METHOD

SiFit: A Method for Inferring Tumor Trees from Single-Cell Sequencing Data under Finite-site Models

Hamim Zafar[1,2], Anthony Tzen[1], Nicholas Navin[2,3], Ken Chen[2] and Luay Nakhleh[1*]

*Correspondence:
nakhleh@rice.edu
[1]Department of Computer Science, Rice University, Houston, Texas, USA
Full list of author information is available at the end of the article

Abstract

Background: Tumor phylogenies provide insightful information on intra-tumor heterogeneity and evolutionary trajectories. Single-cell sequencing (SCS) enables the inference of tumor phylogenies and methods were recently introduced for this task under the infinite-sites assumption.

Results: Violations of this assumption, due to chromosomal deletions and loss of heterozygosity, necessitate the development of statistical inference methods that utilize finite-site models. We propose a statistical inference method for tumor phylogenies from noisy SCS data under a finite-sites model. We demonstrate the performance of our method on synthetic and biological data sets.

Conclusion: Our results suggest that employing a finite-sites model leads to improved inference of tumor phylogenies.

Keywords: Tumor evolution; Intra-tumor heterogeneity; Single-cell sequencing; Finite-site model; Phylogenetic tree

Background

Intra-tumor heterogeneity, as caused by a combination of mutation and selection [1–4], poses significant challenges to the diagnosis and clinical therapy of cancer [5–8]. This heterogeneity can be readily elucidated and understood if the evolutionary history of the tumor cells was known. This knowledge, alas, is not available, since genomic data is most often collected from one snapshot during the evolution of the tumor's constituent cells. Consequently, using computational methods that reconstruct the tumor phylogeny from sequence data is the approach of choice. However, while intra-tumor heterogeneity has been widely studied, the inference of a tumor's evolutionary history remains a daunting task.

Most studies to-date relied on bulk high-throughput sequencing data, which represents DNA extracted from a tissue consisting millions of cells [9–13]. As a result, the admixture signal obtained from such data represents an average of all the distinct subpopulations present in the tumor [14]. This ambiguity makes it difficult to identify the lineage of the tumor from the mixture. In such cases, phylogenetic reconstruction requires a deconvolution of the admixture signal to identify the taxa of the tree [15–17]. This type of data is low-resolution and can not depict cell-to-cell variability that is needed for inference of tumor evolution [14, 18]. Another approach for resolving intra-tumor heterogeneity and reconstructing tumor phylogeny is multi-region sequencing, in which, DNA sampled from multiple spatially

separated regions of the tumor are sequenced [19, 20], however, this approach is restricted to cases where the subpopulations are geographically segregated and can not resolve spatially intermixed heterogeneity [21].

Single-cell DNA sequencing: promises and challenges

With the advent of single-cell DNA sequencing (SCS) technologies, high-resolution data are becoming available, which promise to resolve intra-tumor heterogeneity to a single-cell level [14, 18, 22–25]. These technologies provide sequencing data from single cells, thus allowing for the reconstruction of the cell lineage tree. However, high error rates associated with single-cell sequencing data significantly complicates this task.

The whole-genome amplification (WGA) process, a crucial step in producing single-cell sequencing data, introduces different types of noises that result in erroneous genotype inferences. The prominent WGA errors include: allelic dropout (ADO) errors, false positive errors (FPs), non-uniform coverage distribution and low coverage regions [14]. Allelic dropout is a prominent error in SCS data, which contributes a considerable amount of false negatives in point mutation datasets. ADO is responsible for falsely representing the heterozygous genotypes as homozygous ones and the extent of such errors varies from 0.0972 to 0.43 as reported in different SCS-based studies [22–26]. Even though variant callers have been proposed for reducing ADO errors [27], the extent of such errors is still large. Different single-cell sequencing studies have reported false positive rates varying from 1.2×10^{-6} to 6.7×10^{-5} [22–26], the number of occurrences of which can essentially exceed the number of true somatic mutations. Often consensus-based approach is taken to reduce the number of false positive errors [26–28], in which, variants only observed in more than one single cell are considered. The variants observed in only one single cell are treated as errors and removed. In doing so, this approach also removes the true biological variants unique to a cell whereas, sites of recurrent errors persist. Both ADO and coverage non-uniformity result in unobserved sites. Often more than 50% of the genotypes are reported as missing due to the low quality of SCS data and thus no information regarding the mutation status of that site is conveyed [22].

Existing work

Single-cell-based studies for delineating the tumor phylogeny rely on the single-cell somatic SNV profiles, which are confounded by the technical errors in single-cell sequencing. Even though such errors prohibit the use of classic phylogenetic approaches, many studies have used them. Distance-based methods like UPGMA and neighbor joining have been used by Yu *et al.* [29], and Xu *et al.* [23] respectively. Eirew *et al.* [30] used a popular Bayesian phylogenetic inference tool, MRBAYES [31], for inferring evolutionary history. However, none of these methods account for the SCS specific errors.

BitPhylogeny [32] is a non-parametric Bayesian approach that uses a tree-structured mixture model to infer intra-tumor phylogeny. Even though such an approach is valuable for identifying subclones from bulk sequencing data, it is not suitable in the context of present-day single-cell datasets (fewer than 100

cells) [22–24, 26, 29], which do not provide sufficient data required by the mixture model in order to converge to the target distribution [33]. Furthermore, BitPhylogeny is a flexible framework that can fit different data types but does not specifically model the single-cell errors.

SCITE [34] and OncoNEM [33] are two computational tools that were specifically designed for inference of tumor evolution from SCS data. SCITE is an MCMC algorithm that allows one to infer maximum likelihood tree from imperfect genotype matrix of SCS. It infers the evolutionary history as a mutation tree, proposed by Kim and Simon in [35]. A mutation tree shows the chronological order of the mutations that occur during tumor development. OncoNEM is a likelihood-based method that employs a heuristic search algorithm to find the maximum likelihood clonal tree, a condensed tree that represents the evolutionary relationship between the subpopulations in the data. OncoNEM clusters the cells together into clones and also infers unobserved populations that can improve the likelihood. Both of these methods probabilistically account for technical errors in SCS data and can also estimate the error rates of SCS data. However, both SCITE and OncoNEM suffer by making inferences under the "infinite sites assumption", which posits that each site in the dataset mutates at most once during the evolutionary history [36] and the taxa form a perfect phylogeny [37]. This assumption often gets violated in human tumors due to different events such as chromosomal deletions, loss of heterozygosity (LOH) and convergent evolution [38]. Furthermore, OncoNEM infers clonal trees where cell-to-cell evolution is not displayed, and SCITE is concerned with the order of mutation in the tree but not the lineage of single cells. To the best of our knowledge, there is no method that infers a phylogenetic tree from SCS data under finite-site model of evolution while accounting for the technical errors in SCS.

SiFit

Here we propose SiFit, a likelihood-based approach for inferring tumor trees from imperfect SCS genotype data with potentially missing entries, under finite-site model of evolution. To account for the errors in SCS, SiFit extends the error model of SCITE and OncoNEM. This extension accommodates for the possible genotypes that are excluded by infinite sites model. SiFit also extends the Jukes-Cantor model of evolution [39] to adopt it for cancer phylogeny for single-cell data. SiFit employs a heuristic search algorithm to find the phylogenetic tree that is most likely to produce the observed SCS data. We evaluate SiFit on a comprehensive set of simulated data, where it performs superior to the existing methods in terms of tree reconstruction. Application of SiFit to experimental datasets shows how infinite sites assumption gets violated in real SCS data and how SiFit's reconstructed tumor phylogenies are more comprehensive compared to phylogenies reconstructed under infinite sites assumption. SiFit achieves a major advance in understanding tumor phylogenies from single cells and is applicable to wide variety of available single-cell DNA sequencing datasets.

Results and discussion

Overview of SiFit

We start with a brief explanation of how SiFit infers a tumor phylogeny from noisy genotype data obtained from single-cell sequencing. The input data consist of the

following: (1) an $n \times m$ genotype matrix, which contains the observed genotypes for m single cells at n different loci, the genotype matrix can be binary or ternary depending on the data, and (2) the false positive rate (FPR), α and false negative rate (FNR), β. These error parameters can be learned from the data.

SiFit includes (1) a finite-site model of tumor evolution and an error model for SCS, based on which the likelihood score of a candidate phylogenetic tree and error rate can be quantified and (2) a heuristic algorithm for exploring the joint space of trees and error rates in search of optimal parameters.

SiFit outputs a phylogenetic tree describing the evolutionary relationship between the single cells and the estimated error rates. The single cells are placed at the leaves of the phylogenetic tree. A more detailed technical description of SiFit can be found in the "Methods" section.

Phylogenetic trees and model of tumor evolution

We assume that the observed single cells evolved according to an underlying phylogenetic tree. A phylogeny or phylogenetic tree represents the genealogical relationship among genes, species, populations, etc. [40]. In the context of tumor, it is a rooted binary tree that represents the genealogical relationship among a set of cells. The sequenced single cells are placed at the leaves of the phylogenetic tree. We also assume that the cells evolve according to a finite-site model along the branches of the tree.

The $n \times m$ true genotype matrix G contains the true genotypes of m single cells at n different loci. If the data only contains information about the presence or absence of a mutation at a locus, the matrix is binary, where the absence or presence of a mutation is represented by a 0 or 1 at the entry $G(i, j)$, respectively. Assuming the cells to be diploid, if the data differentiates between heterozygous and homozygous mutations, the genotype matrix is ternary, where a 0, 1 or 2 at entry $G(i, j)$ denotes homozygous reference, heterozygous or homozygous non-reference genotype, respectively. Heterozygous or homozygous non-reference genotypes represent mutations. This ternary representation facilitates the use of mutation profile from modern variant calling algorithms (e.g., Monovar [27] and GATK [41]) that report mutation status of a sample in terms of genotypes.

For the finite-site model of evolution, we extend the Jukes-Cantor model of DNA sequence evolution [39] to accommodate for single-cell data. Adoption of this finite-site model of evolution enables us to account for convergent evolution or reversal of genotypes that are excluded by methods that make the "infinite sites assumption" (SCITE and OncoNEM). OncoNEM also assumes only binary data and does not differentiate between heterozygous and homozygous mutations. This binarization of data might result in loss of information for a dataset with ternary genotypes as heterozygous and homozygous non-reference genotypes can not be distinguished when data is binarized. On the other hand, SCITE assumes that the observation of a homozygous non-reference genotype is due to technical errors only. These assumptions follow from using the infinite sites model and are not made by SiFit.

SCITE also removes the mutations that are present in all cells or in one cell as non-informative in tree reconstruction. SiFit does not remove such mutations as these can be informative in the computation of the likelihood under finite-site models.

Model of single-cell errors

The observed genotype matrix, denoted by D, is an imperfect noisy version of the true genotype matrix G. The false positive errors and the false negative errors are responsible for adding noise in the observed genotype matrix. Considering binary genotype data, false positive errors result in observing a 1 with probability α when the true genotype is 0. Similarly, due to false negative errors, with probability β, we will observe a 0, instead of a 1. These relationships between true and observed genotype matrix are given by

$$Pr(D_{i,j}|G_{i,j}) = \begin{cases} 1-\alpha & \text{if } D_{i,j}=0, G_{i,j}=0 \\ \beta & \text{if } D_{i,j}=0, G_{i,j}=1 \\ \alpha & \text{if } D_{i,j}=1, G_{i,j}=0 \\ 1-\beta & \text{if } D_{i,j}=1, G_{i,j}=1 \end{cases} \tag{1}$$

The error model for ternary data is described in detail in the "Methods" section. The observed genotype matrix can also have missing data because of uneven coverage of single-cell sequencing. SiFit handles missing data by marginalizing over possible genotypes (see "Methods" section for details).

