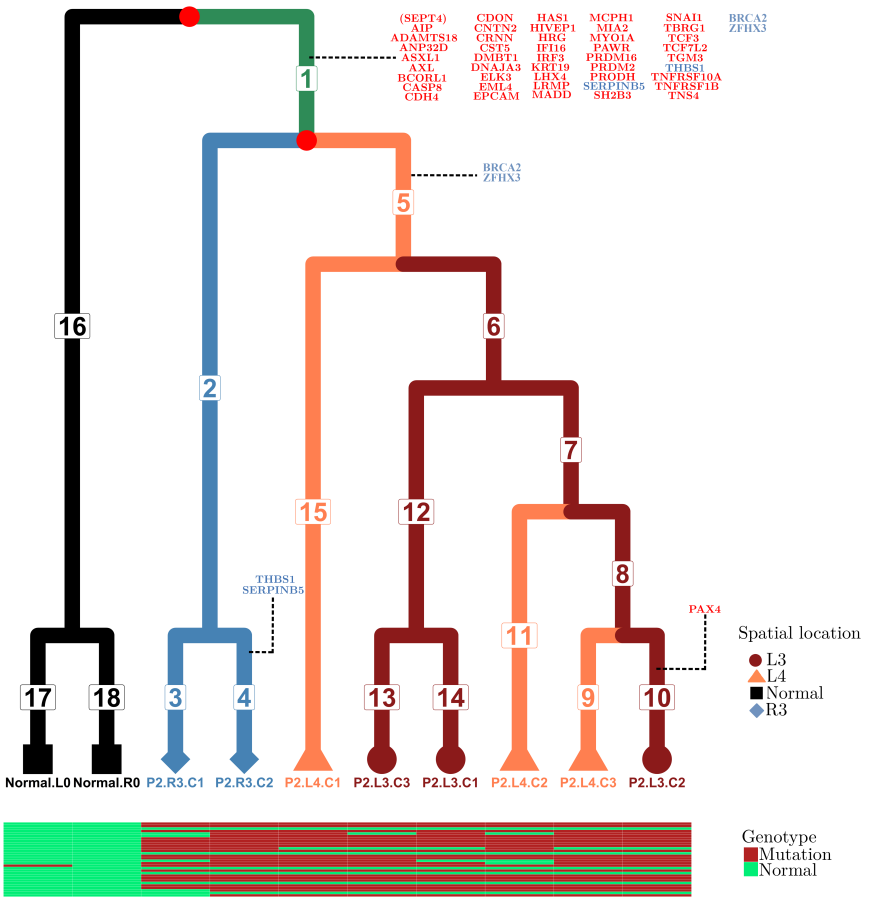


**Figure 9 Patient P2 tumor phylogeny.** P2 tumor phylogenetic tree and inferred temporal order of the mutations. Normal.R0 and Normal.L0 are normal cells from the right side and the left side of tissue, respectively. Normal.R0 and Normal.L0 are set as the outgroup. 18 branches are in this tree. We do not assume the molecular clock when estimating the branch lengths, and branch lengths in this figure are not drawn to scale. The color and tip shape represents the spatial locations of those cells.The temporal order of the mutations is annotated on the branches of the tree. The uncertainty of mutation locations are highlighted with colors. Mutations with very strong signals (probabilities of occurring on one branch greater than 0.7) are marked in red, while genes with moderate signals (probabilities that sum to between 0.1 and 0.7 on more than one branch) are marked in blue.

