$BiTSC^2$: a Bayesian inference of tumor clonal tree by joint analysis of single-cell SNV and CNA data

Supplementary Information

Ziwei Chen[1,2,3] Fuzhou Gong[2,3] Lin Wan[2,3*]
Liang Ma[1,3 *]

[1] Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China

[2]NCMIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

[3]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

[*]Corresponding Authors. Emails: lwan@amss.ac.cn (Lin Wan) and maliang@ioz.ac.cn (Liang Ma).

Contents

Supplementary Methods 4

Supplementary Note 1: Inference of parameters 4

Sampling of subclone assignment C 4

Sampling of SNV and CNA origin matrices L^o, Z^o and updating phase indicator g 5

Sampling of clone tree $\mathcal{T}$ 5

Sampling of missing rate ρ, ADO rate μ, dispersion parameters w and s 6

Supplementary Note 2: Heuristic initialization process for MCMC parallel chains 6

Supplementary Note 3: Derivation of the fully conditional distribution for all model parameters and maximum likelihood inference for g 7

Supplementary Note 4: Model selection 9

Supplementary Note 5: Simulation process 9

Supplementary Note 6: Evaluations of ARI 10

Supplementary Note 7: Preprocessing of raw data in ERBC datasets 10

Supplementary Figures 11

Figure S1: The ground truth of simulation datasets in G1, containing subclonal phylogenetic tree ($\mathcal{T}$) and genotype matrix of CNA (L) and SNV (Z). 11

Figure S2: The ground truth of simulation datasets in G2, containing subclonal phylogenetic tree ($\mathcal{T}$) and genotype matrixes of CNA (L) and SNV (Z). 12

Figure S3: The clone tree can be reconstructed by SNV markers. 13

Figure S4: The ground truth of simulation datasets in G3-G5, containing subclonal phylogenetic tree ($\mathcal{T}$) and genotype matrix of CNA (L) and SNV (Z). 14

Figure S5: The ground truth of simulation datasets in G6, containing subclonal phylogenetic tree ($\mathcal{T}$) and genotype matrix of CNA (L) and SNV (Z). 15

Figure S6: The ground truth of simulation datasets in G7, containing subclonal phylogenetic tree ($\mathcal{T}$) and genotype matrix of CNA (L) and SNV (Z). 16

Figure S7: The simulated coverage heterogeneity in sequencing technologies under different sequencing depths. 17

Figure S8: Comparison of performance on G1 for subclone assignment, subclonal CNA genotype recovery and tree reconstruction between *BiTSC*^{2} with true segment information as input and ground truth. 18

Figure S9: Comparison of performance on G2 for subclone assignment, subclonal CNA genotype recovery and tree reconstruction between *BiTSC*^{2} with true segment information input and ground truth. 19

Figure S10: The estimation of phase indictor g by *BiTSC*^{2} with true segment information as input for dataset G2. 20

Figure S11: Comparison of performance on G1 and G2 for scSNV genotype recovery, subclone assignment and tree reconstruction among *BiTSC*^{2} with locus specific segments as input, RobustClone and BEAM. 21
Figure S12: Comparison of overall performance on G3-G7 for scSNV genotype recovery, subclone assignment and tree reconstruction among *BiTSC*^{2} with real segment information as input, RobustClone and BEAM. 22
Figure S13: Comparison of detail performance on G3-G7 for scSNV genotype recovery, subclone assignment and tree reconstruction among *BiTSC*^{2} with locus specific segments as input, RobustClone and BEAM. 23
Figure S14: Comparison of overall performance on G3-G7 for scSNV genotype recovery, subclone assignment and tree reconstruction among *BiTSC*^{2} with locus specific segments as input, RobustClone and BEAM. 24
Figure S15: The BIC of metastatic colorectal cancer data calculated in Model selection step. 25
Figure S16: The phylogeny tree inferred by SCITE and SCARLET on the dataset from a Metastatic Colorectal Cancer Patient. 26
Figure S17: The BIC of ERBC dataset calculated in model selection step. 27
Figure S18: The missing rate ρ estimated by *BiTSC*^{2} on the datasets from G5 groups. 28
Figure S19: The ground truth where SNV randomly occurs on chromosomes. 29
Figure S20: The tree inferred by SCARLET and *BiTSC*^{2} for the toy model in Figure 1A. 30
Supplementary Tables **31**
Table S1: The definitions of all parameters and examples of main parameters in Figure 1A. 31
Table S2: The setting of simulation parameters for comparison data. 33
Table S3: Prior distribution parameters setting of *BiTSC*^{2} for simulations G1-G7. 34
Table S4: MCMC sampling parameters setting of *BiTSC*^{2} for simulations G1-G7. 34
Table S5: Prior distribution parameters setting of *BiTSC*^{2} for metastatic colorectal cancer dataset. 35
Table S6: MCMC sampling parameters setting of *BiTSC*^{2} for real datasets. 35
Table S7: Prior distribution parameters setting of *BiTSC*^{2} for breast cancer dataset. 36
Supplementary References **37**

Supplementary Methods

Supplementary Note 1: Inference of parameters

For brevity of description, we denote the set of all unknown parameters as Ω, and all unknown parameters except ω as $\Omega_{-\omega}$.

Sampling of subclone assignment C

Due to the sharing of genetic information among homogeneous cells, we assume that there are K latent subclones in the N single cells drawn for sequencing ($K \ll N$). The latent state of cell n is denoted by $C_n = k$ ($n \in \{1, \cdots, N\}, k \in \{1, \cdots, K\}$). We assume that C_n follows the Categorical Distribution with parameter ϕ, where ϕ is a vector of length K and the sum of the elements is 1. ϕ describes the subclonal prevalence, where each element ϕ_k represents the proportion of cells from subclone k. Then we introduce an additional parameter, θ_k, for each ϕ_k, and denote the vector composed of all θ_k ($k \in \{1, \cdots, K\}$) as Θ. We assign each θ_k an independent Gamma$(\gamma, 1)$ prior distribution, and let $\phi_k = \theta_k / \sum_{i=1}^{K} \theta_i$ ($k \in \{1, \cdots, K\}$). This is equivalent to assigning ϕ_k a symmetric Dirichlet$(\gamma, \gamma, \cdots, \gamma)$ prior distribution with mean and mode $(1/K, 1/K, \cdots, 1/K)$, which gives no preference to any subclones. The purpose of introducing θ_k is that we can update one element of Θ at a time, while sampling ϕ_k directly requires updating the entire vector of ϕ due to the restriction of the sum of elements as 1. The former approach usually leads to better mixing of the MCMC (Zeng *et al.*, 2019).

Since the full conditional samples of θ_k, i.e., $p(\theta_k | D, X, \Psi, \varepsilon, \Omega_{-\theta_k})$ can not be directly sampled, we take Metropolis-Hastings sampling method for sampling Θ.

Based on the current sample θ_k in Markov chain, a new θ_k^* is proposed from the transition function $f(\theta_k^* | \theta_k, \lambda)$, where $f(\theta_k^* | \theta_k, \lambda)$ is the density function of Gamma$(\lambda\theta_k, 1/\lambda)$ with center θ_k and variance θ_k/λ. The tuning parameter λ controls the proposing step size, and a larger λ value usually leads to a higher acceptance rate. In our implementation, we adaptively adjust its value to keep the acceptance rate in a reasonable range to ensure effective mixing of Markov chains (Zeng *et al.*, 2019).

Since the sampling space for subclone assignment is discrete and small, we perform Gibbs sampling on C, after updated each value in Θ, by calculating the probabilities of subclone assignment of each cell n. For each cell n, we calculate $p(C_n = k | D, X, \Psi, \varepsilon, \Omega_{-C_n})$ for all possible k ($k \in \{1, 2, \cdots, K\}$) and use them as weights to sample a new state of C_n.

Sampling of SNV and CNA origin matrices L^o, Z^o and updating phase indicator g

Since the sampling spaces for CNV status L^o, SNV status Z^o are discrete and relatively small, we also apply Gibbs sampling by calculating the probabilities of all possible states as weights to build the conditional probabilities.

Since the loci are independent when sampling SNV status, we update Z^o locus by locus. For each m, we calculate the posterior probability $p(Z^o_m = (k, v)|D, X, \Psi, \varepsilon, \Omega_{-Z^o_m})$ for each state combination (k, v). Under scenarios where CNA happens after SNV at overlapping locus m, we calculate the full conditional distribution by integrating over all possible values of phase indicator g_m. That is with 1/2 probability the subsequent CNA happened on the wild type allele ($g_m = 0$) and with 1/2 probability the CNA occurred on the mutant allele ($g_m = 1$), that is, $p(Z^o_m = (k, v)|D, X, \Psi, \varepsilon, \Omega_{-Z^o_m}) = \frac{1}{2}p(Z^o_m = (k, v)|D, X, \Psi, \varepsilon, \Omega_{-Z^o_m}, g_m = 0) + \frac{1}{2}p(Z^o_m = (k, v)|D, X, \Psi, \varepsilon, \Omega_{-Z^o_m}, g_m = 1)$. Then we sample a new Z^o_m from the (posterior) conditional distribution of all possible states.

The sampling process of L^o is similar to that of Z^o. If segmentation information is available, instead of up one locus at a time, all loci within a segment will be collectively updated. For each locus in the segment, under scenarios where CNA happens after SNV at overlapping locus m, we also calculate the full conditional distribution by integrating over all possible values of phase indicator g_m. Then we use the product of weighted posterior probabilities of all loci in the segment as the current sampling probability of state (k, v) for Gibbs sampling of L^o_m.

For the hyper-parameter π of L^o, we adopt Gibbs sampling to update π since we can write the fully conditional distribution in the form of Beta distribution as follows:

$$p(\pi|L^o) \sim \text{Beta}(u + \alpha, S - u + \beta),$$

where S is the number of genome segments, and u is the number of segments without CNA (Zeng *et al.*, 2019).

After performing Gibbs sampling on L^o and Z^o, we estimate each element of g with the maximum probability at each locus.

Sampling of clone tree $\mathcal{T}$

Since the sampling space of phylogenetic tree is discrete, but the size increases rapidly with the growth of the number of subclones K, it will be a huge computational burden to explore all the discrete values of the tree and calculate the posterior probabilities. Here, we adopt a mixed sampling method for the tree, which randomly applies Metropolis-Hastings sampling and slice sampling.

For Metropolis-Hastings sampling, we adopt the following sampling method: randomly select a leaf node and reconnect it to a randomly selected parent node to obtain a new tree structure. However, Metropolis-Hastings sampling may fall into a local tree structure, as so, we also occasionally apply slice sampling:

(1) randomly generate a parameter $\sigma \sim \text{Uniform}(0,\ p(\mathcal{T}|D, X, \Psi, \varepsilon, \Omega_{-\mathcal{T}}))$,

(2) randomly and repeatedly sample a $\mathcal{T}^*$ from tree space and accept $\mathcal{T}^*$ if $p(\mathcal{T}^*|D, X, \Psi, \varepsilon, \Omega_{-\mathcal{T}}) \geq \sigma$.