Tree likelihood

A phylogenetic tree, $\mathscr{T} = (T, \mathbf{t})$ consists of a tree topology T and a vector of the branch lengths $\mathbf{t}$. Assuming the technical errors to be independent of each other, and sites to evolve independently, the likelihood of a phylogenetic tree $\mathscr{T}$, and the error rates $\boldsymbol{\theta} = (\alpha, \beta)$ is given by

$$\mathscr{L}(\mathscr{T}, \boldsymbol{\theta}) = Pr(D|\mathscr{T}, \boldsymbol{\theta}) = \prod_{i=1}^{n} Pr(D_i|\mathscr{T}, \boldsymbol{\theta}), \tag{2}$$

where D_i is the observed data at site i and it is a vector with m values corresponding to m single cells. The likelihood calculation for a particular site is described in detail in the "Methods" section. The maximum likelihood estimate is obtained by

$$(\mathscr{T}, \boldsymbol{\theta})_{ML} = \underset{(\mathscr{T}, \boldsymbol{\theta})}{\arg\max}\, Pr(D|\mathscr{T}, \boldsymbol{\theta}) \tag{3}$$

Heuristic search algorithm

Our model has two main components, the phylogenetic tree $\mathscr{T}$ and the error rates of single-cell data $\boldsymbol{\theta}$. The tree search space has $\frac{(2m-3)!}{2^{m-1}(m-1)!}$ discrete bifurcating tree topologies for m cells, and each topology has a continuous component for branch lengths. The overall search space also has a continuous component for error rates along with the tree space. We designed a heuristic search algorithm to explore the joint search space to infer the maximum likelihood configuration of phylogeny and error rates. In the joint $(\mathscr{T}, \boldsymbol{\theta})$ space, we consider two types of moves to propose a new configuration. In each type of move, one component is changed. Thus from a current configuration $(\mathscr{T}, \boldsymbol{\theta})$, a new configuration of either $(\mathscr{T}', \boldsymbol{\theta})$ or $(\mathscr{T}, \boldsymbol{\theta}')$ is proposed. The new configuration is heuristically accepted according to a ratio of likelihood and proposal. The search procedure terminates when the likelihood does not improve or the maximum number of iterations has been reached.

Performance on simulated data

First, we evaluated the performance of SiFit on extensive simulated datasets. The simulation studies were aimed at analyzing SiFit's accuracy of phylogeny inference under different experimental conditions. We also assessed SiFit's ability to estimate the error rates and its robustness against increased error rates. We compared SiFit's performance to three other methods. To analyze how tree inference process degrades if the inference algorithm fails to account for the SCS errors, we chose a representative of classic phylogeny inference method as used by Eirew *et al.* [30]. Eirew *et al.* used MRBAYES [31], a Bayesian phylogenetic inference method, that reports a set of trees drawn from the posterior distribution. Even though it was applied on SCS data, this method does not account for the errors in SCS data. The trees inferred from this method can be directly compared against the true trees. We also compared against SCITE [34] and OncoNEM [33], methods that infer tumor trees under "infinite sites assumption". SCITE was designed to infer a mutation tree, but it can also infer a binary leaf-labelled tree, where the cells are the leaf labels and edges contain mutations. We used SCITE to infer the binary leaf-labelled tree from simulated datasets so that they can be directly compared against the true trees. OncoNEM infers a clonal tree which can not be directly compared against the simulated trees. OncoNEM first infers a cell lineage tree and then converts it to a clonal tree by clustering nodes. The cell lineage tree inferred by OncoNEM is a different representation of the clonal tree . We convert the cell lineage tree inferred from OncoNEM to an equivalent phylogenetic tree (potentially non-binary) by projecting the internal nodes to leaves (for details see "Methods") enabling us to compare OncoNEM results against true trees.

As the performance metric, we use the tree reconstruction error, which measures the distance of the inferred tree from the true tree. The distance between two binary trees is measured in terms of Robinson-Foulds (RF) distance [42], which counts the number of non-trivial bipartitions that are present in the inferred or the true tree but not in both the trees. We normalize this count using the total number of bipartitions in the two trees. The output of SiFit, SCITE and Bayesian phylogenetic inference algorithm (MRBAYES) is compared against the true tree in terms of the RF distance. The tree inferred by OncoNEM might be non-binary, so for OncoNEM trees, we separately computed FP and FN distances between the true tree and the inferred tree. For binary trees with the same leaf set, the FP and FN distances are equal. For non-binary tree, FP and FN distances could differ from each other. The "Methods" section gives the details of the tree reconstruction error metric for comparing trees.

Accuracy of phylogeny inference

To analyze the accuracy of SiFit's tree inference, we simulated random binary phylogenetic trees for varying number of leaves (single cells). The number of cells, i.e, leaves in the trees, m, was varied as $m = 20$, $m = 40$ and $m = 60$. The number of sites, n, was varied as $n = 100$, $n = 250$ and $n = 500$ respectively. For each combination of n and m, we generated 20 datasets that were simulated from 20 random trees. The simulation for a single dataset was performed as follows. First, a random binary tree is constructed on a leaf set of single cells by a recursive algorithm that

randomly divides the set of cells into two subtrees that are also randomly generated, and then joins them into a single tree by choosing a root that has the two subtrees as the left and right child. The root of the tree has homozygous reference genotype at all sites. The genotype sequence at the root is evolved along the branches of the tree following our finite-site model of evolution. After evolving, the leaves have genotype sequences with true mutations. m genotype sequences corresponding to m single cells constitute the true genotype matrix. Errors are introduced into the true genotype matrix to simulate single-cell errors. The false negative rate for cell c, β_c, is sampled from a normal distribution with mean $\beta_{mean} = 0.2$ and standard deviation $\beta_{sd} = \frac{\beta_{mean}}{10}$. False negatives are introduced in the true genotype matrix with probability β_c for cell c. We introduced false positives to the true genotype matrix with error rate, $\alpha = 2 \times 10^{-3}$, by converting homozygous reference genotypes to heterozygous genotypes with probability α. We used higher false positive rate than reported in previous studies [22, 23] to ensure that false positives are inserted even in datasets with smaller number of sites. After adding noise, the imperfect genotype matrices were used as input to SiFit for learning maximum likelihood tree.

SiFit's tree inference accuracy was compared against three other methods. Same imperfect genotype matrix was used as input to SiFit and SCITE. For OncoNEM and MRBAYES, the genotype matrices were binarized by converting the heterozygous and homozygous non-reference genotypes to 1, i.e., presence of mutation. The comparison is shown in Fig. 1, which shows the tree reconstruction error. For each value of n, the mean error metric over 20 datasets is plotted along with the standard deviation as the error bar. SiFit substantially outperforms the other three methods for all values of m and n. The performance of each algorithm except for OncoNEM improves as the value of n increases. The behavior of OncoNEM is different. For $m = 20$, its accuracy decreases for $n = 250$ compared to $n = 100$ and $n = 500$. Also for $m = 40$, OncoNEM's accuracy slightly decreases when the number of sites n is varied from $n = 250$ to $n = 500$. This might be because, OncoNEM was developed for clonal tree inference and the effect of an additional number of sites cannot be observed in the equivalent phylogenetic tree unless they (the additional sites) are different across the clones. For $n = 250$ and $n = 500$ datasets, SiFit could find the true tree topology for most of the datasets demonstrating its ability to infer correct trees given enough data.

We also tested how SiFit's performance is affected if SCS errors are not accounted for via SiFit's error model. For doing so, we compared the results for SiFit under two experimental conditions (Fig. 2). In the first case, SiFit used both the error model and the model of evolution during inference, while in the second case, SiFit did not employ the error model and inferred based solely on the finite-site model of evolution. As evident from Fig. 2, SiFit achieves higher inference accuracy when it employs the error model along with the model of evolution compared to the case when it excludes the error model. The difference between the two is smaller for datasets with a smaller number of cells ($m = 20$), but the error model plays a substantial role when the datasets get larger ($m = 40$ and $m = 60$). SCS FP errors being non-recurrent, have more pronounced effect as the number of cells increases. With an increase in the number of cells, the accurate reconstruction of phylogeny becomes more difficult for a method that does not account for SCS errors. This

experiment shows that both the model of evolution and the SCS error model are fundamental components of SiFit and both of them play a significant role in tree inference.

Inference with missing data

Due to uneven coverage and amplification bias, current single-cell sequencing datasets are challenged by missing data points where genotype states are unobserved. To investigate how missing data affect phylogeny reconstruction, we performed additional simulation experiments. For $m = 20$ and $n = \{100, 250, 500\}$, we generated datasets using the same error rates as before. For each combination of n and m, we generated 20 datasets, for each of which, three other datasets with missing data $= \{10\%, 25\%, 50\%\}$ were generated. To generate the datasets with missing data, genotype information of sites were removed with probability $0.1, 0.25, 0.5$ for missing data $= \{10\%, 25\%, 50\%\}$ respectively. SiFit's results were compared against SCITE and OncoNEM, the results are shown in Fig. 3. As the missing data rate increases from 0 to 25%, we observe a slight increase in tree reconstruction error as compared to the datasets without missing data. As missing data rate becomes 50%, tree reconstruction error increases slightly more, even though for $n = \{250, 500\}$ the increase in tree reconstruction error remains consistent. This is expected as 50% missing data results in removing half of the data points. In each case, SiFit performs substantially better than SCITE and OncoNEM. SiFit's likelihood calculation treats each missing data as contributing a marginal probability of 1, effectively making it equivalent to reducing the number of sites n. Even for very high rates (50%) of missing data, SiFit's performance is very good, especially for datasets with 250 and 500 sites.

Robustness to increasing error rates

Allelic dropout is the major source of error in single-cell sequencing data resulting in false negatives [14]. To test the robustness of SiFit to increase in false negative rate, β, we simulated datasets with increased false negative rates. The number of cells, m was set to 20 and the number of sites, n, was set to 250. Mean false negative rate, β_{mean}, was varied from 0.1 to 0.4 in steps of 0.1 i.e, $\beta_{mean} \in \{0.1, 0.2, 0.3, 0.4\}$. The false negative rate of cell c, β_c was sampled from a normal distribution as described in the previous experiment. The false positive rate was set to $\alpha = 2 \times 10^{-3}$. With these settings, for each value of $\beta_{mean} \in \{0.1, 0.2, 0.3, 0.4\}$, 20 datasets were simulated for phylogeny reconstruction. For this and subsequent experiments, the performance of SiFit was compared against that of only SCITE as previous experiments showed that SCITE performed the best among three competitor algorithms. With the increase in the false negative rate, the tree inference error increases slightly. For different settings of false negative rates, SiFit performs better than SCITE in reducing the tree reconstruction error (Fig. 4). The rate of increase in tree reconstruction error for SiFit is also much lower as compared to that of SCITE. This suggests that SiFit is more robust against technical errors as compared to SCITE. As the false negative rate increases, the standard deviation of inference error also increases. This might be because for higher error rate, the chance of another tree with different topology fitting the data increases.