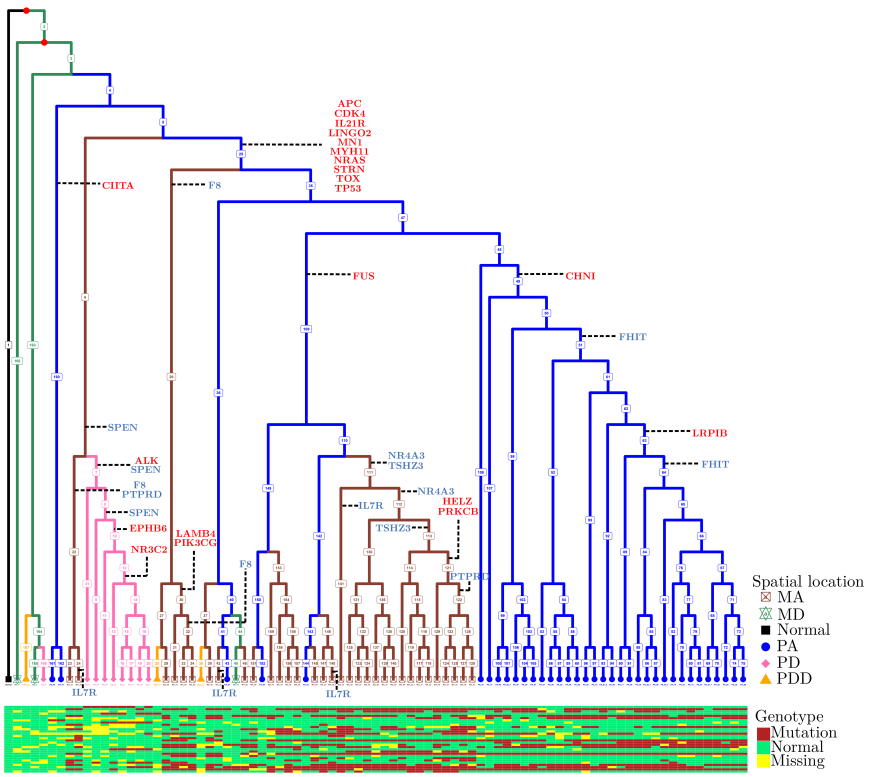


**Figure 10 Patient CRC1 tumor phylogeny.** CRC1 tumor phylogenetic tree and inferred temporal order of the mutations. The color and tip shape represent the spatial locations of the sampled cells. The uncertainty of mutation locations are highlighted with colors. The temporal order of the mutations is annotated on the branches of the tree. The uncertainty of mutation locations are highlighted with colors. Mutations with very strong signals (probabilities of occurring on one branch greater than 0.7) are marked in red, while genes with moderate signals (probabilities that sum to between 0.1 and 0.7 on more than one branch) are marked in blue. The branch lengths are not scaled.