Slice sampling enables our sampler to make big jumps which can avoid the chain being trapped into local mode. According to empirical analysis, the combination of Metropolis-Hastings sampling and slice sampling not only increases the sampler's mobility, but also produces a higher acceptance rate and is robust in a variety of simulation situations.

Sampling of missing rate ρ, ADO rate μ, dispersion parameters w and s

Since the full conditional distribution of ρ is difficult to sample directly, we use Metropolis sampling with uniform prior on interval $[0, 1]$ to update ρ. Assuming that ρ_0 is the sample of the current estimated missing rate in MCMC chain, we randomly and uniformly sample a new sample ρ_1 in the interval $[0, 1]$, calculate the ratio of the posterior probability of ρ (i.e., $p(\rho|D, X, \Psi, \varepsilon, \Omega_{-\rho})$) before and after sampling to judge whether to accept the new sample ρ_1 according to the Metropolis sampling criterion. The sampling processes of μ is the same to that of ρ.

For the dispersion parameters s and w of the Negative Binomial distribution and Beta-Binomial distribution, we also use Metropolis sampling with Gamma prior. Specifically, assuming that s_0 is the currently estimated sample in the MCMC chain, we generate a new sample s_1 from the normal distribution with mean s_0 and variance s_d (given in advance). Then we calculate the ratio of the posterior probability of s before and after sampling to judge whether to accept the new sample s_1 (Marass *et al.*, 2016). The sampling processes of w is the same to that of s.

Supplementary Note 2: Heuristic initialization process for MCMC parallel chains

We use heuristic initialization for each parallel chain before MCMC sampling. We calculate the variant reads frequency (VRF) at each locus in each cell based on the total reads and mutant reads. Generally, cells from the same subclone have similar VRF. Therefore, based on the VRF matrix, we use Gaussian mixture model to cluster the cells and obtain the subclone cell assignment to initialize C. Then we obtain the VRF of each subclone at each locus, and use the minimum spanning tree (MST) algorithm to construct the subclonal evolutionary path to initialize the clone tree $\mathcal{T}$. After initializing C and $\mathcal{T}$ for each chain, CNA and SNV are randomly allocated on the tree, and then MCMC sampling optimization is performed.

Supplementary Note 3: Derivation of the fully conditional distribution for all model parameters and maximum likelihood inference for g

(1) π

the posterior distribution of π is (with prior Beta(α, β)):

$$\begin{aligned} p(\pi|L^o) &\propto p(L^o|\pi)p(\pi) \\ &\propto \pi^u(1-\pi)^{S-u}p(\pi) \\ &\propto \text{Beta}(u+\alpha, S-u+\beta), \end{aligned}$$

where S is the number of genome segments, and u is the number of segments without CNA.

(2) Θ

the posterior distribution of Θ is:

$$\begin{aligned} p(\Theta|D, X, \Psi, \varepsilon, \Omega_{-\Theta}) &= p(\Theta|C) \\ &\propto p(C|\Theta)p(\Theta) \\ &\propto p(\Theta)\prod_k(\frac{\theta_k}{\sum_k \theta_k})^{N_k}, \end{aligned}$$

where N_k is the number of cells belonging to subclone k.

(3) C

we update C one by one:

$$\begin{aligned} p(C_n|D, X, \Psi, \varepsilon, \Omega_{-C_n}) &= p(C_n|D, X, \Psi, L^o, Z^o, \mathcal{T}, g, \rho, \mu, s, w, \varepsilon) \\ &\propto p(C_n)\prod_m p(x_{mn}|d_{mn}, Z^o_m, L^o_m, C_n, \mathcal{T}, \mu, w, \varepsilon, g_m)p(d_{mn}|\psi_n, L^o_m, C_n, \mathcal{T}, \rho, \mu, s, \varepsilon) \end{aligned}$$

(4) Z^o

we update Z^o row by row:

$$\begin{aligned} p(Z^o_m|D, X, \Psi, \varepsilon, \Omega_{-Z^o_m}) &= p(Z^o_m|D, X, L^o_m, C, \mathcal{T}, w, \mu, \varepsilon) \\ &\propto p(Z^o_m)\int_{g_m}\prod_n p(x_{mn}|d_{mn}, Z^o_m, L^o_m, C_n, \mathcal{T}, w, \mu, \varepsilon, g_m) \end{aligned}$$

(5) L^o

we update rows of L^o in the same segment together. Consider the loci in segment Δ_i and assume they share the same CNA status $L^o_{\Delta_i}$:

$$\begin{aligned} p(L^o_{\Delta_i}|D, X, \Psi, \varepsilon, \Omega_{-L^o_{\{m:m\in\Delta_i\}}}) &= p(L^o_{\Delta_i}|D, X, \Psi, Z^o, C, \mathcal{T}, \rho, \mu, \varepsilon, w, s) \\ &\propto p(L^o_{\Delta_i}) \prod_{m\in\Delta_i} \int_{g_m} \prod_n p(x_{mn}|d_{mn}, Z^o_m, L^o_m = L^o_{\Delta_i}, C_n, \mathcal{T}, w, \mu, \varepsilon, g_m) \\ &\quad \times \prod_{m\in\Delta_i} \prod_n p(d_{mn}|\psi_n, L^o_m = L^o_{\Delta_i}, C_n, \mathcal{T}, \rho, \mu, s, \varepsilon). \end{aligned}$$

(6) $\mathcal{T}$

the posterior distribution of $\mathcal{T}$ is:

$$\begin{aligned} p(\mathcal{T}|D, X, \Psi, \varepsilon, \Omega_{-\mathcal{T}}) &= p(\mathcal{T}|D, X, \Psi, C, L^o, Z^o, g, \rho, \mu, w, s, \varepsilon) \\ &\propto p(\mathcal{T}) \prod_{m,n} p(x_{mn}|d_{mn}, Z^o_m, L^o_m, C_n, \mathcal{T}, \mu, w, \varepsilon, g_m) p(d_{mn}|\psi_n, L^o_m, C_n, \mathcal{T}, \rho, \mu, s, \varepsilon) \end{aligned}$$

(7) ρ

the posterior distribution of ρ (with uniform prior on interval $[0, 1]$) is:

$$\begin{aligned} p(\rho|D, X, \Psi, \varepsilon, \Omega_{-\rho}) &= p(\rho|D, \Psi, L^o, C, \mathcal{T}, \mu, s, \varepsilon) \\ &\propto p(\rho) \prod_{m,n} p(d_{mn}|\psi_n, L^o_m, C_n, \mathcal{T}, \rho, \mu, s, \varepsilon) \end{aligned}$$

(8) μ

the posterior distribution of μ (with uniform prior on interval $[0, 1]$) is:

$$\begin{aligned} p(\mu|D, X, \Psi, \varepsilon, \Omega_{-\mu}) &= p(\mu|D, X, \Psi, L^o, Z^o, C, g, \rho, w, s, \varepsilon) \\ &\propto p(\mu) \prod_{m,n} p(x_{mn}|d_{mn}, Z^o_m, L^o_m, C_n, \mathcal{T}, \mu, w, \varepsilon, g_m) p(d_{mn}|\psi_n, L^o_m, C_n, \mathcal{T}, \rho, \mu, s, \varepsilon) \end{aligned}$$

(9) s

the posterior distribution of s (with Gamma prior) is:

$$\begin{aligned} p(s|D, X, \Psi, \varepsilon, \Omega_{-s}) &= p(s|D, \Psi, L^o, C, \mathcal{T}, \rho, \mu, \varepsilon) \\ &\propto p(s) \prod_{m,n} p(d_{mn}|\psi_n, L^o_m, C_n, \mathcal{T}, \rho, \mu, s, \varepsilon) \end{aligned}$$

(10) w

the posterior distribution of w (with Gamma prior) is:

$$\begin{aligned} p(w|D, X, \Psi, \varepsilon, \Omega_{-w}) &= p(w|D, X, L^o, Z^o, C, \mathcal{T}, g, \mu, \varepsilon) \\ &\propto p(w) \prod_{m,n} p(x_{mn}|d_{mn}, Z^o_m, L^o_m, C_n, \mathcal{T}, \mu, w, \varepsilon, g_m) \end{aligned}$$

(11) g

the inference of g with the maximum probability at each locus m is:

$$g_m = argmax_{g_m \in \{0,1\}} \prod_{n} p(x_{mn}|d_{mn}, Z^o_m, L^o_m, C_n, \mathcal{T}, w, \mu, \varepsilon, g_m)$$

Supplementary Note 4: Model selection

After performing inference on each fixed number of subclones (k), we need to solve the model selection problem to find the best k. Selecting the model with maximum likelihood will lead to overfitting, because the models with more subclones are more likely to yield an improved likelihood. The Bayesian Information Criterion (BIC), which proposed by Schwarz *et al.* (1978) as,

$$BIC = -2\ln f(x|\hat{\theta}) + p\ln n,$$

not only prefers the model with large loglikelihood values, but also adds the penalty on the number of free model parameters to punish the complexity of the model and shows heavily penalization on more complicated models for large samples. In addition, to avoid simulation-based maximization in performing model selection, we choose to apply a modified version of BIC (Carlin and Louis, 2009; Turkman *et al.*, 2019), defined as,

$$BIC = -2E_{\theta|x}[\ln f(x|\theta)] + p\ln n.$$

Supplementary Note 5: Simulation process

We first designed ground truth information under different group parameters, including the tree structure $\mathcal{T}$, the subclonal SNV genotype matrix Z and CNA genotype matrix L, which are displayed in Figure S1-S2 and S4-S6. We allocate cells almost evenly to each subclone, thereby obtaining the real cellular SNV and CNA genotypes according to the subclones they belong to. For each locus j in each cell i, we introduce allelic drop-out rate μ: we randomly lose an allele with the probability of μ. Then we use Negative Binomial distribution and Beta-Binomial distribution with sequencing errors ε and sequencing depths Ψ to generate total reads matrix D and mutant reads matrix X. Finally, we randomly select some sites in D and X with the probability ρ, and change the total reads and mutant reads to 0 to simulate missing events.

Supplementary Note 6: Evaluations of ARI

For comparison of the accuracy of subclone assignment, we used ARI (Rand, 1971; Qiu *et al.*, 2017) to measure the similarity between ground truth and estimation of C. Assume there are two partitions, $P^{(1)} = \{P_1^{(1)}, P_2^{(1)}, \cdots, P_r^{(1)}\}$ and $P^{(2)} = \{P_1^{(2)}, P_2^{(2)}, \cdots, P_s^{(2)}\}$ that divide set A into r and s groups, respectively. Let n_{ij} represents the overlap number in $P_i^{(1)}$ and $P_j^{(2)}$, that is, $n_{ij} = |P_i^{(1)} \cap P_j^{(2)}|$, then the set $\{n_{ij} | i \in \{1, \cdots, r\}, j \in \{1, \cdots, s\}\}$ describes overlaps between all possible pairs in $P^{(1)}$ and $P^{(2)}$. We then define the number of cells within group i from partition $P^{(1)}$, as $a_i = \sum_{j=1}^{s} n_{ij}$, and the number of cells within subclone j from partition $P^{(2)}$, as $b_j = \sum_{i=1}^{r} n_{ij}$. Then ARI of the two partitions can be calculated by

$$\mathrm{ARI}(P^{(1)}, P^{(2)}) = \frac{\sum_{ij} \binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}] / \binom{n}{2}}.$$

The value of ARI is in the region of [0,1]. A larger value indicates better assignment.