Estimation of error rates

In addition to the phylogenetic tree, SiFit also learns the error parameters from the data. To examine SiFit's capability to estimate the false negative rate from data, we simulated 50 datasets from 50 random binary trees. For these datasets, the number of cells was set to 20 and the number of sites was set to 250 and the false positive rate was set to $\alpha = 0.002$. The false negative rate, β was varied from 0.1 to 0.4. These imperfect data matrices were given to SiFit for inference of tree and false negative rate.

SiFit performed very well for estimating false negative rate as shown in Fig. 5. The maximum likelihood value of β learned from the data were highly correlated (0.9689) to the ones that generated the data. This experiment demonstrates SiFit's ability to infer error parameters from data.

SCITE can also learn false negative rate from data. Since, it assumes infinite sites model, any deviation from that model should be treated as an error by SCITE. To examine this, we used SCITE for learning the false negative rates from the same datasets. As expected, SCITE overestimated the false negative rates for most these datasets (Fig. 5) because any site that violates the infinite sites assumption is treated as having an error by SCITE. The correlation of 0.5070 between the SCITE's estimates and original values of false negative rate was much lower than that of SiFit.

Inference of tumor phylogeny from real tumor SCS data

We applied SiFit to three experimental single-cell DNA exome sequencing datasets: a JAK2-negative myeloproliferative neoplasm, a muscle-invasive bladder cancer and an estrogen receptor positive breast cancer patients. From these data we inferred the phylogenetic lineages of the tumor and ordered the chronology of mutations. These studies used different single-cell DNA sequencing methods and had different samples sizes and error rates, which we selected to show that SiFit can be applied broadly to different single-cell exome datasets.

Phylogenetic lineage of a myeloproliferative neoplasm

SiFit was applied to single-cell exome sequencing data from a JAK2-negative myeloproliferative neoplasm [22] patient. In this dataset, 58 tumor cells were sequenced, which resulted in the detection of 712 somatic SNVs. The average ADO rate was originally estimated to be $\beta = 0.4309$ and the false positive rate was estimated to be $\alpha = 6.04 \times 10^{-5}$ [22]. Using SiFit, we estimated $\beta = 0.301176$, which is slightly lower than the value reported in the original study. In total, approximately 58% of the values were missing in the dataset. The reported genotypes were binary values, representing the presence or absence of a mutation at the SNV sites and were obtained from another published study [32].

To test whether the genotype matrix violates the "infinite sites assumption", we ran the four-gamete test. The four-gametes theorem states that an $m \times n$ binary matrix, M, has an undirected perfect phylogeny if and only if no pair of columns contain all four binary pairs $(0,0; 0,1; 1,0$ and $1,1)$, where m represents the number of taxa (leaves of the tree) and n represents genomic sites [43]. The perfect phylogeny model conveys the biological feature that every genomic site mutates at most once

in the phylogeny [43] and that mutations are never lost. The existence of perfect phylogeny shows that the data could fit the infinite sites model of evolution. The binary mutation matrix from JAK2-negative myeloproliferative neoplasm violated the four-gamete test, with 307 pairs of SNV sites that contained all four binary pairs.

The maximum likelihood tree inferred by SiFit on 712 SNVs is shown in Fig. 6 and has a log-likelihood of -12556.469516. The tree shows that the distance of the normal bulk tissue (LN.T1) is the shortest from the root. The tree is linear near the root, and branching is observed in the later stages of tumor progression, which resulted in the divergence of two major subpopulations (B and C). We performed k-medoids clustering using silhouette score (see "Method" for details) on the ML tree-based distance matrix, which identified three tumor subpopulations (A, B, C). More than 70% of the mutations occur at the trunk of the tree suggesting that they occurred at the earliest stages of tumor progression. In the original study, the authors identified 8 key cancer genes that were predicted to have a functional impact and an important role in tumor progression: *SESN2, ST13, NTRK1, ABCB5, FRG1, ASNS, TOP1MT* and *DNAJC17* (Hou *et al.* [22], Table 3). However the authors could not resolve the order of these mutations during tumor progression or the clonal subpopulations in which they occurred. The SiFit tumor lineage shows that *SESN2, NTRK1, DNAJC17* and *TOP1MT* were early mutations that occurred in the trunk of the lineage. In the later stages of the tumor lineage, driver mutations were acquired in *ABCB5* and *PDE4DIP*, which led to a major expansion of a new subpopulation (B). Additional cancer gene mutations arose in *ST13* and *FRG1* within the subbranches of subpopulation B as this subpopulation continued to expand during tumor evolution. The same *FRG1* mutation occurred in two different branches within subpopulation B, indicating possibility of convergent evolution. This result clearly differentiates our method from others and is a consequence of the finite-site model uniquely implemented in SiFit. From population B, subpopulation C diverged later in the evolution of the tumor and acquired a number of new mutations, including a cancer gene mutation in *ASNS*. In addition to the driver mutations, this tree also estimated the timing of many additional mutations that occurred during the evolution of the tumor (Fig. 6). We also applied SCITE to this dataset, which infers a linear mutation tree with two major subpopulations diverging (14 and 12 cells) (Additional file 1: Fig. S1).

Phylogenetic lineage of a muscle-invasive bladder cancer

We also applied SiFit to single-cell exome sequencing data from a muscle-invasive bladder transitional cell carcinoma [26]. The dataset consisted of 44 single tumor cells, 12 single normal cells, in addition to bulk exome sequencing data from normal and tumor tissue. In the original study, Li *et al.* [26] detected 443 somatic SNVs across the cells using a consensus-based filtering method. The average ADO rate was estimated to be 0.4 and the false positive rate was estimated to be 6.7×10^{-5}. SiFit estimated $\beta = 0.535172$ which is slightly higher than the false negative rate reported in the original study. 55.2% entries were missing in the final genotype matrix. The genotypes represented the presence or absence of a mutation at the site.

We ran the four-gamete test on this dataset, which identified 123 pairs of SNV sites violating the "infinite sites assumption" as a perfect phylogeny can not exist because of the violation of four-gamete test.

SiFit was run on this dataset and the maximum likelihood tree was constructed (Fig. 7). The tree shows a long linear trunk of tumor cells (A) that emerged from the normal cells, that eventually bifurcated into two subtrees (B and C) in the later stages of tumor evolution. K-medoids clustering analysis on the ML tree-based distance matrix identified three tumor subpopulations (A, B and C). These results are consistent with the original study that reported three major subpopulations in addition to the normal cells [26]. In the original study, four key genes were identified as driver mutations: *CFTR, NIPBL, ASTN1* and *DHX57*, but their chronology or population substructure was not delineated. The SiFit lineage showed that *CFTR, NIPBL, ASTN1* and *DHX57*, all occurred at the earliest stages of tumor evolution, in the base on the tumor trunk, before the first tumor cell was sampled (BC-58). In addition to these key driver mutations, our annotations using TCGA and COSMIC also identified *PDE4DIP, ATM* and *BMPR1A* as potential driver mutations that occurred at the earliest stages of tumor evolution and were located in the base of the evolutionary trunk. The SiFit tree also revealed the mutation occurrence and order for other nonsynonymous mutations, including 72 in clone A (red), 11 in clone B (blue) and 23 in clone C (green). In the later stages of tumor evolution, the tree bifurcated into two major subpopulations (B and C), after having acquired 17 nonsynonymous mutations. However, the role of these mutations in tumor progression remains unclear. We also ran SCITE on this dataset which inferred two trees with a linear series of tumor cells that diverged from a normal subpopulation (Additional file 1: Fig. S2, Fig. S3). Both tree structures contained a long trunkal branch, but differed in the placement of a few of the normal single cells. The SCITE tree did not resolve the bifurcation of the two major tumor lineages (B and C) or the three major subpopulations that were identified by SiFit.

Phylogenetic lineage of an ER positive breast cancer

We further selected an invasive ductal carcinoma from an oestrogen-receptor positive (ER^+) breast cancer patient for phylogenetic analysis using SiFit [24]. This dataset consisted of single-cell exome sequencing data from 47 tumor cells and 12 normal cells. We focused our analysis on 40 nonsynonymous mutation sites that were reported in the original study and were represented as a binary genotype matrix. In the original study, the estimated false positive rate was 1.24×10^{-6} and the estimated allelic drop out was 9.72%. SiFit estimated the false negative rate to be $\beta = 0.139126$, which was very close to the value reported in the original study.

The four-gamete test failed indicating the violation of "infinite sites assumption" for 21 pairs of SNV sites. The maximum likelihood phylogenetic tree constructed from this dataset (Fig. 8) shows a linear evolution of the cells at earlier stages of the tumor in a very narrow trunk, followed by a highly branched tree structure that resulted in multiple subpopulations. The tumor cells emerged from the normal breast cells after acquiring driver mutations in *PIK3CA* and *CASP3* which lead to an expansion of the initial subpopulation via a linear trajectory. By performing k-medoids clustering on the ML tree-based distance matrix, we identified five tumor

subpopulations (A, B, C, D and E). Important driver mutations in *FBN2* and *FGFR2* emerged in clone A. However, the *FGFR2* clones did not undergo further expansion and were detected in only three cells, while the *FBN2* clones continued to expand and diverged to form additional subpopulations (B, C, D and E). The long branch lengths closer to the root of the tree and the placement of fewer tumor cells in the linear branches suggests a highly branched tree structure and ongoing mutational evolution in this tumor. The highly branched tree structure in this ER^+ breast tumor agrees with the findings of Miller *et al.* [44], who investigated the clonal landscape of 22 oestrogen-receptor-positive (ER^+) breast cancer samples via bulk sequencing and reported that more than 80% samples contained multiple (varying from 2 to 5) subclonal cell populations with extensive intratumor heterogeneity, which changed in response to aromatase inhibitor (AI) treatment. The tree inferred by SCITE from this dataset had a linear structure in the trunk and then bifurcated into two subtrees (Additional file 1: Fig. S4).

Conclusions

Tumor phylogenies provide insight into the clonal substructure of tumors and the chronological order of mutations that arose during tumor progression. These lineages have direct applications in clinical oncology, for both diagnostic applications in measuring the amount of intra-tumor heterogeneity in tumors and for improving targeted therapy by helping oncologists identify mutations that are present in the majority of tumor cells. Single-cell DNA sequencing data provides an unprecedented opportunity to reconstruct tumor phylogenies at the highest possible resolution, however are challenged by extensive technical errors that arise during genome amplification. In this paper, we introduced SiFit, a probabilistic method for recreating the evolutionary histories of tumors under finite-site model of evolution from imperfect mutation profiles of single cells. This likelihood-based approach can infer the maximum likelihood phylogeny that best fits the noisy single-cell datasets. SiFit can also estimate the error rates of the single-cell DNA sequencing experiments. SiFit employs a resilient error model that can account for various technical artifacts in single-cell sequencing data, including allelic dropout (ADO), false positives and missing data. Our model is adaptable and can be easily extended to include position-specific error rates. SiFit also provides this flexibility in choosing the model of evolution, for which we extended the Jukes-Cantor model of evolution [39] to accommodate it for tumor phylogeny from single cell data. SiFit is robust to variation in error rates and performs consistently with varying number of cells in the dataset making it widely applicable to SCS datasets that vary in error rates and the number of cells sequenced.