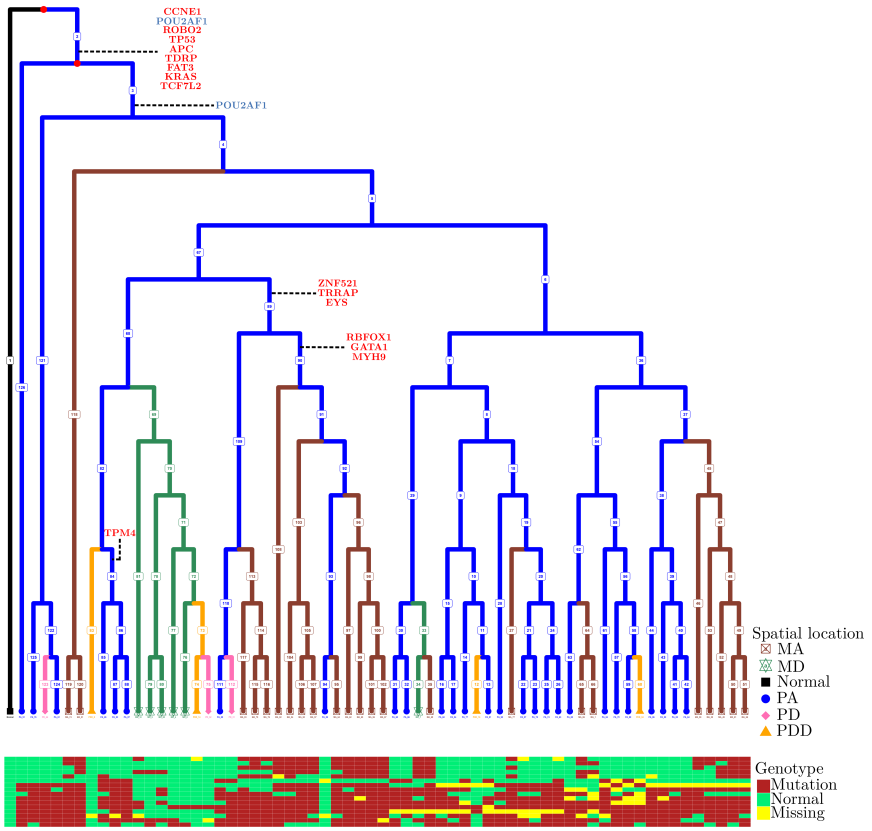


**Figure 11 Patient CRC2 tumor phylogeny.** CRC1 tumor phylogenetic tree and inferred temporal order of the mutations. The color and tip shape represent the spatial locations of the sampled cells. The uncertainty of mutation locations are highlighted with colors. The temporal order of the mutations is annotated on the branches of the tree. The uncertainty of mutation locations are highlighted with colors. Mutations with very strong signals (probabilities of occurring on one branch greater than 0.7) are marked in red, while genes with moderate signals (probabilities that sum to between 0.1 and 0.7 on more than one branch) are marked in blue. The branch lengths are not scaled.

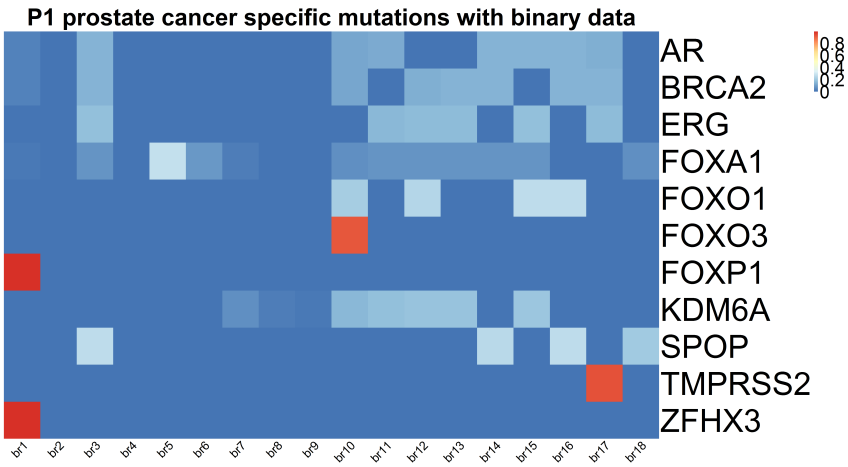


**Figure 12 Heatmap of posterior probabilities of mutation branch for P1 using binary genotype data**. This heatmap is for prostate cancer specific genes. Colors indicate the magnitude of the probability, with red indicating probability close to 1 and blue indicating probability close to 0. For P1, prior of $\alpha$ is set as $\alpha|\mathbf{S}_i \sim Beta(0.29, 0.71)$ (larger variance). The prior of $\beta$ is set as $\beta|\mathbf{S}_i \sim Beta(0.02, 0.98)$ (larger variance). Distribution of mutation rate $\lambda_1$ ($0 \to 1$ for ternary genotype) is set as $\lambda_1|\mathbf{S}_i \sim Gamma(10^{-5}, 10^{-2})$ (larger variance). Distribution of mutation rate $\lambda_2$ ($1 \to 2$ for ternary genotype) is set as $\lambda_2|\mathbf{S}_i \sim Gamma(10^{-1}, 10^{-1})$ (larger variance).