Supplementary Note 7: Preprocessing of raw data in ERBC datasets

In order to obtain mutant reads and total reads information from FASTQ files as the input of ***BiTSC***^{2}, the preprocessing steps of read alignment and SNV calling are described as follows. For read alignment, single-cell and bulk reads are aligned to human reference GRCh37 using the MEM algorithm in BWA software. Then, following a standardized best-practices pipeline, mapped reads are processed by filtering reads with low mapping-quality, performing local realignment around indels and removing PCR duplicates. For SNV calling, the single-cell SNV calls are obtained using Monovar software (Zafar *et al.*, 2016), a variant caller specifically designed for single-cell data, with default parameter settings. The SNV calls of bulk datasets are obtained by the paired-sample variant-calling approach implemented with VarDict software. The low-quality bulk SNV calls are removed by the SelectVariants tool in Genome Analysis Toolkit (GATK). Then the bulk SNVs passed QC are further divided into two categories: "germline" SNVs which present in both tumor and normal bulk samples, and "Somatics" SNVs which can only be found in tumor bulk samples. We exclude the small indels and other complex structural rearrangements in our final list of "gold-standard" bulk SNVs. Finally, the gold-standard SNVs present in single-cell calls are extracted, resulting in the mutant reads and total reads in 55 single cells with 1137 SNVs obtained. The corresponding accession codes can refer to the Supplementary Note of Alves and Posada (2018).

Supplementary Figures

Figure S1: The ground truth of simulation datasets in G1, containing subclonal phylogenetic tree ($\mathcal{T}$) and genotype matrix of CNA (L) and SNV (Z).

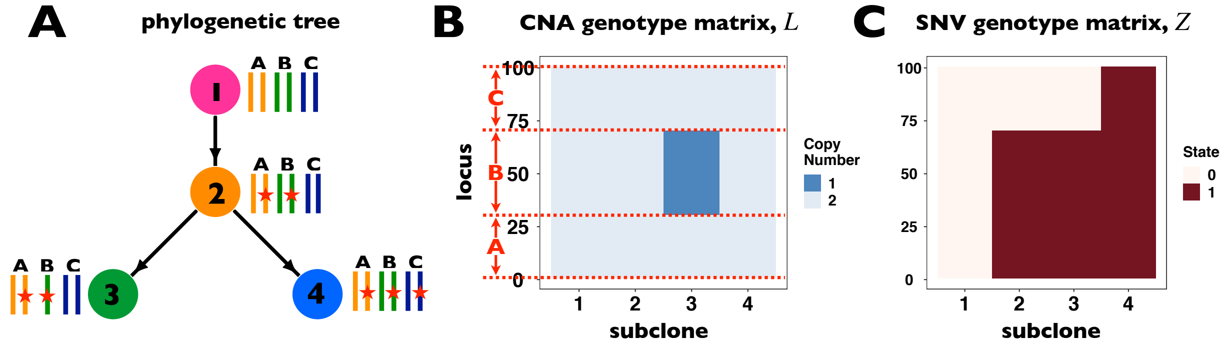

Fig. S1. The ground truth of simulation datasets in G1, containing phylogenetic tree ($\mathcal{T}$) and subclonal genotype matixes of CNA (L) and SNV (Z).

Figure S2: The ground truth of simulation datasets in G2, containing subclonal phylogenetic tree ($\mathcal{T}$) and genotype matrixes of CNA (L) and SNV (Z).

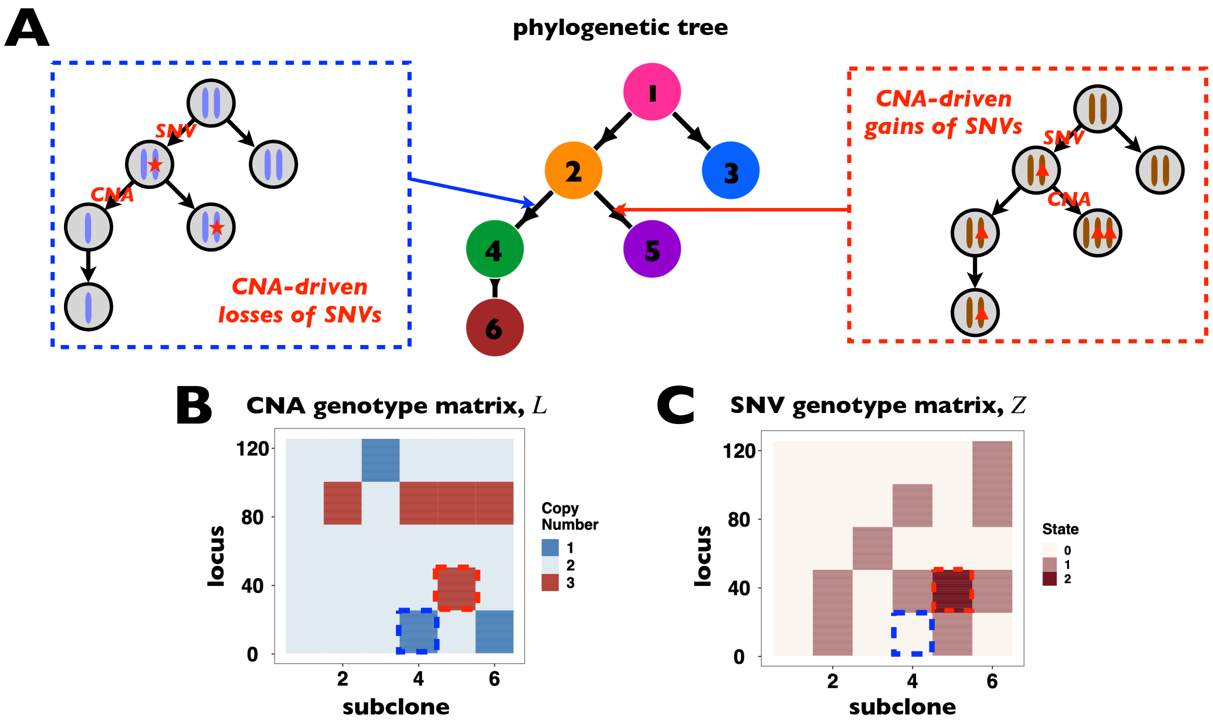

Fig. S2. The ground truth of simulation datasets in G2, containing phylogenetic tree ($\mathcal{T}$) and subclonal genotype matrixes of CNA (L) and SNV (Z). In panel A, the blue box visualizes the CNA-driven losses of mutations on a genomic segment happening in subclone4. The red box visualizes the CNA-driven gains of mutations on another genomic segment occurring in subclone5. Panels B and C are the true genotypes of CNA and SNV. The blue and red boxes in panels B and C highlight the corresponding changes in the genotypes of subclone4 and subclone5.

Figure S3: The clone tree can be reconstructed by SNV markers.

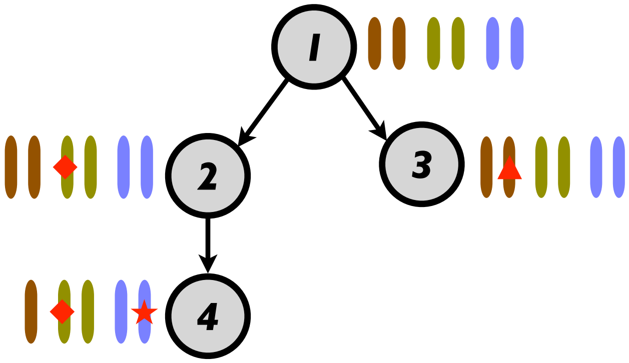

Fig. S3. The clone tree can be reconstructed by SNV markers. The copy loss in the first locus and the mutation in the third locus provide same information for recovering the true topology of the tree, then information of CNA is redundant with respect to the tree topology.

Figure S4: The ground truth of simulation datasets in G3-G5, containing subclonal phylogenetic tree ($\mathcal{T}$) and genotype matrix of CNA (L) and SNV (Z).

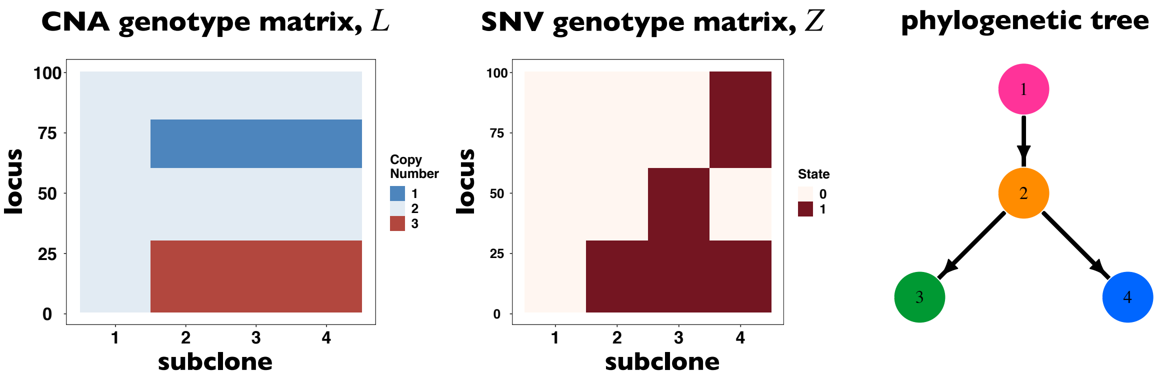

Fig. S4. The ground truth of simulation datasets in G3-G5, containing subclonal phylogenetic tree ($\mathcal{T}$) and genotype matrix of CNA (L) and SNV (Z).

Figure S5: The ground truth of simulation datasets in G6, containing subclonal phylogenetic tree ($\mathcal{T}$) and genotype matrix of CNA (L) and SNV (Z).