The main difference of SiFit from existing methods, specifically SCITE [34] and OncoNEM [33] is that SiFit introduces a finite-site model of evolution. Both SCITE and OncoNEM makes the "infinite sites assumption" that is frequently violated in cases of convergent evolution or reversal of genotypes, events that occur in human tumors due to LOH and chromosomal deletions [38]. SiFit also makes use of the high-resolution SCS data by utilizing the single cells as the taxonomic units of the reconstructed phylogenetic tree. On the other hand, SCITE reports a mutation tree, in which the lineage of the cells are not shown. OncoNEM reports a clonal

tree, which is a condensed tree with multiple cells clustered into a clone. This type of clonal clustering and the use of clones as the taxonomic units, though useful for finding genealogical relationships between clones, is low-resolution as a clone represents a consensus of information from multiple single cells. The utilization of mutation information from each individual cell makes SiFit's tree reconstruction method both robust and high-resolution.

SiFit performs accurately as evident from a comprehensive set of simulation studies that takes into account different aspects of modern SCS datasets by experimenting with varying number of cells in the dataset, wide range of error rates and different fractions of missing data. The simulation studies also demonstrated that SiFit substantially outperformed the state-of-the-art methods and is more robust to technical errors from WGA. We also applied SiFit to reconstruct the phylogeny for three real SCS tumor datasets. SiFit accurately reconstructed the phylogenetic lineages of these tumors, and identified points in which subpopulations diverged from the main tumor lineages. These trees also provided insight into the order of mutations and the chronology in which they arose during tumor progression, which were not inferred in the original studies.

SiFit's phylogeny inference can potentially be improved by incorporating copy number variations along with single nucleotide variants. Recent studies [45] indicate that copy number appears to follow punctuated evolutionary model and are likely to provide insight into possible loss of heterozygosity (LOH) events and can facilitate in tree inference. Such an approach has previously been used in the context of bulk sequencing data [16] and can be incorporated for SCS data under a finite-site model of evolution. SiFit currently uses fixed error rates at every site. The error model can be further extended using position-specific error rates, where sites with lower-confidence mutations will have higher error rates and vice versa. The error model will have higher complexity in that situation and systematic model selection has to be performed.

As single-cell DNA sequencing becomes more high-throughput [46, 47] enabling hundreds of cells to be analyzed in parallel at reduced cost and throughput, SiFit is well positioned to analyze the resulting large-scale datasets to understand the evolution of clones during tumor progression. SiFit adds a major step forward in understanding the tumor phylogeny from SCS data and will have important translational applications for improving cancer diagnosis, treatment and personalized therapy. Although the current study is focused on cancer, SiFit can potentially also be applied to single-cell mutation profiles from a wide variety of fields including immunology, neurobiology, microbiology and tissue mosaicism. These applications are expected to provide new insights into our understanding of cancer and other human diseases.

Methods

Input data

The input to SiFit is a matrix $D_{n\times m} = (D_{ij})$ of observed genotypes, where $i \in \{1, ..., n\}$ denotes the index of genomic locus, $j \in \{1, ..., m\}$ is the index of the single cell and D_{ij} is the observed genotype at the i^{th} site of cell j. The genotype matrix can be binary or ternary depending on the representation of the data. For binary

matrix, $D_{ij} \in \{0, 1, X\}$, where 0, 1 and X denote the absence of mutation, presence of mutation and missing data respectively. For a ternary matrix, D_{ij} can take value from the set, $\{0, 1, 2, X\}$, where 0 denotes homozygous reference genotype, 1 and 2 denote heterozygous and homozygous non-reference genotypes respectively and X denotes missing data.

Model of single-cell errors

False positive errors and false negative errors are the two different types of noises that could be present in the genotype matrix. If α is the false positive error rate and β is the false negative error rate, then for a ternary genotype matrix, the relationship between the true and observed genotype matrices is given by:

$$Pr(D_{i,j}|G_{i,j}) = \begin{cases} 1-\alpha-\frac{\alpha\beta}{2} & \text{if } D_{i,j}=0, G_{i,j}=0 \\ \alpha & \text{if } D_{i,j}=1, G_{i,j}=0 \\ \frac{\alpha\beta}{2} & \text{if } D_{i,j}=2, G_{i,j}=0 \\ \frac{\beta}{2} & \text{if } D_{i,j}=0, G_{i,j}=1 \\ 1-\beta & \text{if } D_{i,j}=1, G_{i,j}=1 \\ \frac{\beta}{2} & \text{if } D_{i,j}=2, G_{i,j}=1 \\ 0 & \text{if } D_{i,j}=0, G_{i,j}=2 \\ 0 & \text{if } D_{i,j}=1, G_{i,j}=2 \\ 1 & \text{if } D_{i,j}=2, G_{i,j}=2 \end{cases} \quad (4)$$

where $G_{i,j}$ is the unobserved true genotype at the i^{th} site of cell j. A true homozygous non-reference genotype (site with true homozygous mutation) is affected by neither false positive error nor allelic dropout. A false negative error can affect the heterozygous genotype and combined with false positive error can also affect homozygous reference genotype. False positive error can affect homozygous reference genotypes.

Single-cell datasets also contain missing data, sites for which genotype information is missing. In our computation, we take $Pr(D_{i,j}|G_{i,j}) = 1$ whenever $D_{i,j} = X$. By doing so, we marginalize the effect of missing data over three possible true genotypes and is reflected in the likelihood computation.

Likelihood of a phylogenetic tree

Phylogenetic tree

We consider that the phylogenetic tree for single cells is a rooted directed binary tree $\mathscr{T} = (T, \mathbf{t})$. It has two components, a tree topology T and a vector of branch lengths $\mathbf{t}$. The phylogenetic tree represents the genealogical relationship among a set of single cells. The root of this tree has homozygous reference genotypes at all sites. The leaves of the tree represent the observed single cells. The internal nodes represent ancestral cells that are not observed in the data. Cells evolve along the branches of the tree following a model of evolution and the branch length denotes expected number of mutations per site.

Model of evolution

We assume that the sites evolve identically and independently. As the model of evolution in single cells, we extend the single parameter Jukes Cantor Model [39], a finite-site Markov model, to accommodate for two copies of DNA present in a diploid cell. Given three possible genotypes $\{0, 1, 2\}$, the transition matrix P_t is given by,

$$\begin{aligned} P_t(0,0) = P_t(2,2) &= (\frac{1}{4} + \frac{3}{4}e^{-\mu t})^2 \\ P_t(0,1) = P_t(2,1) &= (\frac{1}{4} + \frac{3}{4}e^{-\mu t})(\frac{1}{4} - \frac{1}{4}e^{-\mu t}) \\ P_t(0,2) = P_t(2,0) &= (\frac{1}{4} - \frac{1}{4}e^{-\mu t})^2 \\ P_t(1,0) = P_t(1,2) &= \frac{1}{2}(\frac{1}{4} + \frac{3}{4}e^{-\mu t})(\frac{1}{4} - \frac{1}{4}e^{-\mu t}) \\ P_t(1,1) &= 1 - (\frac{1}{4} + \frac{3}{4}e^{-\mu t})(\frac{1}{4} - \frac{1}{4}e^{-\mu t}) \end{aligned} \tag{5}$$

where $P_t(i, j)$ represents the probability that genotype i will mutate to genotype j over a branch of length t, μ is the parameter of the model. For a binary genotype matrix, the transition matrix is given by,

$$\begin{aligned} P_t(0,0) &= (\frac{1}{4} + \frac{3}{4}e^{-\mu t})^2 \\ P_t(0,1) = P_t(1,0) &= (\frac{1}{4} + \frac{3}{4}e^{-\mu t})(\frac{1}{4} - \frac{1}{4}e^{-\mu t}) \\ P_t(1,1) &= 1 - (\frac{1}{4} + \frac{3}{4}e^{-\mu t})(\frac{1}{4} - \frac{1}{4}e^{-\mu t}) \end{aligned} \tag{6}$$

In Eq. (5) and Eq. (6), we assume that each copy of a chromosome evolves independently.

Likelihood

Since we assume that each site evolves independently and the technical errors affect each site independently, the likelihood for the observed genotype matrix given a phylogenetic tree $\mathscr{T}$ and error rates, $\boldsymbol{\theta}$ is given by

$$\mathscr{L}(\mathscr{T}, \boldsymbol{\theta}) = Pr(D|\mathscr{T}, \boldsymbol{\theta}) = \prod_{i=1}^{n} Pr(D_i|\mathscr{T}, \boldsymbol{\theta}) \tag{7}$$

where D_i is the observed data at site i and it is a vector with m values corresponding to m single cells. Let γ be the set of possible genotypes. If v be an internal node of the tree with children u, w, and let $L_i^v(g), g \in \gamma$ denote the partial conditional likelihood defined by

$$L_i^v(g) = Pr(D_i^v|\mathscr{T}, \boldsymbol{\theta}, \hat{D}_i(v) = g) \tag{8}$$

where D_i^v is the restriction of data D_i to the descendants of node v and $\hat{D}_i(v)$ is the ancestral genotype for i^{th} site at node v. $L_i^v(g)$ is the likelihood at site i for the subtree rooted at node v, given that the genotype at v is g.

The likelihood of the complete observed data D_i at the i^{th} site is given by:

$$Pr(D_i|\mathscr{T},\boldsymbol{\theta}) = L_i^r(0) \tag{9}$$

where r is the root of the tree and since we consider that the genotypes at the root are all homozygous reference (0), the probability $Pr(\hat{D}_i(r) = 0)$ equals 1. The partial conditional likelihood function satisfies the recursive relation

$$L_i^v(g) = \left[\sum_{h\in\gamma} P_{t_{vu}}(g,h)L_i^u(h)\right]\left[\sum_{h\in\gamma} P_{t_{vw}}(g,h)L_i^w(h)\right] \tag{10}$$

for all internal nodes v with children u and w. t_{vu} and t_{vw} are the branch lengths corresponding to branches that connect v to u and w respectively. $P_{t_{vu}}(g,h)$ and $P_{t_{vw}}(g,h)$ are transition probabilities that are calculated using either Eq. (5) or Eq. (6) with argument t_{vu} and t_{vw} respectively. For a leaf of the tree that denotes single cell j, the partial likelihood is given by

$$L_i^j(g) = Pr(D_{i,j}|G_{i,j} = g)$$

where $Pr(D_{i,j}|G_{i,j})$ is calculated using either Eq. (2) or Eq. (4) depending on the data. The partial likelihood values at the leaves are computed based on the error rates of SCS data.

The log-likelihood for the observed genotype matrix given a phylogenetic tree $\mathscr{T}$ and error rates, $\boldsymbol{\theta}$ becomes a summation over n sites as in Eq. (11)

$$\log \mathscr{L}(\mathscr{T},\boldsymbol{\theta}) = \sum_{i=1}^{n} \log L_i^r(0) \tag{11}$$

This likelihood computation uses Felsenstein's pruning algorithm [48] for calculating the likelihood of a phylogenetic tree with the transition probabilities given by Eq. (5) or Eq. (6). For the calculation of the partial likelihoods for leaves, we use the SCS error model instead of values suggested in [48].