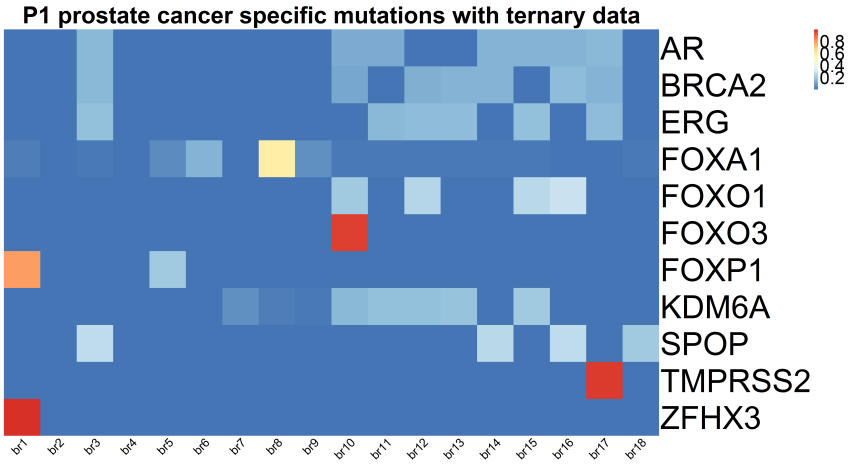


**Figure 13 Heatmap of posterior probabilities of mutation branch for P1 using ternary genotype data.** This heatmap is for prostate cancer specific genes. Colors indicate the magnitude of the probability, with red indicating probability close to 1 and blue indicating probability close to 0. Prior distribution of error rates and mutation rates are set as in Fig 12

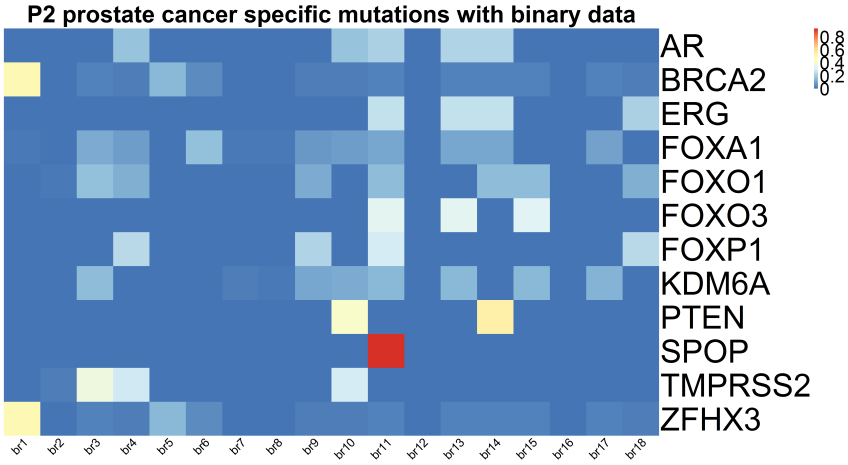


**Figure 14 Heatmap of posterior probabilities of mutation branch for P2 using binary genotype data.** This heatmap is for prostate cancer specific genes. Colors indicate the magnitude of the probability, with red indicating probability close to 1 and blue indicating probability close to 0. For P2, prior distribution for the mutation rate parameters are as in Figure 12. Prior of $\alpha$ is set as $\alpha|\mathbf{S}_i \sim Beta(0.31, 0.69)$ (larger variance). The prior of $\beta$ is set as $\beta|\mathbf{S}_i \sim Beta(0.02, 0.98)$ (larger variance).

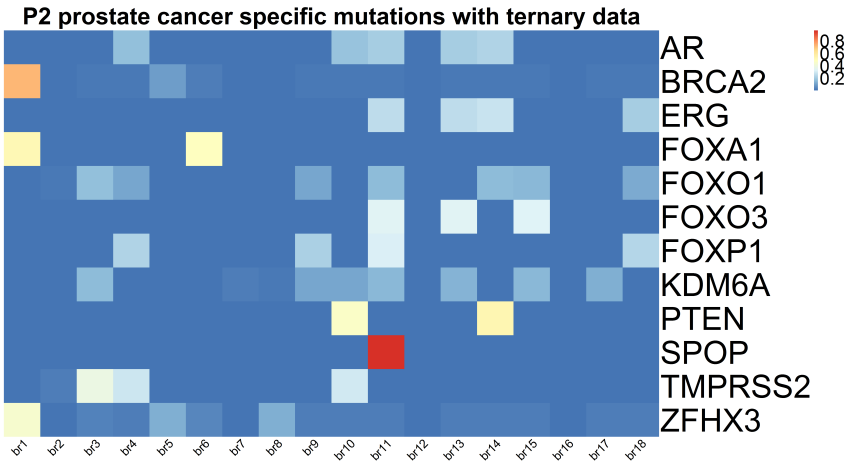


**Figure 15 Heatmap of posterior probabilities of mutation branch for P2 using ternary genotype data.** This heatmap is for prostate cancer specific genes. Colors indicate the magnitude of the probability, with red indicating probability close to 1 and blue indicating probability close to 0. Prior distribution of error rates and mutation rates are set as in Fig 14.