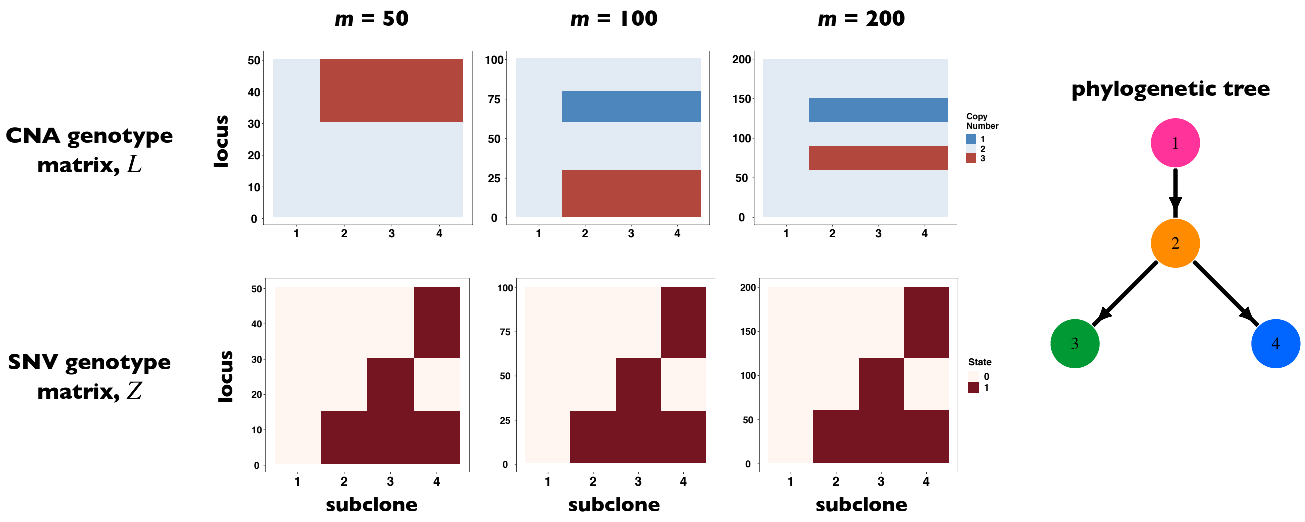

Fig. S5. The ground truth of simulation datasets in G6, containing subclonal phylogenetic tree ($\mathcal{T}$) and genotype matrix of CNA (L) and SNV (Z).

Figure S6: The ground truth of simulation datasets in G7, containing subclonal phylogenetic tree ($\mathcal{T}$) and genotype matrix of CNA (L) and SNV (Z).

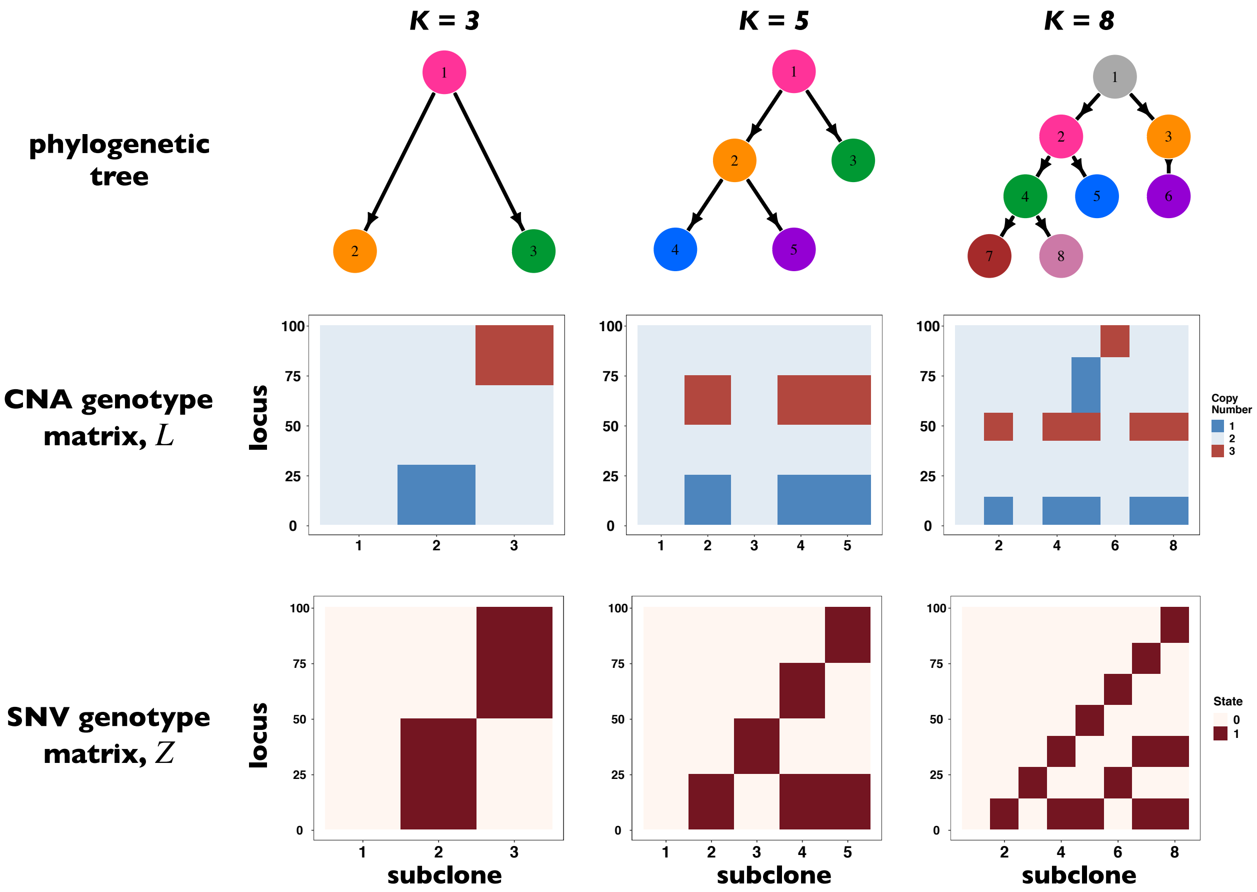

Fig. S6. The ground truth of simulation datasets in G7, containing subclonal phylogenetic tree ($\mathcal{T}$) and genotype matrix of CNA (L) and SNV (Z).

Figure S7: The simulated coverage heterogeneity in sequencing technologies under different sequencing depths.

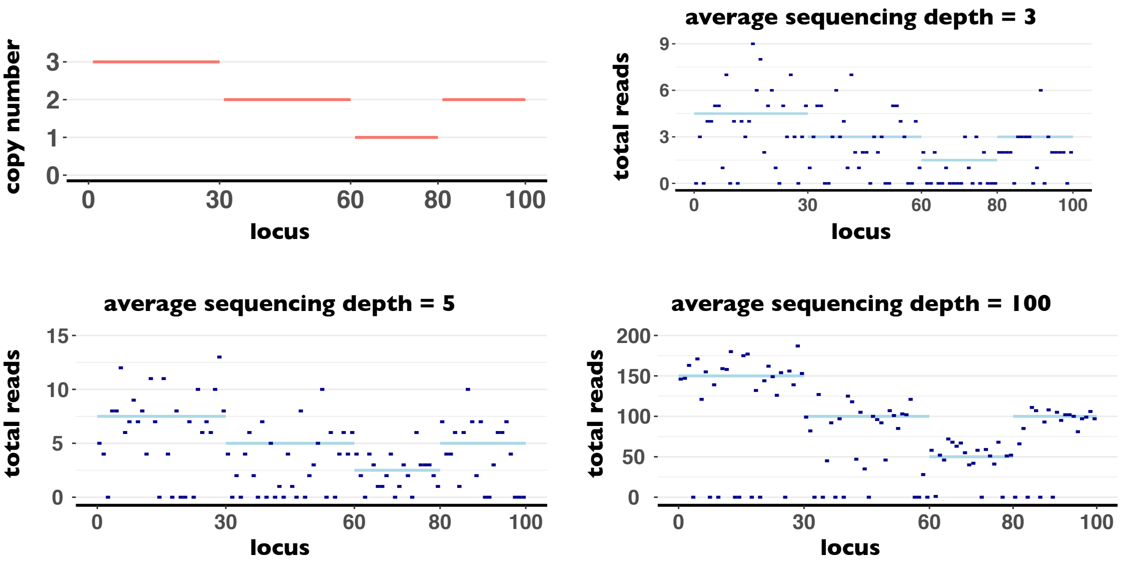

Fig. S7. The simulated coverage heterogeneity in sequencing technologies under different sequencing depths.

Figure S8: Comparison of performance on G1 for subclone assignment, subclonal CNA genotype recovery and tree reconstruction between *BiTSC*^{2} with true segment information as input and ground truth.

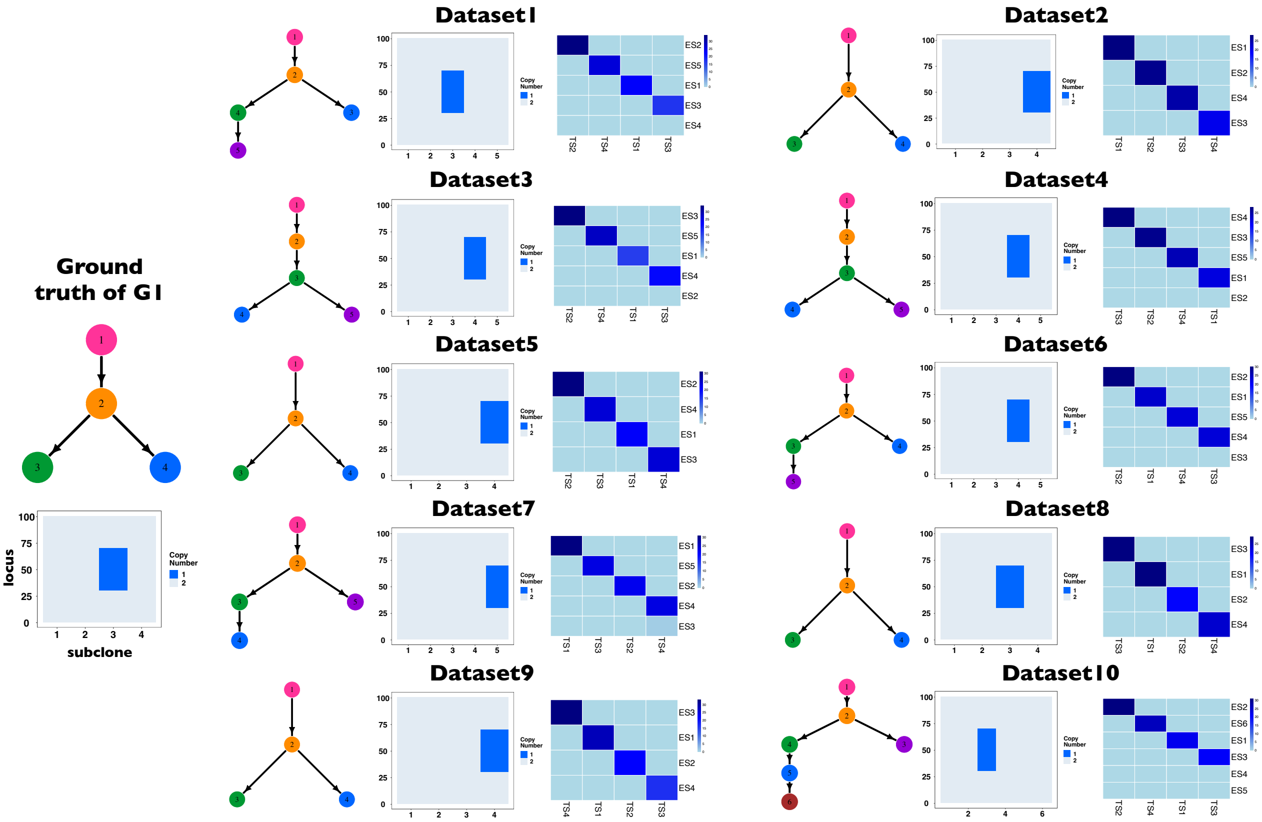

Fig. S8. Comparison of performance on G1 for subclone assignment, subclonal CNA genotype recovery and tree reconstruction between $BiTSC^2$ with true segment information as input and ground truth. On the far left of the figure, we show the ground truth information of G1, which includes the phylogenetic tree and the subclonal CNA genotype matrix. For each dataset, from left to right, we display the phylogenetic tree reconstructed by $BiTSC^2$ with true segment information as input, the CNA subclonal genotype matrix (the horizontal axis represents the subclone, the vertical axis represents the locus), and the heatmap of the corresponding relationship between true subclones and estimated subclones (the shade of the color indicates the number of cells overlapped by the true subclones and the estimated subclones, where the darker the color indicates the more overlapped cells, and the lighter the color indicates the less overlapping cells. ES stands for "estimated subclone", TS stands for "true subclone"). $BiTSC^2$ completely recovers the cellular CNA genotypes for G1.