Search algorithm to infer phylogeny

We developed a heuristic search algorithm to stochastically explore the joint space of phylogenetic trees and error rates. In the joint $(\mathscr{T},\boldsymbol{\theta})$ space, we need to consider two different types of moves to propose a new configuration. In tree changing moves, a new phylogenetic tree, $\mathscr{T}'$ is proposed from current state $\mathscr{T}$. In error rate changing moves, a new error rate, $\boldsymbol{\theta'}$ is proposed from current error rate $\boldsymbol{\theta}$. The proposed configuration is accepted or rejected based on an acceptance ratio. The acceptance ratio for proposing a new phylogenetic tree is given by,

$$\rho_T = min\left\{\frac{Pr(D|\mathscr{T}',\boldsymbol{\theta})q_T(\mathscr{T}|\mathscr{T}')}{Pr(D|\mathscr{T},\boldsymbol{\theta})q_T(\mathscr{T}'|\mathscr{T})}, 1\right\} \tag{12}$$

which involves calculating the ratio of the likelihood of new configuration and current configuration. Acceptance ratio also requires a proposal ratio which is computed

based on q_T, the proposal distribution for proposing new tree. A new error rate $\boldsymbol{\theta'}$ is accepted with ratio given by,

$$\rho_\theta = min\left\{\frac{Pr(D|\mathscr{T},\boldsymbol{\theta'})p_\theta(\boldsymbol{\theta'})q_\theta(\boldsymbol{\theta}|\boldsymbol{\theta'})}{Pr(D|\mathscr{T},\boldsymbol{\theta})p_\theta(\boldsymbol{\theta})q_\theta(\boldsymbol{\theta'}|\boldsymbol{\theta})},1\right\} \tag{13}$$

which takes into account the ratio of likelihoods of new and current configurations, the ratio of prior probability of new and current error rates and also a proposal ratio. p_θ is the prior distribution on error rate and q_θ is the proposal distribution for proposing new error rate. Even though the details of this algorithm is motivated by Metropolis-Hastings algorithm [49] for doing Markov Chain Monte Carlo (MCMC) sampling, we search for the maximum likelihood configuration. The inference algorithm is shown in Algorithm 1

Algorithm 1 Algorithm for phylogeny and error rate inference. D is the observed genotype matrix, θ_p is the starting value of error rates. The algorithm runs for n_{iter} iterations. With probability π, error rate changing moves are proposed.

function PhyloTreeSearch(D, θ_p, n_{iter}, π, m)
  Initialize:
  $\mathscr{T}^0$ to a random tree with m leaves
  $\boldsymbol{\theta}^0$ to θ_p
  $\mathscr{L}^0 \leftarrow$ Likelihood of $(\mathscr{T}^0, \boldsymbol{\theta}^0)$
  $\mathscr{L}^{best} \leftarrow \mathscr{L}^0$, $\mathscr{T}^{best} \leftarrow \mathscr{T}^0$, $\boldsymbol{\theta}^{best} \leftarrow \boldsymbol{\theta}^0$
  for $i = 1...n_{iter}$ **do**
    Define $\mathscr{T} \leftarrow \mathscr{T}^{i-1}$, $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}^{i-1}$
    Sample $r \sim U(0,1)$
    if $r \leq \pi$ **then**
      Sample $\boldsymbol{\theta'} \sim q_\theta(\boldsymbol{\theta'}|\boldsymbol{\theta})$
      $\rho_\theta \leftarrow min\left\{\frac{Pr(D|\mathscr{T},\boldsymbol{\theta'})p_\theta(\boldsymbol{\theta'})q_\theta(\boldsymbol{\theta}|\boldsymbol{\theta'})}{Pr(D|\mathscr{T},\boldsymbol{\theta})p_\theta(\boldsymbol{\theta})q_\theta(\boldsymbol{\theta'}|\boldsymbol{\theta})},1\right\}$
      accept $\boldsymbol{\theta'}$ with probability ρ_θ
      $\boldsymbol{\theta}^i \leftarrow \boldsymbol{\theta'}$, $\mathscr{T}^i \leftarrow \mathscr{T}$
      $\mathscr{L}^i \leftarrow$ Likelihood of $(\mathscr{T}, \boldsymbol{\theta'})$
    else
      Sample $\mathscr{T}' \sim q_T(\mathscr{T}'|\mathscr{T})$
      $\rho_T \leftarrow min\left\{\frac{Pr(D|\mathscr{T}',\boldsymbol{\theta})q_T(\mathscr{T}|\mathscr{T}')}{Pr(D|\mathscr{T},\boldsymbol{\theta})q_T(\mathscr{T}'|\mathscr{T})},1\right\}$
      accept $\mathscr{T}'$ with probability ρ_T
      $\boldsymbol{\theta}^i \leftarrow \boldsymbol{\theta}$, $\mathscr{T}^i \leftarrow \mathscr{T}'$
      $\mathscr{L}^i \leftarrow$ Likelihood of $(\mathscr{T}', \boldsymbol{\theta})$
    end if
    if $\mathscr{L}^i > \mathscr{L}^{best}$ **then**
      $\mathscr{L}^{best} \leftarrow \mathscr{L}^i$, $\mathscr{T}^{best} \leftarrow \mathscr{T}^i$, $\boldsymbol{\theta}^{best} \leftarrow \boldsymbol{\theta}^i$
    end if
  end for
  return $(\mathscr{L}^{best}, \mathscr{T}^{best}, \boldsymbol{\theta}^{best})$
end function

Tree proposals

To explore the space of trees we need efficient moves that can make small and big changes in the tree topology. Also, we need moves that change only the branch lengths instead of changing the topology. To ensure that our search does not get stuck to a local optimum, we use a combination of different types of moves. Lakner et al. [50] described several tree proposal mechanisms that are effective in Bayesian phylogenetic inference. Since our goal is to effectively search the tree space, we can employ the same tree proposals in our search algorithm. We adopt two different types of tree proposals described in [50] in our search process, branch change

proposals that alter branch lengths and branch-rearrangement proposals that alter the tree topology. The branch-rearrangement proposals can be divided into two subtypes: the prune and reattach moves and the swapping moves.

For proposing new branch length, we draw a sample u from a uniform distribution on $[0, 1)$ and then get a random number r^* by applying the transformation $r^* = e^{\eta(u-0.5)}$. The new branch length l^* is a product of current branch length l and r^*. In this way, we update branch length of all branches. This ensures that branch lengths are locally changed the proposal ratio becomes a product $\prod_k r_k^*$, where k is the total number of branches in the tree. η is a tuning parameter that is set to the value suggested in [50].

We consider two types of pruning-regrafting moves, namely Random Subtree Pruning and Regrafting (rSPR) and Extending Subtree Pruning and Regrafting (eSPR), which were described in [50]. The pruning-regrafting moves randomly select an interior branch, prune a subtree attached to that branch, and then reattach the subtree to another regrafting branch present in the other subtree. For rSPR, the regrafting branch is chosen randomly. For eSPR, an extension probability guides the movement of the point of regrafting across one branch at a time. The eSPR move favors local rearrangements.

We consider three types of swapping moves, namely Stochastic Nearest Neighbor Interchange (stNNI), Random Subtree Swapping (rSTS) and Extending Subtree Swapping (eSTS). stNNI chooses an internal branch as the focal branch and stochastically swaps the subtrees attached to the focal branch. eSTS also involves the swap of two subtrees but not necessarily nearest neighbors. The subtrees are chosen according to an extension mechanism similar to eSPR. For rSTS, two randomly chosen subtrees are swapped.

At each step of the search algorithm, one of these six moves is chosen with a fixed probability. The proposal ratio associated with each branch-rearrangement proposal is described in detail in [50].

Estimation of error rate

During the search process, we also update error rates. The estimates of error rates that are input to SiFit are used to design the prior probability $p(\theta)$. The error rate being a probability (value between 0 and 1), we choose a beta prior. The mean of the prior is estimated from the input error rate and observed genotype matrix. We choose a large standard deviation to cover a wide range of values. We choose a normal distribution as the proposal distribution for proposing new error rate. At each generation, the normal distribution is centered on the current value of error rate. A user specified fixed probability determines whether, in a particular iteration, a new error rate will be proposed.

Tree inference error metric

To measure the accuracy of tree inference, we used a metric that compares the topology of the inferred tree to that of the true tree and computes a distance between the two. This metric on general phylogenetic trees was proposed in [42] and it is based on the symmetric difference between the bipartitions of the two trees. The topology of a tree can be represented by the bipartitions present in the

tree. A bipartition of a tree based on an edge gives us two set of leaves that would be formed by deleting the edge. If $\mathscr{E}$ is the set of edges of $\mathscr{T}$, then the bipartition encoding of $\mathscr{T}$, denoted by $C(\mathscr{T}) = \{\xi(e) : e \in \mathscr{E}\}$, is the set of bipartitions defined by each edge in $\mathscr{T}$. $\xi(e)$ is the bipartition on the leaf set of $\mathscr{T}$ produced by removing the edge e from $\mathscr{T}$. We consider three distances between two trees.

If $\mathscr{T}_t$ is the true tree on a set of single cells $\mathscr{S}$ and $\mathscr{T}_i$ is the inferred tree, then the following are the three inference error metrics.

False Negative (FN) distance, This counts the edges in $\mathscr{T}_t$ that induce bipartitions that are not present in $C(\mathscr{T}_i)$. This distance is normalized by dividing by the total number of bipartitions in $\mathscr{T}_t$, i.e. $\frac{|C(\mathscr{T}_t) \setminus C(\mathscr{T}_i)|}{|C(\mathscr{T}_t)|}$

False Positive (FP) distance, This counts the edges in $\mathscr{T}_i$ that induce bipartitions that are not present in $C(\mathscr{T}_t)$. This distance is normalized by dividing by the total number of bipartitions in $\mathscr{T}_i$, i.e. $\frac{|C(\mathscr{T}_i) \setminus C(\mathscr{T}_t)|}{|C(\mathscr{T}_i)|}$

Robinson-Foulds (RF) distance. The Robinson-Foulds distance is the average of FP and FN distance. This is the most common error metric.

If the two trees to compare are binary then we use RF distance between them as the error metric. For binary trees, FP, FN and RF distances are equal to each other. To compare a true binary tree to an inferred non-binary tree, we compute FP and FN distances separately.

SiFit, SCITE and MRBAYES output binary tree which can be compared against the true tree in terms of RF distance. For OncoNEM, we consider the cell lineage tree that it infers and then we convert the cell lineage tree to an equivalent phylogenetic tree by projecting the observed single cells to leaves (shown in Additional file 1: Fig. S5). The equivalent phylogenetic tree might be binary or non-binary and we compute both FP and FN distances for it when comparing to the true tree.

Inference of ancestral sequences and order of mutations

The inference of the chronological order of mutations in the tumor lineage requires the inference of mutation status of the internal nodes so that the mutations can be placed on the branches of the phylogeny. We infer the mutational profiles of the internal nodes using a likelihood-based approach that finds the most likely mutational profile for an internal node given the phylogenetic tree and error rates. We extend the dynamic programming algorithm for inferring ancestral sequences described in Pupko *et. al* [51] to account for the error rates of the single cells.

For a single cell c at the leaf of the tree, the partial likelihood for a genotype g at site i is calculated as $L_c(g) = \arg\max_h P_{t_{vc}}(g,h) Pr(D_{i,c}|G_{i,c} = h)$ and mutation state, $m_c(g)$, is set to h that attains the maximum value for partial likelihood. v is the parent of c and t_{vc} is the branch length connecting v to c. For a missing data, $Pr(D_{i,c}|G_{i,c} = h)$ becomes 1. For a nonroot internal node, u, with children y and z, the partial likelihood is calculated as $L_u(g) = \arg\max_h P_{t_{wu}}(g,h) L_y(h) L_z(h)$ and the mutation state, $m_u(g)$, is set to h that attains the maximum value. For the root of the tree, mutation state $m_r = 0$ and the mutation state for an internal node, u, whose parent w's mutation state is already determined as g, is chosen as $m_u(g)$.

