Accounting for errors in data improves timing in single-cell cancer evolution

Kylie Chen[1*], Jiří C. Moravec[2], Alex Gavryushkin[2,3], David Welch[1], Alexei J. Drummond[1,4]

1 School of Computer Science, University of Auckland, Auckland, New Zealand
2 Department of Computer Science, University of Otago, Dunedin, New Zealand
3 School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand
4 School of Biological Sciences, University of Auckland, Auckland, New Zealand

* kche309@aucklanduni.ac.nz

Abstract

Single-cell sequencing provides a new way to explore the evolutionary history of cells. Compared to traditional bulk sequencing, where a population of heterogeneous cells is pooled to form a single observation, single-cell sequencing isolates and amplifies genetic material from individual cells, thereby preserving the information about the origin of the sequences. However, single-cell data is more error-prone than bulk sequencing data due to the limited genomic material available per cell. Here, we present error and mutation models for evolutionary inference of single-cell data within a mature and extensible Bayesian framework, BEAST2. Our framework enables integration with biologically informative models such as relaxed molecular clocks and population dynamic models. Our simulations show that modeling errors increase the accuracy of relative divergence times and substitution parameters. We reconstruct the phylogenetic history for a colorectal cancer patient and a healthy patient from single-cell DNA sequencing data. We find that the estimated times of terminal splitting events are shifted forward in time compared to models which ignore errors. We observed that not accounting for errors can overestimate the phylogenetic

diversity in single-cell DNA sequencing data. We estimate that 30-50% of the apparent diversity can be attributed to error. Our work enables a full Bayesian approach capable of accounting for errors in the data within the integrative Bayesian software framework BEAST2.

Introduction

The growth of cancer cells can be viewed as an evolutionary process where mutations accumulate along cell lineages over time. Within each cell, single nucleotide variants (SNVs) act as markers for the evolutionary process. By sampling and sequencing cells, we can reconstruct the possible evolutionary histories for these cell lineages. This can provide insight into the timing of events and modes of evolution.

Currently, there are two main methods for obtaining genomic sequences, bulk sequencing, and single-cell sequencing. Bulk sequencing data is traditionally used in genomic studies. By pooling the genetic material from many cells to form a single observation, greater coverage and thus genetic signal is retained. However, in the context of cancer phylogenetics, the analysis of bulk data poses challenges. Firstly the intermixing of tumor and normal cells affects the genomic signal. Secondly, the pooled sample may be heterogeneous and thus contain a mixture of different genomic variants (Dagogo-Jack and Shaw, 2018; de Bruin et al., 2014; Liu et al., 2018).

In contrast, single-cell sequencing isolates and amplifies the genetic material within a single cell (Kuipers et al., 2017). The isolation step alleviates the mixture problem. However, errors are more problematic for single-cell sequencing due to the insufficient coverage and limited amount of genetic material available. Sources of errors include (1) allelic dropout, where one of the alleles fails to be amplified; (2) sequencing error; (3) amplification error; and (4) cell doublets, where two cells are sequenced as one by mistake (Kuipers et al., 2017; Lähnemann et al., 2020; Woodworth et al., 2017). Various error models have been proposed to address these issues, such as models based on false positive/false negative probabilities (Jahn et al., 2016; Ross and Markowetz, 2016; Zafar et al., 2017, 2019), models of allelic dropout/sequencing errors (Kozlov et al., 2021), and models of read count errors (Satas et al., 2020).

To enable easy integration with molecular clock and phylogeography models that are commonly used in other areas of phylogenetics (Kearns et al., 2018;

Malmstrøm et al., 2016; Meijer et al., 2012), we implemented error models described in (Jahn et al., 2016; Kozlov et al., 2021; Ross and Markowetz, 2016; Zafar et al., 2017, 2019) within BEAST2 (Bouckaert et al., 2019), a mature Bayesian evolutionary framework. Our motivation is to enable inference of both the evolutionary history and model parameters for single-cell phylogenetics. We show our implementation is well-calibrated (Dawid, 1982) and demonstrate the models on both real and simulated single-cell DNA data. We also show through simulation studies that not accounting for errors leads to inaccurate estimation of timing and substitution parameters when data is error-prone. Our results suggest that using a model that is not error-aware can significantly overestimate the number of substitutions and hence the evolutionary time scale. Consequently, 30-50% of the phylogenetic diversity can be attributed to errors. We note that although these methods are developed with cancer analysis in mind, they are also applicable to non-cancer single-cell phylogenetics such as somatic cell evolution.