Figure S9: Comparison of performance on G2 for subclone assignment, subclonal CNA genotype recovery and tree reconstruction between *BiTSC*^{2} with true segment information input and ground truth.

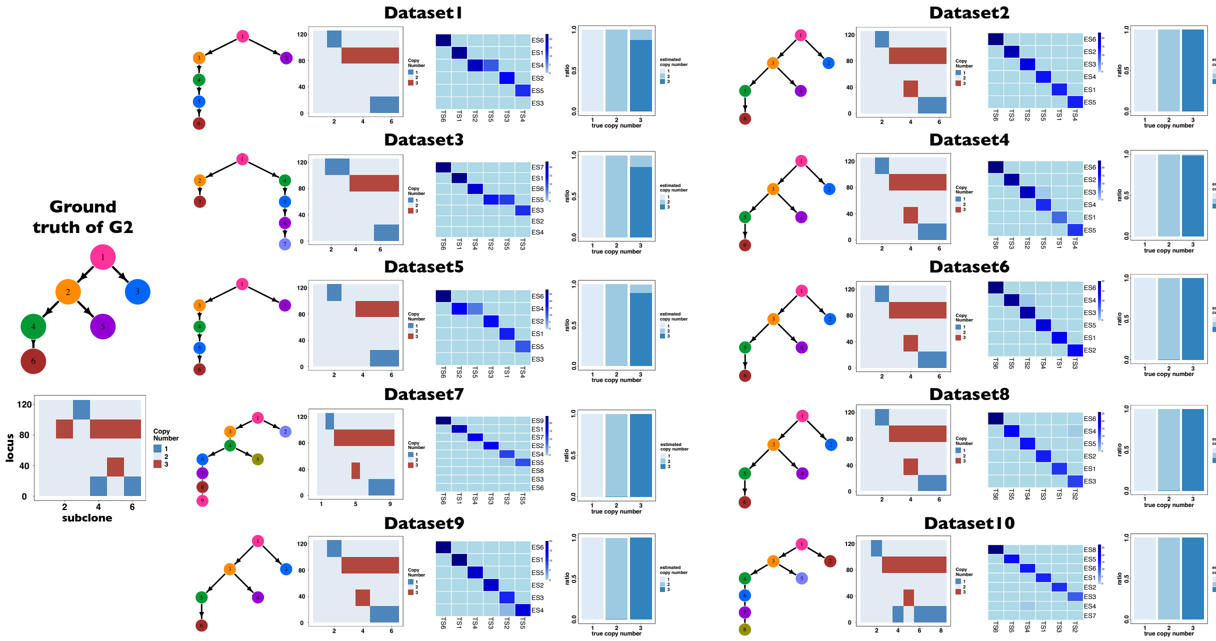

Fig. S9. Comparison of performance on G2 for subclone assignment, subclonal CNA genotype recovery and tree reconstruction between *BiTSC*^{2} with true segment information input and ground truth. On the far left of the figure, we show the ground truth information of G2, which includes the phylogenetic tree and the subclonal CNA genotype matrix. For each dataset, from left to right, we display the phylogenetic tree reconstructed by *BiTSC*^{2} with true segment information as input, the CNA subclonal genotype matrix (the horizontal axis represents the subclone, the vertical axis represents the locus), the heatmap of the corresponding relationship between true subclones and estimated subclones (the shade of the color indicates the number of cells overlapped by the true subclones and the estimated subclones, where the darker the color indicates the more overlapped cells, and the lighter the color indicates the less overlapping cells. ES stands for ”estimated subclone”, TS stands for ”true subclone”) and the mapping of true copy number and estimated copy number. Beside replicates 1, 3 and 5, which have 87.5%, 85.7% and 89.2% of true positives when the true copy number is 3, all other replicates have near or equal to 100% true positive rate. The accuracies of the recovered cell CNA genotypes for each datasets are 98.2%, 100%, 97.8%, 99.8%, 98.4%, 99.8%, 99.8%, 99.8%, 99.6% and 100%, respectively, with a mean of 99.32%.

Figure S10: The estimation of phase indictor g by $\boldsymbol{BiTSC^2}$ with true segment information as input for dataset G2.

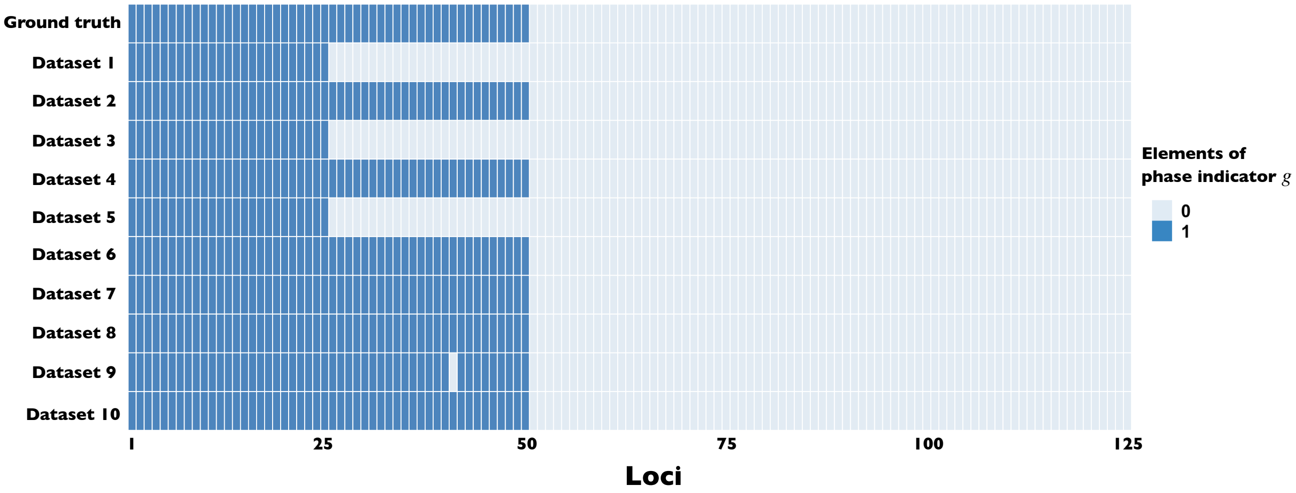

Fig. S10. The estimation of phase indictor g by $BiTSC^2$ with true segment information as input for dataset G2.

Figure S11: Comparison of performance on G1 and G2 for scSNV genotype recovery, subclone assignment and tree reconstruction among *BiTSC*^{2} with locus specific segments as input, RobustClone and BEAM.

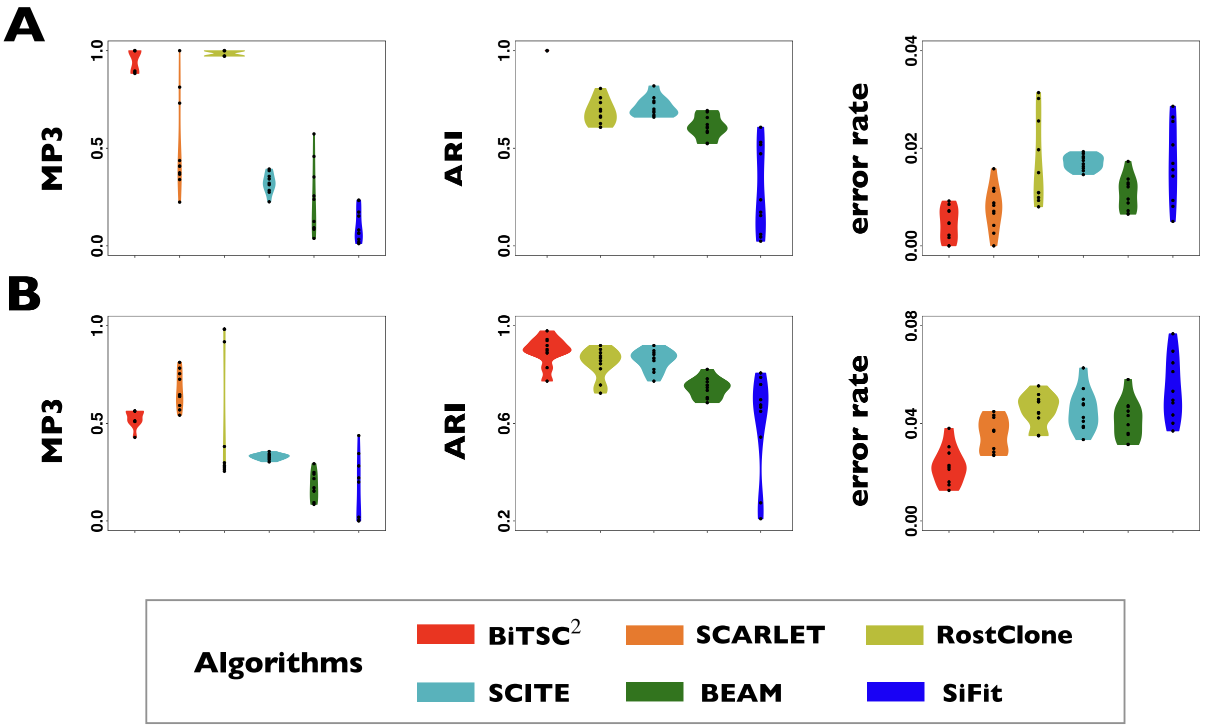

Fig. S11. Comparison of performance on G1 and G2 for scSNV genotype recovery, subclone assignment and tree reconstruction among *BiTSC*^{2} with locus specific segments as input, Robust-Clone and BEAM. **(A)** The violin plot of the algorithms for error rate of recovered scSNV genotype matrix, ARI of subclone assignment and MP3 similarity on G1 dataset. **(B)** The violin plot of the algorithms for error rate of recovered scSNV genotype matrix, ARI of subclone assignment and MP3 similarity on G2 dataset.