After inferring the mutational profiles of the internal nodes, the mutations on a branch can be found by finding the SNV sites for which the mutational status of the two nodes at the two ends of the branch differ.

Clustering of cells

To cluster the cells into subpopulations for the tumor datasets, we used k-medoids clustering with silhouette scores. A distance matrix was obtained from the ML tree reconstructed by SiFit, in which, an entry represents the distance between two cells. The distance between two cells was calculated by summing the branch lengths on the path that connects the two cells. K-medoids clustering was performed on the resulting distance matrix using 'clustering' library of R (http://www.r-project.org) and the number of clusters was varied from 2 to 5. In each case, the average silhouette score was measured and the number of clusters that maximized silhouette score was reported as the optimal number of clusters.

Software availability

SiFit has been implemented in Java and is freely available at https://bitbucket.org/hamimzafar/sifit.

Availability of data and materials

The published human tumor datasets that are analyzed are available from the supplementary material of [22, 26] and in Fig. 2f of [24].

Competing interests

The authors declare that they have no competing interests.

Author's contributions

HZ, NN, KC and LN designed the study. HZ and AT developed the algorithm and implemented the software. HZ ran the experiments. All authors wrote and approved the manuscript.

Funding

The study was supported by the National Cancer Institute R01 CA172652 (K.C.), cancer center support grant P30 CA016672 and Andrew Sabin Family Foundation.

Author details

[1]Department of Computer Science, Rice University, Houston, Texas, USA. [2]Department of Bioinformatics and Computational Biology, the University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA. [3]Department of Genetics, the University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA.

References

1. Nowell, P.: The clonal evolution of tumor cell populations. Science **194**(4260), 23–28 (1976)
2. Merlo, L.M.F., Pepper, J.W., Reid, B.J., Maley, C.C.: Cancer as an evolutionary and ecological process. Nat Rev Cancer **6**(12), 924–935 (2006)
3. Pepper, J.W., Scott Findlay, C., Kassen, R., Spencer, S.L., Maley, C.C.: Synthesis: Cancer research meets evolutionary biology. Evolutionary Applications **2**(1), 62–70 (2009)
4. Yates, L.R., Campbell, P.J.: Evolution of the cancer genome. Nat Rev Genet **13**(11), 795–806 (2012)
5. Greaves, M., Maley, C.C.: Clonal evolution in cancer. Nature **481**(7381), 306–313 (2012)
6. Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D., McMichael, J.F., Wallis, J.W., Lu, C., Shen, D., Harris, C.C., Dooling, D.J., Fulton, R.S., Fulton, L.L., Chen, K., Schmidt, H., Kalicki-Veizer, J., Magrini, V.J., Cook, L., McGrath, S.D., Vickery, T.L., Wendl, M.C., Heath, S., Watson, M.A., Link, D.C., Tomasson, M.H., Shannon, W.D., Payton, J.E., Kulkarni, S., Westervelt, P., Walter, M.J., Graubert, T.A., Mardis, E.R., Wilson, R.K., DiPersio, J.F.: Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature **481**(7382), 506–510 (2012)
7. Gillies, R.J., Verduzco, D., Gatenby, R.A.: Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. Nat Rev Cancer **12**(7), 487–493 (2012)
8. Burrell, R.A., McGranahan, N., Bartek, J., Swanton, C.: The causes and consequences of genetic heterogeneity in cancer evolution. Nature **501**(7467), 338–345 (2013)
9. Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., Beerenwinkel, N.: Reliable detection of subclonal single-nucleotide variants in tumour cell populations. Nature Communications **3** (2012)
10. Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Cote, A., Shah, S.P.: PyClone: statistical inference of clonal population structure in cancer. Nat Meth **11**(4), 396–398 (2014)
11. Ha, G., Roth, A., Khattra, J., Ho, J., Yap, D., Prentice, L.M., Melnyk, N., McPherson, A., Bashashati, A., Laks, E., Biele, J., Ding, J., Le, A., Rosner, J., Shumansky, K., Marra, M.A., Gilks, C.B., Huntsman, D.G., McAlpine, J.N., Aparicio, S., Shah, S.P.: TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. Genome Research **24**(11), 1881–1893 (2014)

12. Zare, H., Wang, J., Hu, A., Weber, K., Smith, J., Nickerson, D., Song, C., Witten, D., Blau, C.A., Noble, W.S.: Inferring clonal composition from multiple sections of a breast cancer. PLoS Comput Biol **10**(7), 1–15 (2014)
13. El-Kebir, M., Oesper, L., Acheson-Field, H., Raphael, B.J.: Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. Bioinformatics **31**(12), 62–70 (2015)
14. Navin, N.: Cancer genomics: one cell at a time. Genome Biology **15**(8), 452–465 (2014)
15. Jiao, W., Vembu, S., Deshwar, A.G., Stein, L., Morris, Q.: Inferring clonal evolution of tumors from single nucleotide somatic mutations. BMC Bioinformatics **15**(1), 1–16 (2014)
16. Deshwar, A.G., Vembu, S., Yung, C.K., Jang, G.H., Stein, L., Morris, Q.: Phylowgs: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biology **16**(1), 1–20 (2015)
17. El-Kebir, M., Satas, G., Oesper, L., Raphael, B.: Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures. Cell Systems **3**(1), 43–53 (2016)
18. Jiang, Y., Qiu, Y., Minn, A.J., Zhang, N.R.: Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proceedings of the National Academy of Sciences **113**(37), 5528–5537 (2016)
19. Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N.Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C.R., Nohadani, M., Eklund, A.C., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P.A., Swanton, C.: Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. New England Journal of Medicine **366**(10), 883–892 (2012)
20. Yates, L.R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T., Alexandrov, L.B., Larsimont, D., Davies, H., Li, Y., Ju, Y.S., Ramakrishna, M., Haugland, H.K., Lilleng, P.K., Nik-Zainal, S., McLaren, S., Butler, A., Martin, S., Glodzik, D., Menzies, A., Raine, K., Hinton, J., Jones, D., Mudie, L.J., Jiang, B., Vincent, D., Greene-Colozzi, A., Adnet, P.-Y., Fatima, A., Maetens, M., Ignatiadis, M., Stratton, M.R., Sotiriou, C., Richardson, A.L., Lonning, P.E., Wedge, D.C., Campbell, P.J.: Subclonal diversification of primary breast cancer revealed by multiregion sequencing. Nat Med **21**(7), 751–759 (2015). Article
21. Navin, N.E.: The first five years of single-cell cancer genomics and beyond. Genome Res **25**(10), 1499–1507 (2015)
22. Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., Wu, H., Ye, X., Ye, C., Wu, R., Jian, M., Chen, Y., Xie, W., Zhang, R., Chen, L., Liu, X., Yao, X., Zheng, H., Yu, C., Li, Q., Gong, Z., Mao, M., Yang, X., Yang, L., Li, J., Wang, W., Lu, Z., Gu, N., Laurie, G., Bolund, L., Kristiansen, K., Wang, J., Yang, H., Li, Y., Zhang, X., Wang, J.: Single-cell exome sequencing and monoclonal evolution of a jak2-negative myeloproliferative neoplasm. Cell **148**(5), 873–885 (2012)
23. Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., He, W., Zeng, L., Xing, M., Wu, R., Jiang, H., Liu, X., Cao, D., Guo, G., Hu, X., Gui, Y., Li, Z., Xie, W., Sun, X., Shi, M., Cai, Z., Wang, B., Zhong, M., Li, J., Lu, Z., Gu, N., Zhang, X., Goodman, L., Bolund, L., Wang, J., Yang, H., Kristiansen, K., Dean, M., Li, Y., Wang, J.: Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell **148**(5), 886–895 (2012)
24. Wang, Y., Waters, J., Leung, M.L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., Multani, A., Zhang, H., Zhao, R., Michor, F., Meric-Bernstam, F., Navin, N.E.: Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature **512**(7513), 155–160 (2014)
25. Gawad, C., Koh, W., Quake, S.R.: Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. Proceedings of the National Academy of Sciences **111**(50), 17947–17952 (2014)
26. Li, Y., Xu, X., Song, L., Hou, Y., Li, Z., Tsang, S., Li, F., Im, K., Wu, K., Wu, H., Ye, X., Li, G., Wang, L., Zhang, B., Liang, J., Xie, W., Wu, R., Jiang, H., Liu, X., Yu, C., Zheng, H., Jian, M., Nie, L., Wan, L., Shi, M., Sun, X., Tang, A., Guo, G., Gui, Y., Cai, Z., Li, J., Wang, W., Lu, Z., Zhang, X., Bolund, L., Kristiansen, K., Wang, J., Yang, H., Dean, M., Wang, J.: Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. GigaScience **1**(1), 12 (2012)
27. Zafar, H., Wang, Y., Nakhleh, L., Navin, N., Chen, K.: Monovar: single-nucleotide variant detection in single cells. Nat Meth **13**(6), 505–507 (2016)
28. Zhang, C.-Z., Adalsteinsson, V.A., Francis, J., Cornils, H., Jung, J., Maire, C., Ligon, K.L., Meyerson, M., Love, J.C.: Calibrating genomic and allelic coverage bias in single-cell sequencing. Nature Communications **6**, 6822 (2015)
29. Yu, C., Yu, J., Yao, X., Wu, W.K., Lu, Y., Tang, S., Li, X., Bao, L., Li, X., Hou, Y., Wu, R., Jian, M., Chen, R., Zhang, F., Xu, L., Fan, F., He, J., Liang, Q., Wang, H., Hu, X., He, M., Zhang, X., Zheng, H., Li, Q., Wu, H., Chen, Y., Yang, X., Zhu, S., Xu, X., Yang, H., Wang, J., Zhang, X., Sung, J.J., Li, Y., Wang, J.: Discovery of biclonal origin and a novel oncogene slc12a5 in colon cancer by single-cell sequencing. Cell Res **24**(6), 701–712 (2014)
30. Eirew, P., Steif, A., Khattra, J., Ha, G., Yap, D., Farahani, H., Gelmon, K., Chia, S., Mar, C., Wan, A., Laks, E., Biele, J., Shumansky, K., Rosner, J., McPherson, A., Nielsen, C., Roth, A.J.L., Lefebvre, C., Bashashati, A., de Souza, C., Siu, C., Aniba, R., Brimhall, J., Oloumi, A., Osako, T., Bruna, A., Sandoval, J.L., Algara, T., Greenwood, W., Leung, K., Cheng, H., Xue, H., Wang, Y., Lin, D., Mungall, A.J., Moore, R., Zhao, Y., Lorette, J., Nguyen, L., Huntsman, D., Eaves, C.J., Hansen, C., Marra, M.A., Caldas, C., Shah, S.P., Aparicio, S.: Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. Nature **518**(7539), 422–426 (2015)
31. Huelsenbeck, J.P., Ronquist, F.: Mrbayes: Bayesian inference of phylogenetic trees. Bioinformatics **17**(8), 754–755 (2001)
32. Yuan, K., Sakoparnig, T., Markowetz, F., Beerenwinkel, N.: Bitphylogeny: a probabilistic framework for reconstructing intra-tumor phylogenies. Genome Biology **16**(1), 1–16 (2015)
33. Ross, E.M., Markowetz, F.: OncoNEM: inferring tumor evolution from single-cell sequencing data. Genome Biology **17**(1), 1–14 (2016)
34. Jahn, K., Kuipers, J., Beerenwinkel, N.: Tree inference for single-cell data. Genome Biology **17**(1), 1–17 (2016)
35. Kim, K.I., Simon, R.: Using single cell sequencing data to model the evolutionary history of a tumor. BMC