Related work

For bulk sequencing, there are many tools that estimate the clonal compositions in each bulk sample (Jiang et al., 2016; Miura et al., 2018; Popic et al., 2015) and infer their clonal history (Alves et al., 2019; Cooper et al., 2015; Heide et al., 2019). These clone inference tools are most applicable to bulk sequencing samples that contain a mixture of clones, and phylogenetic reconstruction is performed on the identified clones. However, with the recent availability of single-cell technology, variations between cells can be studied more directly (Schwartz and Schäffer, 2017). This has led to the development of tools for single-cell phylogenetics.

As errors present a key challenge to the analysis of single-cell data, there is a need for models that account for errors introduced during the sequencing process, missing data, coverage discrepancies (Lee et al., 2020), and the ability to quantify uncertainty (Lähnemann et al., 2020). A range of error models have been proposed: binary/ternary mutation models parameterized by false positive (FP) and false negative (FN) errors such as OncoNEM (Ross and Markowetz, 2016), SCITE (Jahn et al., 2016) and SiFit algorithms (Zafar et al., 2017, 2019); genotype models with sequencing/allelic dropout errors in CellPhy (Kozlov et al., 2021); and read count error models as in SCARLET (Satas et al., 2020).

Both OncoNEM (Ross and Markowetz, 2016) and SCITE (Jahn et al., 2016) are infinite site models which assume mutations can only occur at a site once. The SiFit algorithms (Zafar et al., 2017, 2019) relax the infinite sites constraint. SiFit uses a deletion-aware finite site ML model, and SiFitClone extends this to infer clonal clusters from mutation data using a Bayesian mixture model, while accounting for FP, FN, and cell doublets. SCARLET (Satas et al., 2020) uses a ML model combining copy number and SNVs, where errors are modeled on the read counts. First, a copy number tree is inferred, then the copy number tree is refined using ML with SNV information. CellPhy (Kozlov et al., 2021) is the first model that makes use of single-cell genotypes and incorporates sequencing and allelic dropout errors within its model.

Besides CellPhy, which uses an ML phylogenetic framework, most models are standalone implementations. ML approaches typically only provide uncertainty for the tree topology, and not for other model parameters. Incorporating error models in a Bayesian phylogenetic framework would allow uncertainty in model parameters such as error parameters to be quantified. Additionally, using a phylogenetic framework enables recent developments from other areas of phylogenetics to be used for single-cell phylogenetics (e.g., population dynamics models and location-aware phylogeography models).

In this paper, we implement error models for binary and diploid genotype data and investigate whether errors impact the time scale of evolutionary trees inferred from single-cell data. We perform preliminary analysis on real single-cell data and show how error models can be used with population growth and molecular clock models. The next section describes the evolutionary models used in this study.

Materials and Methods

We implement two sets of models: a binary model, which handles data with two states, and a GT16 model, which handles diploid nucleotide data with sixteen states. The mutation process is modeled as a substitution process evolving along the branches of a tree τ, with mutation rates defined by the substitution matrix Q. Errors are modeled as a noisy process on tip sequences of the tree, where the true genotype is obfuscated according to probabilities of the error matrix E. To perform inference on data, we sample the posterior distribution of trees and other model parameters using Markov chain Monte Carlo (MCMC).

Binary model

The presence or absence of a mutation within a diploid cell is represented as a binary state $S = \{1, 0\}$. The mutation rate $0 \to 1$ is fixed at 1, and a parameter λ for the relative rate of $1 \to 0$, so that the rate matrix Q is

$$Q = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} -1 & 1 \\ \lambda & -\lambda \end{pmatrix} \end{array},$$

with equilibrium frequencies

$$\pi_0 = \lambda/(\lambda + 1),$$
$$\pi_1 = 1/(\lambda + 1).$$

To account for false-positive (FP) and false-negative (FN) errors, we implement the binary error model described in (Jahn et al., 2016; Ross and Markowetz, 2016; Zafar et al., 2017) with the error matrix E, which has the following entries

$$E_{00} = P(0|0) = 1 - \alpha,$$
$$E_{10} = P(1|0) = \alpha,$$
$$E_{01} = P(0|1) = \beta,$$
$$E_{11} = P(1|1) = 1 - \beta,$$

where $P(x|y)$ is the probability of observing data x given that the true state is y so that α is the FP probability, and β is the FN probability.