Figure S12: Comparison of overall performance on G3-G7 for scSNV genotype recovery, subclone assignment and tree reconstruction among *BiTSC*^{2} with real segment information as input, RobustClone and BEAM.

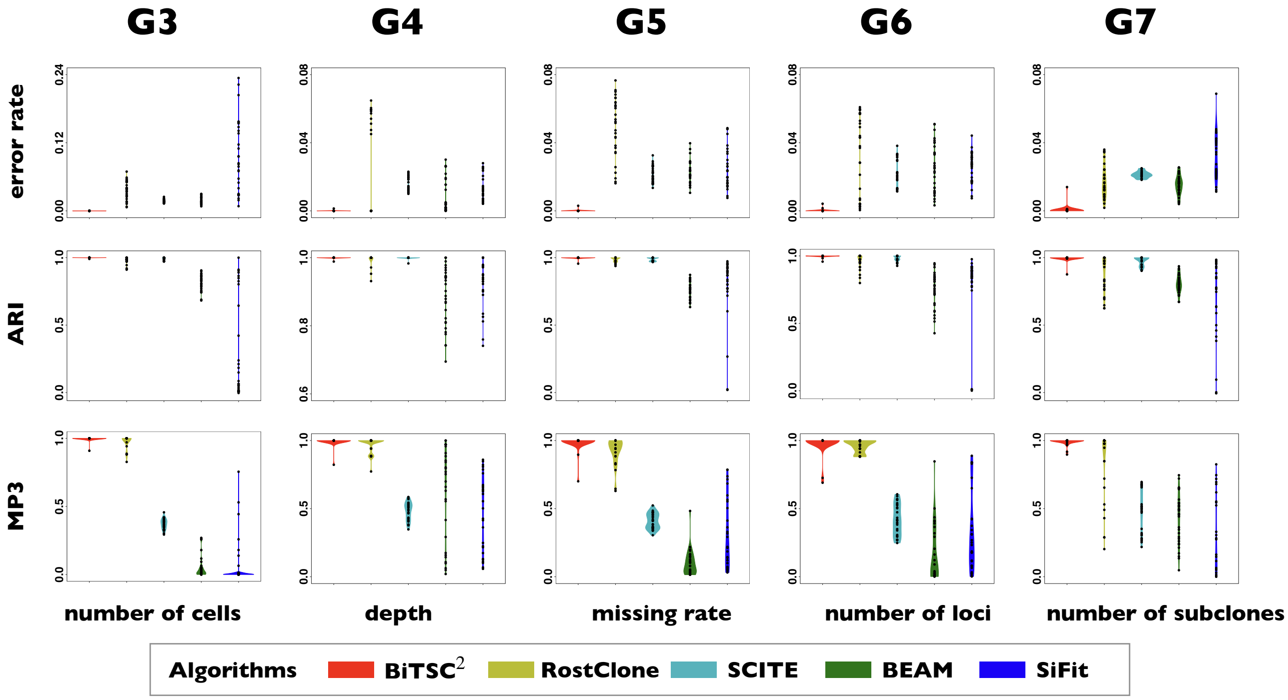

Fig. S12. Comparison of overall performance on G3-G7 for scSNV genotype recovery, subclone assignment and tree reconstruction among *BiTSC*^{2} with real segment information as input, RobustClone and BEAM.

Figure S13: Comparison of detail performance on G3-G7 for scSNV genotype recovery, subclone assignment and tree reconstruction among *BiTSC*^{2} with locus specific segments as input, RobustClone and BEAM.

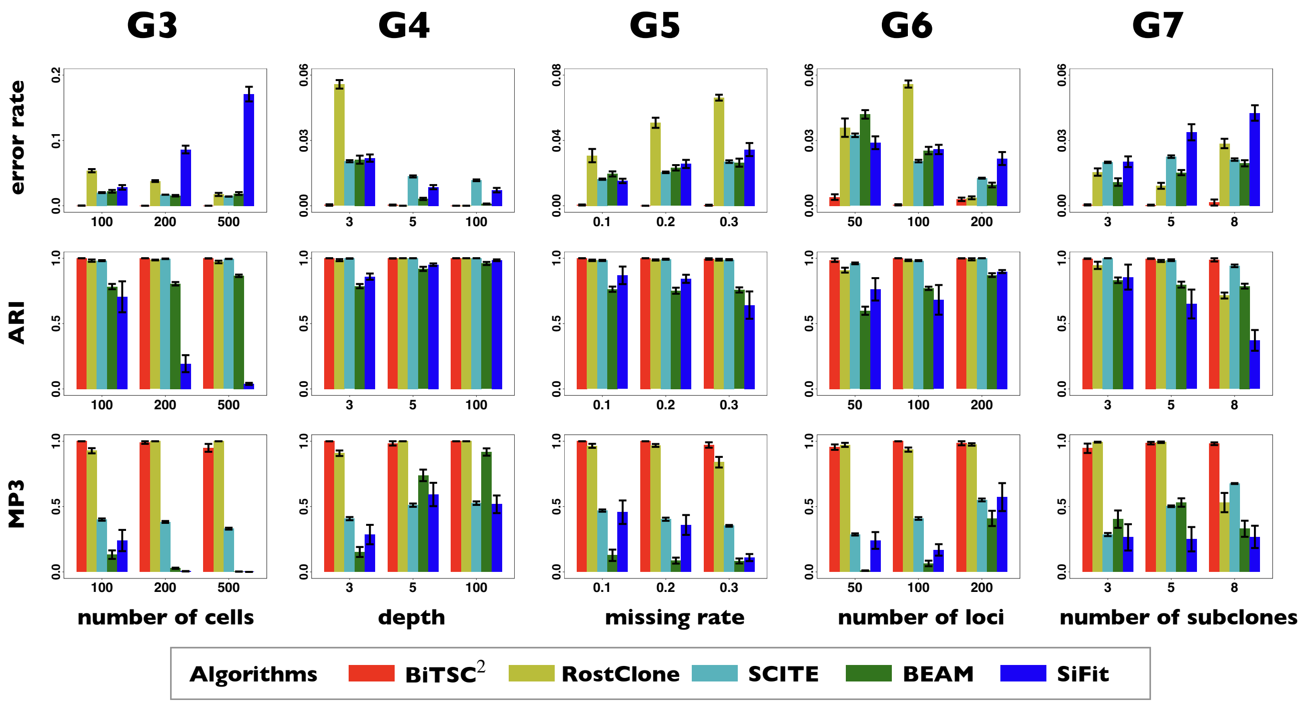

Fig. S13. Comparison of detail performance on G3-G7 for scSNV genotype recovery, subclone assignment and tree reconstruction among *BiTSC*^{2} with locus specific segments as input, Robust-Clone and BEAM.

Figure S14: Comparison of overall performance on G3-G7 for scSNV genotype recovery, subclone assignment and tree reconstruction among $BiTSC^2$ with locus specific segments as input, RobustClone and BEAM.

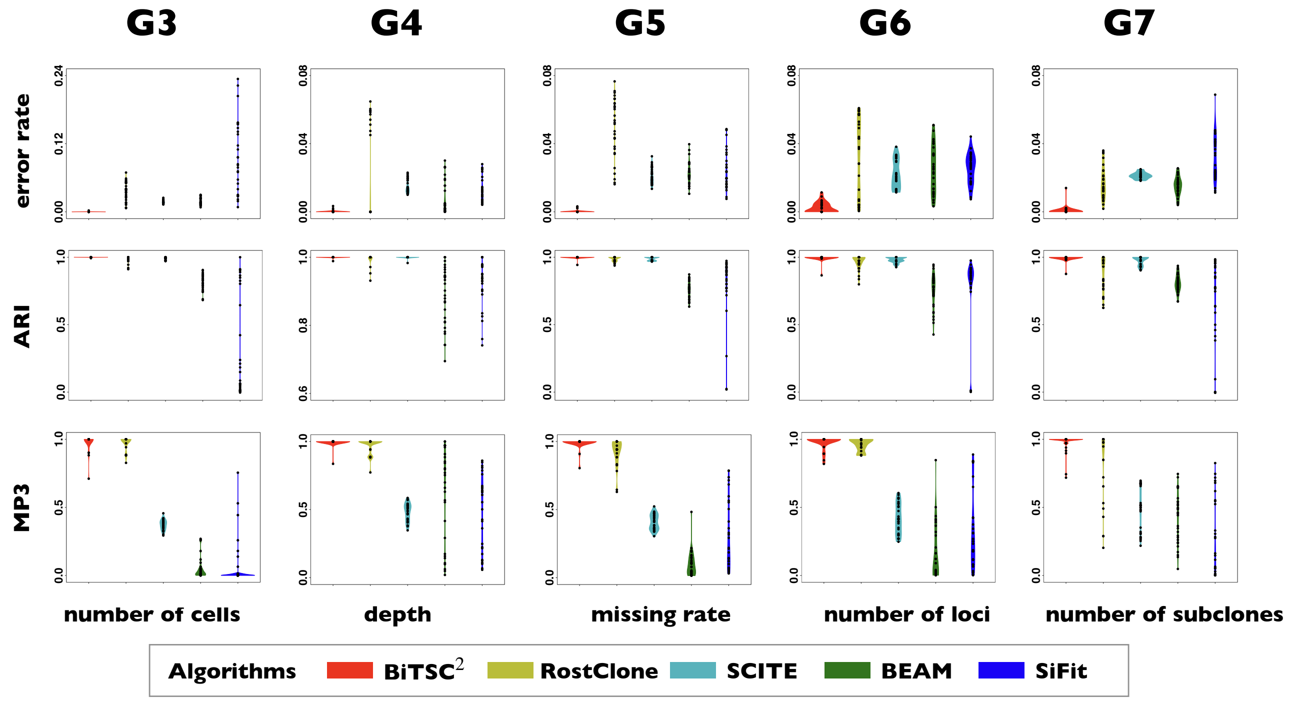

Fig. S14. Comparison of overall performance on G3-G7 for scSNV genotype recovery, subclone assignment and tree reconstruction among $BiTSC^2$ with locus specific segments as input, RobustClone and BEAM.

Figure S15: The BIC of metastatic colorectal cancer data calculated in Model selection step.