Bioinformatics **15**(1), 27 (2014)
36. Ma, J., Ratan, A., Raney, B.J., Suh, B.B., Miller, W., Haussler, D.: The infinite sites model of genome evolution. Proceedings of the National Academy of Sciences **105**(38), 14254–14261 (2008)
37. Gusfield, D.: Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology. Cambridge University Press, Cambridge (1997)
38. Davis, A., Navin, N.E.: Computing tumor trees from single cells. Genome Biology **17**(1), 1–4 (2016)
39. Jukes, T.H., Cantor, C.R.: Chapter 24 - evolution of protein molecules. In: Mammalian Protein Metabolism, pp. 21–132. Academic Press, Cambridge (1969)
40. Yang, Z., Rannala, B.: Molecular phylogenetics: principles and practice. Nat Rev Genet **13**(5), 303–314 (2012)
41. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J.: A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet **43**(5), 491–498 (2011)
42. Robinson, D.F., Foulds, L.R.: Comparison of phylogenetic trees. Mathematical Biosciences **53**(1), 131–147 (1981)
43. Gusfield, D.: ReCombinatorics: The Algorithmics of Ancestral Recombination Graphs and Explicit Phylogenetic Networks. The MIT Press, Cambridge (2014)
44. Miller, C.A., Gindin, Y., Lu, C., Griffith, O.L., Griffith, M., Shen, D., Hoog, J., Li, T., Larson, D.E., Watson, M., Davies, S.R., Hunt, K., Suman, V.J., Snider, J., Walsh, T., Colditz, G.A., DeSchryver, K., Wilson, R.K., Mardis, E.R., Ellis, M.J.: Aromatase inhibition remodels the clonal architecture of estrogen-receptor-positive breast cancers. Nature Communications **7** (2016)
45. Gao, R., Davis, A., McDonald, T.O., Sei, E., Shi, X., Wang, Y., Tsai, P.-C., Casasent, A., Waters, J., Zhang, H., Meric-Bernstam, F., Michor, F., Navin, N.E.: Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. Nat Genet **48**(10), 1119–1130 (2016)
46. Baslan, T., Kendall, J., Ward, B., Cox, H., Leotta, A., Rodgers, L., Riggs, M., D'Italia, S., Sun, G., Yong, M., Miskimen, K., Gilmore, H., Saborowski, M., Dimitrova, N., Krasnitz, A., Harris, L., Wigler, M., Hicks, J.: Optimizing sparse sequencing of single cells for highly multiplex copy number profiling. Genome Research **25**(5), 714–724 (2015)
47. Leung, M.L., Wang, Y., Kim, C., Gao, R., Jiang, J., Sei, E., Navin, N.E.: Highly multiplexed targeted dna sequencing from single nuclei. Nat. Protocols **11**(2), 214–235 (2016). Protocol
48. Felsenstein, J.: Evolutionary trees from dna sequences: A maximum likelihood approach. Journal of Molecular Evolution **17**(6), 368–376 (1981)
49. Hastings, W.K.: Monte carlo sampling methods using markov chains and their applications. Biometrika **57**(1), 97–109 (1970)
50. Lakner, C., van der Mark, P., Huelsenbeck, J.P., Larget, B., Ronquist, F.: Efficiency of markov chain monte carlo tree proposals in bayesian phylogenetics. Systematic Biology **57**(1), 86–103 (2008)
51. Pupko, T., Pe, I., Shamir, R., Graur, D.: A fast algorithm for joint reconstruction of ancestral amino acid sequences. Molecular Biology and Evolution **17**(6), 890–896 (2000)

Figures

Additional Files

Additional file 1 — Supplementary Material
This file contains supplementary figures.

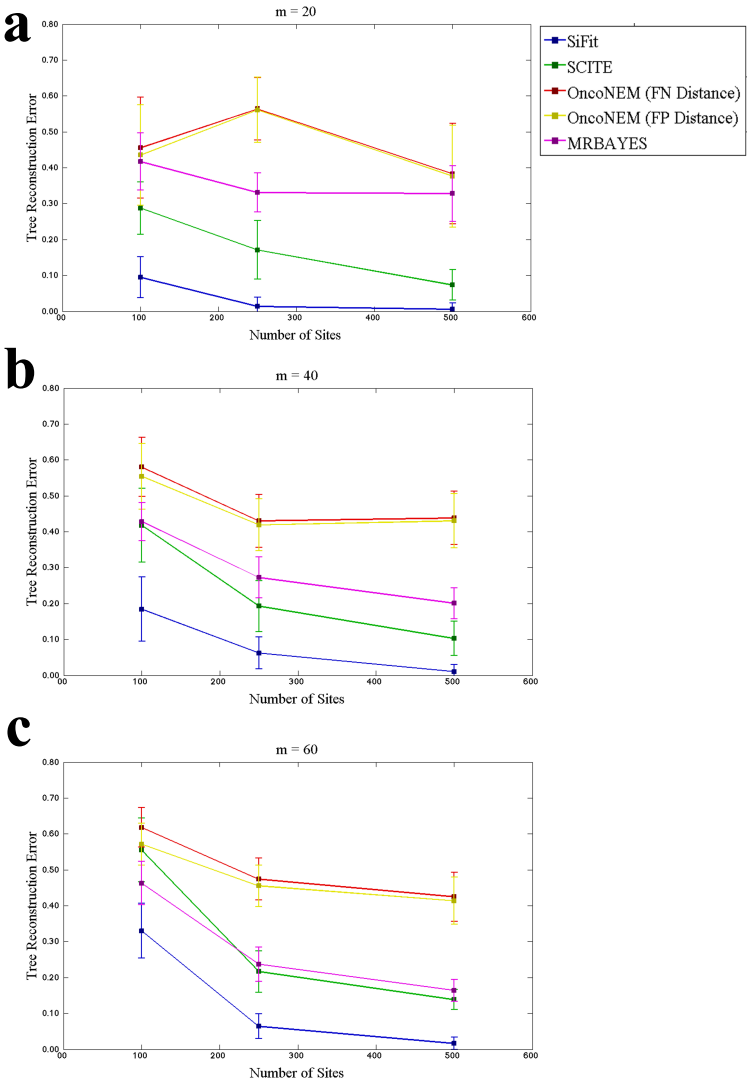

Figure 1 Performance comparison on datasets with varying number of cells. SiFit's tree reconstruction accuracy is compared against that of SCITE, OncoNEM and MRBAYES. The y-axis denotes the tree reconstruction error that measures the distance of the inferred tree from the ground truth. Three points for $n = 100$, $n = 250$ and $n = 500$ are plotted on the x-axis. In each case, the mean tree reconstruction error over 20 datasets is plotted. The vertical error bar indicates the standard deviation of tree reconstruction error over 20 datasets. **a** Performance comparison for datasets with 20 cells. **b** Performance comparison for datasets with 40 cells. **c** Performance comparison for datasets with 60 cells.

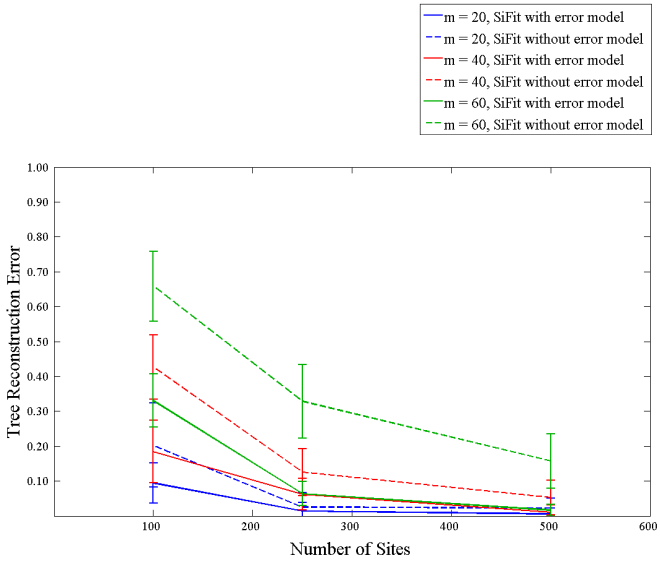

Figure 2 Effect of SiFit's SCS error model. SiFit's tree reconstruction accuracy is compared under two conditions, in the first case, SiFit employs both the error model and model of evolution, whereas in the second case, the error model is excluded. The y-axis denotes the tree reconstruction error that measures the distance of the inferred tree from the ground truth. Three points for $n = 100$, $n = 250$ and $n = 500$ are plotted on the x-axis. In each case, the mean tree reconstruction error over 20 datasets is plotted. The vertical error bar indicates the standard deviation of tree reconstruction error over 20 datasets. The solid lines represent the cases when SiFit accounts for SCS errors while the dotted lines represent the cases when SiFit does not account for SCS errors. The blue, red and green lines compare results for datasets with 20, 40 and 60 cells respectively.

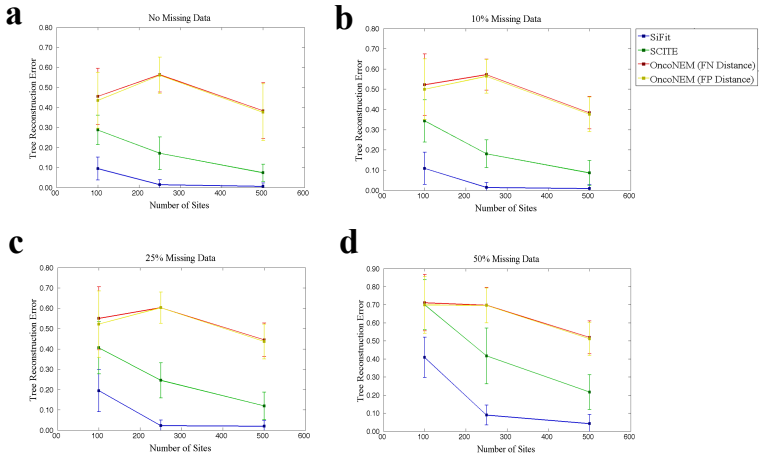

Figure 3 Performance comparison on datasets with missing data. SiFit's tree reconstruction accuracy is compared against that of SCITE and OncoNEM on datasets with missing data. The y-axis denotes the tree reconstruction error that measures the distance of the inferred tree from the ground truth. Three points for $n = 100$, $n = 250$ and $n = 500$ are plotted on the x-axis. In each case, the mean tree reconstruction error over 20 datasets is plotted. The vertical error bar indicates the standard deviation of tree reconstruction error over 20 datasets. **a** Comparison for datasets without any missing data. **b** Comparison for datasets with 10% missing data. **c** Comparison for datasets with 25% missing data. **d** Comparison for datasets with 50% missing data.