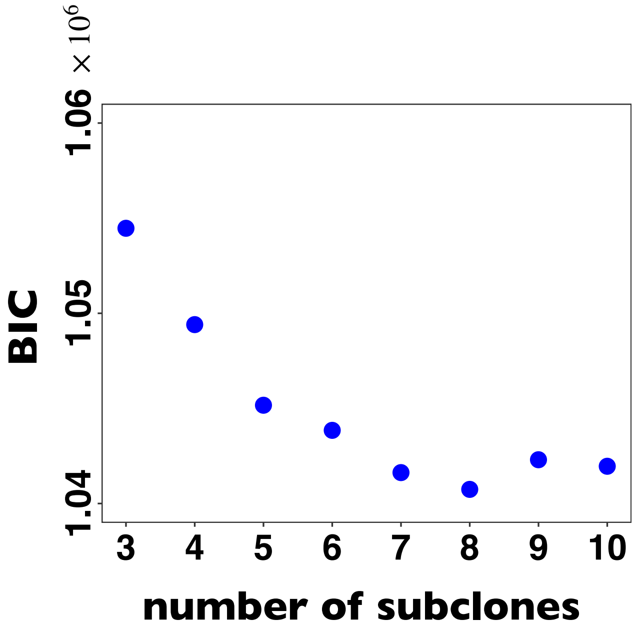

Fig. S15. The BIC of metastatic colorectal cancer data calculated in Model selection step.

Figure S16: The phylogeny tree inferred by SCITE and SCARLET on the dataset from a Metastatic Colorectal Cancer Patient.

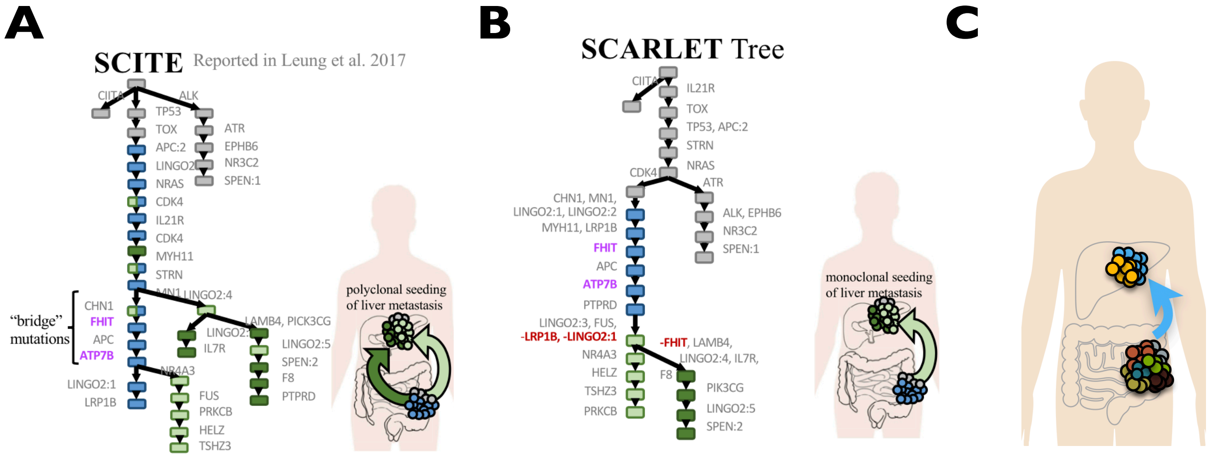

Fig. S16. The phylogeny tree inferred by SCITE and SCARLET on the dataset from a Metastatic Colorectal Cancer Patient. **(A)**The phylogeny tree inferred by SCITE in Leung *et al.* (2017), where two distinct branches of metastatic cells suggest polyclonal seeding of liver metastasis. **(B)** The phylogeny tree inferred by SCARLET with all metastatic cells contained in a single branch, which suggests monoclonal seeding of the liver metastasis. **(C)** The phylogeny tree inferred by $BiTSC^2$ in Figure 5A suggests monoclonal seeding of the liver metastasis. Figure AB are adapted from Satas *et al.* (2020).

Figure S17: The BIC of ERBC dataset calculated in model selection step.

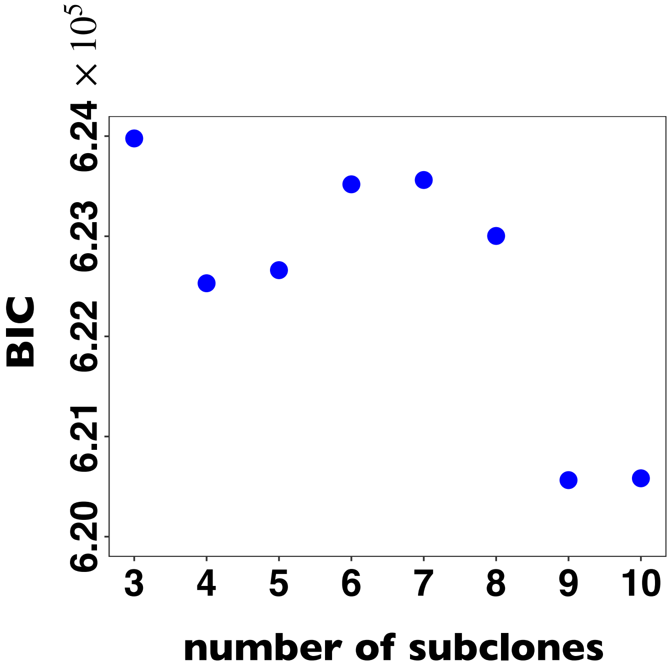

Fig. S17. The BIC of ERBC dataset calculated in model selection step.

Figure S18: The missing rate ρ estimated by $BiTSC^2$ on the datasets from G5 groups.

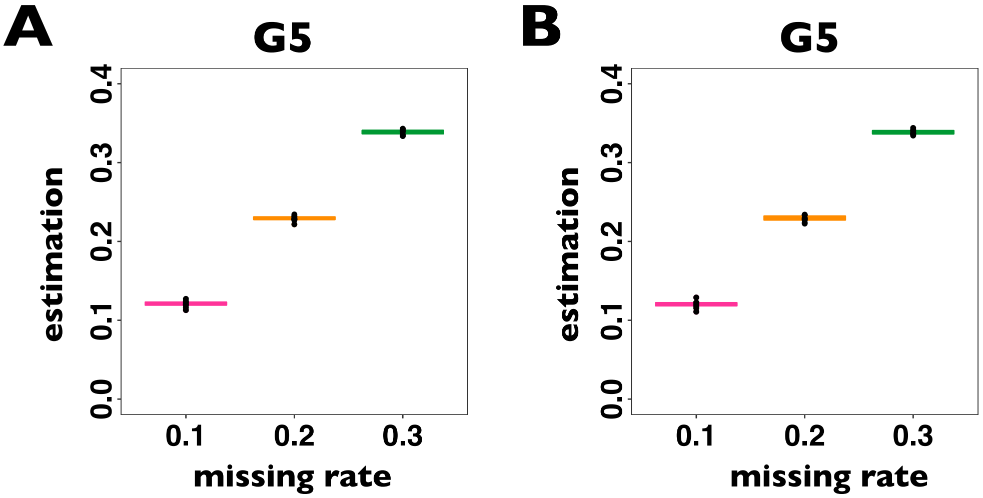

Fig. S18. The missing rate ρ estimated by $BiTSC^2$ with real segmentation as input **(A)** and with locus specific segments as input **(B)** on the datasets from G5 groups.

Figure S19: The ground truth where SNV randomly occurs on chromosomes.

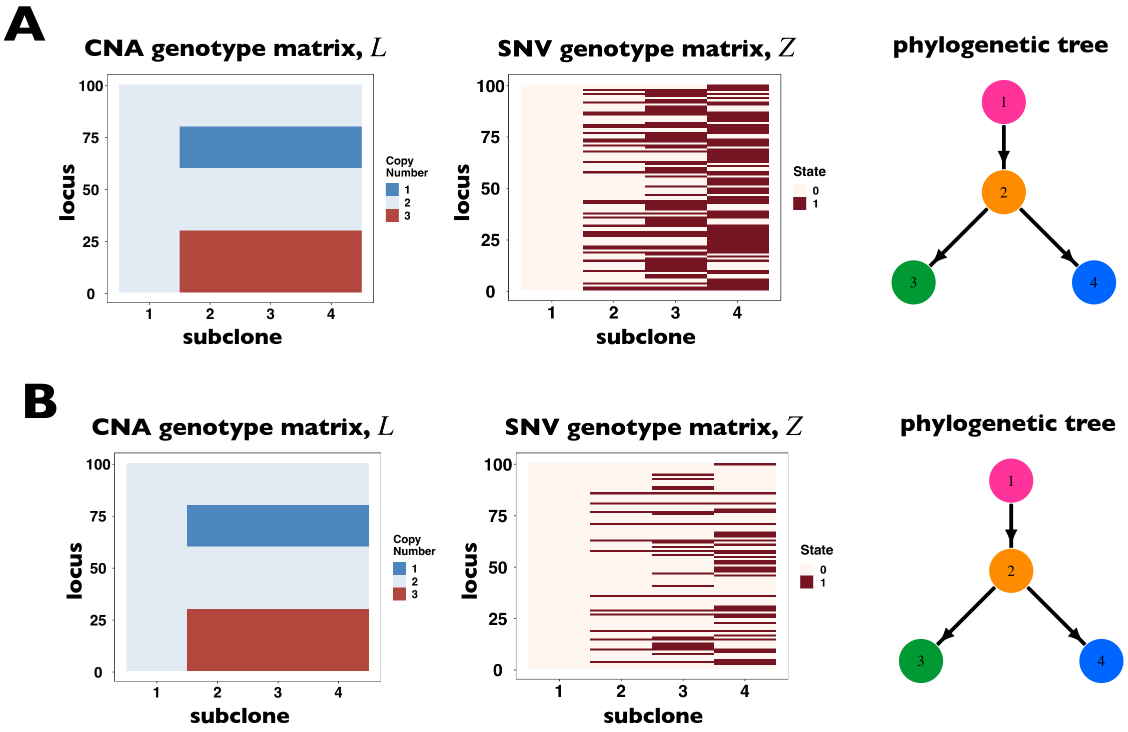

Fig. S19. The ground truth where SNV randomly occurs on chromosomes. (A) The ground truth where SNVs randomly and uniformly occurs on all genomic regions. (B) The ground truth where SNVs are randomly and sparsely distributed within each CNA segment.

Figure S20: The tree inferred by SCARLET and $BiTSC^2$ for the toy model in Figure 1A.