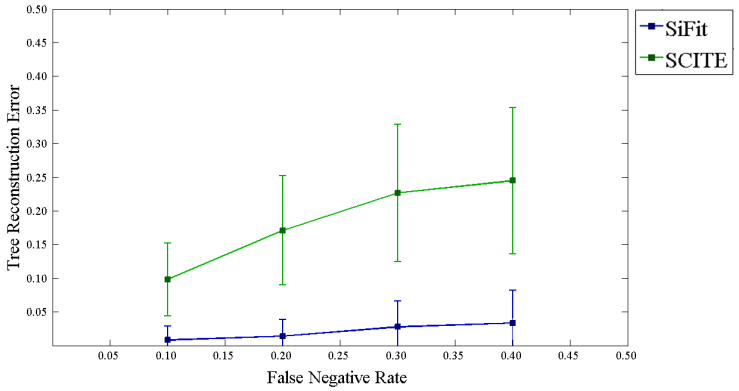

Figure 4 Effect of increase in error rates. SiFit's tree reconstruction accuracy is compared against that of SCITE for increasing false negative rate. The y-axis denotes the tree reconstruction error that measures the distance of the inferred tree from the ground truth. Four points corresponding to false negative rate $= \{0.1, 0.2, 0.3, 0.4\}$ are plotted. In each case, the mean tree reconstruction error over 20 datasets is plotted. The vertical error bar indicates the standard deviation of tree reconstruction error over 20 datasets.

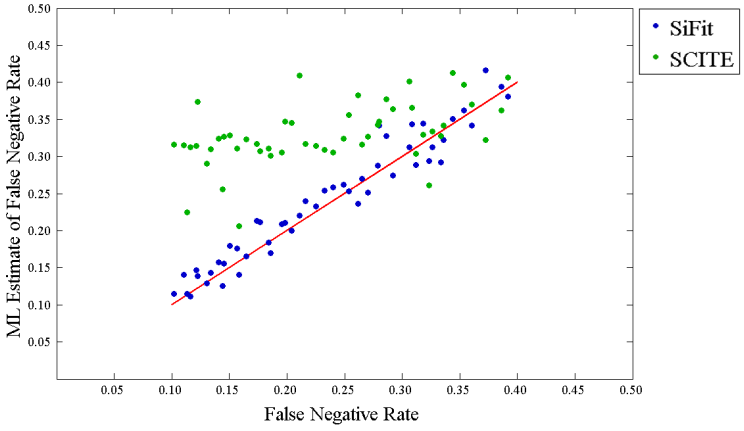

Figure 5 Estimation of error rates. The ML estimate of false negative rate is compared against the false negative rate used for generating the data. The red line represents the perfect estimate (correlation coefficient $= 1$). The blue dots represent the estimates by SiFit, the green dots represent the estimates by SCITE.

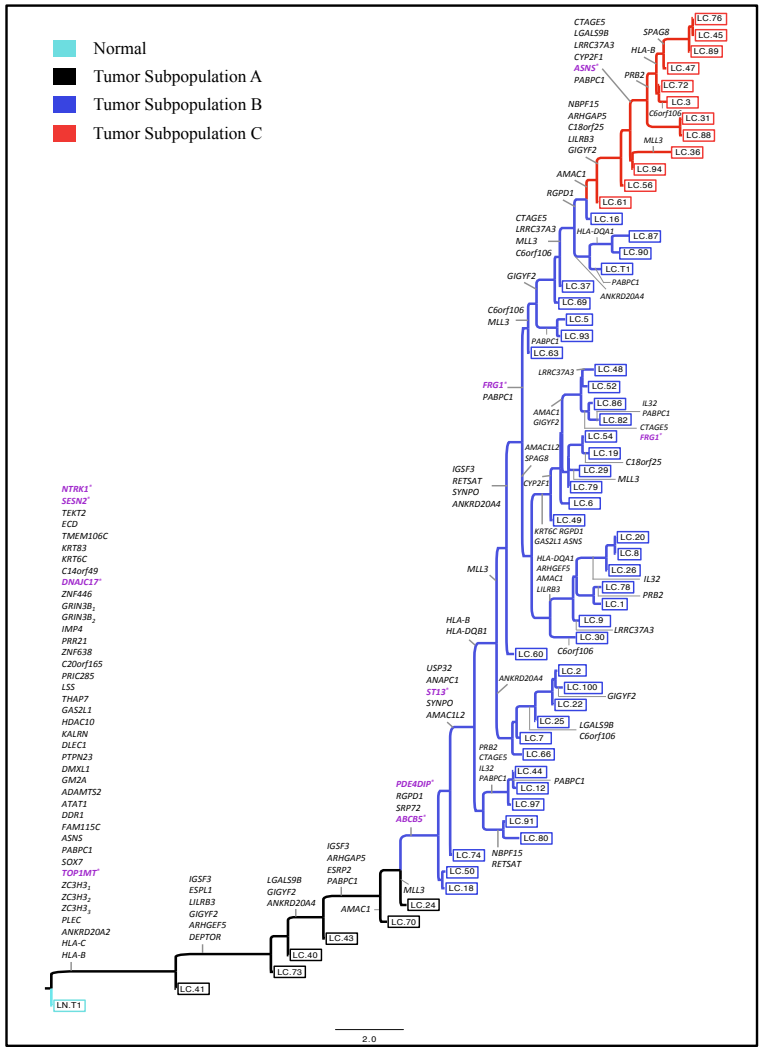

Figure 6 Maximum Likelihood phylogenetic tree reconstructed by SiFit for JAK2-negative myeloproliferative neoplasm. LN.T1 represents the normal bulk tissue. All other leaves are single tumor cells. Non-synonymous mutations are annotated on the branches of the tree. The driver genes reported in the original study and additional cancer genes are marked in purple.

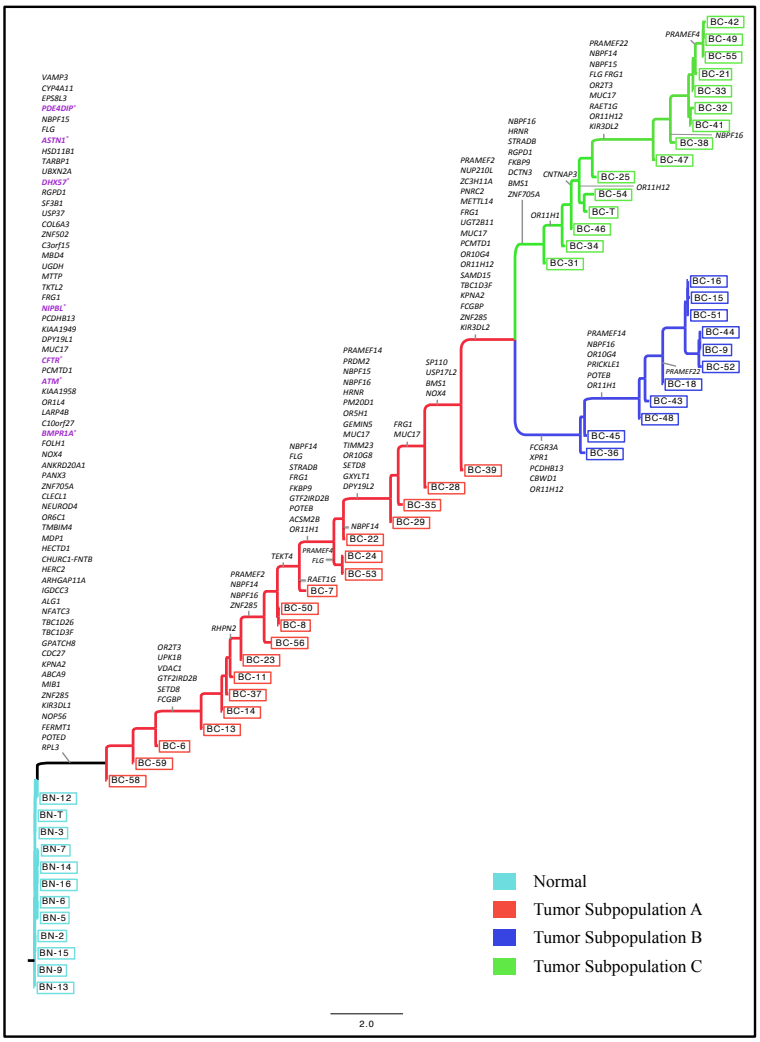

Figure 7 Maximum Likelihood phylogenetic tree reconstructed by SiFit for muscle-invasive bladder transitional cell carcinoma. The leaves with legend marked with "BN" represent the normal single cells. The leaves with legend marked with "BC" are tumor cells. Nonsynonymous mutations are annotated on the branches of the tree. The driver genes reported in the original study and additional cancer genes are marked in purple.

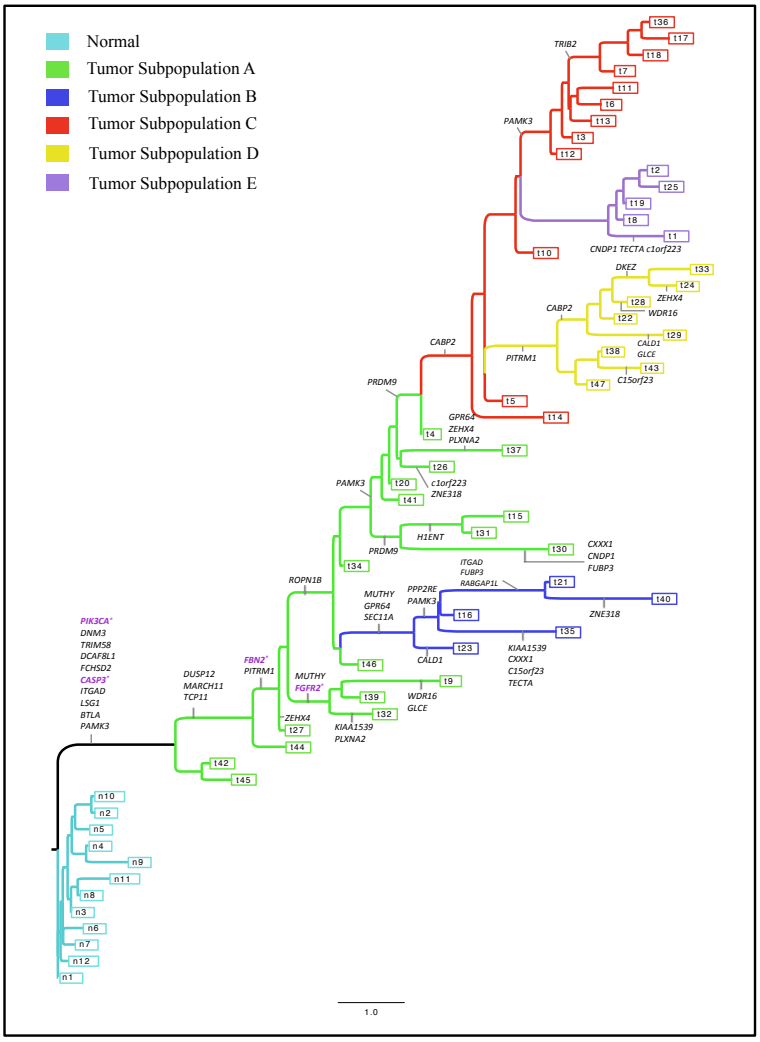

Figure 8 Maximum Likelihood phylogenetic tree reconstructed by SiFit for Estrogen-receptor positive (ER^+) breast cancer. The leaves marked with 'n' are normal cells, the leaves marked with 't' are tumor cells. Nonsynonymous mutations are annotated on the branches of the tree. The cancer genes are marked in purple.