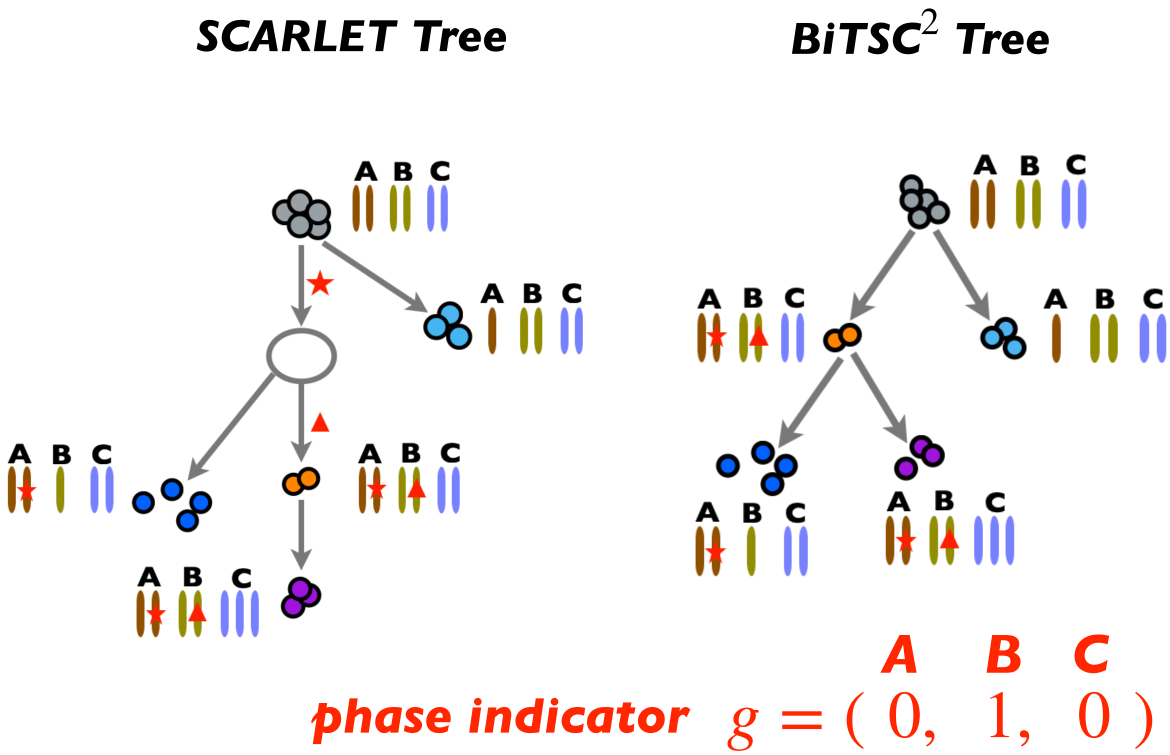

Fig. S20. The tree inferred by SCARLET and $BiTSC^2$ for the toy model in Figure 1A.

Supplementary Tables

Table S1: The definitions of all parameters and examples of main parameters in Figure 1A.

Table S1: The definitions of all parameters and examples of main parameters in Figure 1A. Assume that the input matrixes consist of N cells measured at M loci and there exist K latent subclones in the cells drawn for sequencing ($K \ll N$).

Parameters	Definitions	Value in Figure 1A
Z^o	SNV origin matrix with dimensions of $M \times 2$. For the m-th row of Z^o, $Z^o_m = (k, v)$ indicates mutation at locus m occurs from subclone k and gains v mutant copies.	$Z^o = \begin{pmatrix} 2 & 1 \\ 2 & 1 \\ 0 & 0 \end{pmatrix}$
L^o	CNA origin matrix with dimensions of $M \times 2$. For the m-th row of L^o, $L^o_m = (k, v)$ indicates the CNA at locus m arises in subclone k and gains (or losses if v is negative) v normal or mutant copies.	$L^o = \begin{pmatrix} 3 & -1 \\ 4 & -1 \\ 5 & 1 \end{pmatrix}$
g	Phase indicator vector with length M. For the m-th element of g, $g_m = 1$ indicates CNA happened on the mutant allele at locus m, and $g_m = 0$ otherwise.	$g = (0,\ 1,\ 0)$
$\mathscr{T}$	Clone tree vector with length K. For the i-th element of $\mathscr{T}$, $\mathscr{T}_i = k$ indicates the parent of subclone i is subclone k.	$\mathscr{T} = (0,\ 1,\ 1,\ 2,\ 2)$
Z	SNV subclone genotypes matrix with dimensions of $M \times K$. The element at the i-th row and j-th column of Z, Z_{ij}, represents the number of mutant copies at the i-th locus of the j-th subclone.	$Z = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$
L	CNA subclone genotypes matrix with dimensions of $M \times K$. The element at the i-th row and j-th column of L, L_{ij}, represents the total number of copies at the i-th locus of the j-th subclone.	$L = \begin{pmatrix} 2 & 2 & 1 & 2 & 2 \\ 2 & 2 & 2 & 1 & 2 \\ 2 & 2 & 2 & 2 & 3 \end{pmatrix}$
C	Subclone assignment vector with length N. For the n-th element of C, $C_n = k$ indicates the cell n is from subclone k.	$C = (1,\ 2,\ 1,\ 1,\ 4,\ 5,\ 4,\ 4,\ 1,\ 5,\ 1,\ 4,\ 3,\ 2,\ 5,\ 3,\ 3)$
ϕ	The parameter of the categorical distribution.	$\phi = (\frac{5}{17}, \frac{2}{17}, \frac{3}{17}, \frac{4}{17}, \frac{3}{17})$

Parameters	Definitions	Value in Figure 1A
θ	Parameters vector with length K. $\theta_k \sim \text{Gamma}(\gamma,1)$, and $\phi_k = \theta_k / \sum_{i=1}^{K} \theta_i$	–
γ	The parameter of Gamma distribution.	–
s	The dispersion parameter of Negative Binomial distribution.	–
w	The dispersion parameter of Beta-Binomial distribution.	–
ρ	The zero-inflation parameter of zero-inflated negative binomial (ZINB) distribution, i.e., missing rate.	–
ε	Sequencing error rate.	–
μ	Allelic dropout rate.	–
ψ	Sequencing depth vector with length N.	–
π	Prior probability for a segment with no CNA.	–
α, β	The given hyperparameters of Beta distribution.	–
ζ	The somatic point mutation rate.	–

Table S2: The setting of simulation parameters for comparison data.

Table S2: The setting of simulation parameters for comparison data.

	change factor			control factors				
G3	n			ψ	ρ	m	K	μ
	100	200	500	3	0.2	100	4	0.1
G4	ψ			n	ρ	m	K	μ
	3	5	100	100	0.2	100	4	0.1
G5	ρ			n	ψ	m	K	μ
	0.1	0.2	0.3	100	3	100	4	0.1
G6	m			n	ψ	ρ	K	μ
	50	100	200	100	3	0.2	4	0.1
G7	K			n	ψ	ρ	m	μ
	3	5	8	100	3	0.2	100	0.1

Table S3: Prior distribution parameters setting of $\boldsymbol{BiTSC^2}$ for simulations G1-G7.

Table S3: Prior distribution parameters setting of $BiTSC^2$ for simulations G1-G7.

Parameters	Value
Maximum possible mutant copies (M_s)	1
Maximum possible total copies (M_c)	4
Dirichlet prior parameter of ϕ (γ)	1.5
Beta prior parameter of π (α, β)	(10000,1)
the standard deviation of the Normal distribution used to propose a new w and s	18
the shape parameter of the prior Gamma distribution of w and s	11
the rate parameter of the prior Gamma distribution of w and s	0.1
prior parameter of Z^o (ζ)	0.01

Table S4: MCMC sampling parameters setting of $\boldsymbol{BiTSC^2}$ for simulations G1-G7.

Table S4: MCMC sampling parameters setting of $BiTSC^2$ for simulations G1-G7.

Parameters	Value
Number of chains	5
Temperature increment (ΔT)	0.35
Sample size for posterior inference	500
Burn-in sample size	500
Sample size for tuning adaptive parameter	500
Interval to perform chain swap	30
Probability of tree slice sampling	0.15
Probability of tree Metropolis-Hastings sampling	0.85

Table S5: Prior distribution parameters setting of *BiTSC*^{2} for metastatic colorectal cancer dataset.

Table S5: Prior distribution parameters setting of $BiTSC^2$ for metastatic colorectal cancer dataset.

Parameters	Value
Maximum possible mutant copies (M_s)	1
Maximum possible total copies (M_c)	10
Dirichlet prior parameter of ϕ (γ)	1.5
Beta prior parameter of π (α, β)	(10000,1)
the standard deviation of the Normal distribution used to propose a new w and s	18
the shape parameter of the prior Gamma distribution of w and s	11
the rate parameter of the prior Gamma distribution of w and s	0.1
prior parameter of Z^o (ζ)	0.01

Table S6: MCMC sampling parameters setting of *BiTSC*^{2} for real datasets.

Table S6: MCMC sampling parameters setting of $BiTSC^2$ for real datasets.

Parameters	Value
Number of chains	5
Temperature increment (ΔT)	0.35
Sample size for posterior inference	1000
Burn-in sample size	1000
Sample size for tuning adaptive parameter	1000
Interval to perform chain swap	30
Probability of tree slice sampling	0.15
Probability of tree Metropolis-Hastings sampling	0.85

Table S7: Prior distribution parameters setting of *BiTSC*^{2} for breast cancer dataset.

Table S7: Prior distribution parameters setting of *BiTSC*^{2} for breast cancer dataset.

Parameters	Value
Maximum possible mutant copies (M_s)	1
Maximum possible total copies (M_c)	10
Dirichlet prior parameter of ϕ (γ)	1.5
Beta prior parameter of π (α, β)	(100,1)
the standard deviation of the Normal distribution used to propose a new w and s	18
the shape parameter of the prior Gamma distribution of w and s	11
the rate parameter of the prior Gamma distribution of w and s	0.1
prior parameter of Z^o (ζ)	0.01

Supplementary References

Alves, J. M. and Posada, D. (2018). Sensitivity to sequencing depth in single-cell cancer genomics. *Genome medicine*, **10**(1), 1–11.

Carlin, B. P. and Louis, T. A. (2009). Bayesian methods for data analysis. *Journal of the Royal Statal Society*, **172**(4), 935?936.

Leung, M. L., Davis, A., Gao, R., Casasent, A., Wang, Y., Sei, E., Vilar, E., Maru, D., Kopetz, S., and Navin, N. E. (2017). Single-cell dna sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome research*, **27**(8), 1287–1299.

Marass, F., Mouliere, F., Yuan, K., Rosenfeld, N., Markowetz, F., *et al.* (2016). A phylogenetic latent feature model for clonal deconvolution. *The Annals of Applied Statistics*, **10**(4), 2377–2404.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature methods*, **14**(10), 979.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, **66**(336), 846–850.

Satas, G., Zaccaria, S., Mon, G., and Raphael, B. J. (2020). Scarlet: Single-cell tumor phylogeny inference with copy-number constrained mutation losses. *Cell Systems*, **10**(4), 323–332.

Schwarz, G. *et al.* (1978). Estimating the dimension of a model. *The annals of statistics*, **6**(2), 461–464.

Turkman, M. A. A., Paulino, C. D., and Müller, P. (2019). *Computational Bayesian Statistics: An Introduction*, volume 11. Cambridge University Press.

Zafar, H., Wang, Y., Nakhleh, L., Navin, N., and Chen, K. (2016). Monovar: single-nucleotide variant detection in single cells. *Nature methods*, **13**(6), 505–507.

Zeng, L., Warren, J. L., Zhao, H., *et al.* (2019). Phylogeny-based tumor subclone identification using a bayesian feature allocation model. *Annals of Applied Statistics*, **13**(2), 1212–1241.