GT16 diploid nucleotide model

To model diploid nucleotide sequences, we implement the GT16 substitution and error model described in (Kozlov et al., 2021). The GT16 substitution model is an extension of the four-state general time-reversible nucleotide GTR model (Tavaré et al., 1986) to sixteen states where each state represents one phasing of the diploid genotype. Let a, b, c, d denote different alleles chosen from the nucleotide set $S \in \{A, C, G, T\}$, r_{ab} be the rate of going from allele a to b, and $\pi = (\pi_{AA}, \ldots, \pi_{TT})$ be the equilibrium genotype frequencies. The elements in

the instantaneous rate matrix Q can be written as:

$$
\begin{aligned}
Q_{aa\to ab} &= r_{ab} \cdot \pi_{ab}, \\
Q_{aa\to ba} &= r_{ab} \cdot \pi_{ba}, \\
Q_{ab\to aa} &= r_{ab} \cdot \pi_{aa}, \\
Q_{ab\to bb} &= r_{ab} \cdot \pi_{bb}.
\end{aligned}
$$

Other non-diagonal entries not listed above have a rate of zero. The diagonals are the outgoing rates which are the sum of the ingoing rates

$$
\begin{aligned}
Q_{aa\to aa} &= -\sum_{b,c\in S\setminus a} Q_{aa\to bc}, \\
Q_{ab\to ab} &= -\sum_{\substack{c,d\in S \\ c\neq a \text{ or } d\neq b}} Q_{ab\to cd}.
\end{aligned}
$$

The relative rates of the Q matrix are

$$
\begin{aligned}
r_{AC} &= r_{CA} = \alpha, \\
r_{AG} &= r_{GA} = \beta, \\
r_{AT} &= r_{AT} = \gamma, \\
r_{CG} &= r_{GC} = \kappa, \\
r_{CT} &= r_{TC} = \lambda, \\
r_{GT} &= r_{TG} = \mu.
\end{aligned}
$$

The GT16 error model for diploid nucleotides described in (Kozlov et al., 2021) accounts for amplification errors and biases in single-cell sequencing. This model has two parameters, the amplification and sequencing error ϵ, and the allelic dropout error δ. The entries of the error matrix E are derived in (Kozlov et al., 2021).

Let a, b, c denote different alleles, ϵ denote the amplification/sequencing error, and δ the allelic dropout error such that $\epsilon, \delta \in [0, 1)$. If the true genotype is homozygous aa, then the probabilities of observing the genotypes aa, ab, bb are

$$
\begin{aligned}
P(aa|aa) &= 1 - \epsilon + (1/2) \cdot \delta\epsilon, \\
P(ab|aa) &= (1 - \delta) \cdot (1/6) \cdot \epsilon, \\
P(bb|aa) &= (1/6) \cdot \delta\epsilon.
\end{aligned}
$$

If the true genotype is heterozygous ab, then the probabilities of observing the genotypes aa, ab, cc are

$$\begin{aligned}
P(aa|ab) &= (1/2) \cdot \delta + (1/6) \cdot \epsilon - (1/3) \cdot \delta\epsilon, \\
P(cc|ab) &= (1/6) \cdot \delta\epsilon, \\
P(ac|ab) &= (1-\delta)(1/6) \cdot \epsilon, \\
P(ab|ab) &= (1-\delta)(1-\epsilon).
\end{aligned}$$

For other combinations of observed/true alleles, the probability is assumed to be zero. Note that the genotypes given above assume phased data, $ab \neq ba$. For unphased data, these can be encoded as ambiguities for heterozygous states, $P(ab^*) = P(ab) + P(ba)$, where ab^* represents ab without phasing information.

Likelihood calculation

We use the binary model to illustrate how the likelihood is calculated. Suppose we observe a sequence D of length n. The likelihood can be calculated independently at each site D_i using Felsenstein's peeling algorithm (Felsenstein, 1981)

$$P(D|\tau, \lambda, \alpha, \beta) = \prod_{i=1}^{n} P(D_i|\tau, \lambda, \alpha, \beta), \tag{1}$$

where the tree is τ, substitution matrix is Q and model parameters are λ, α, β. Let $S = \{0, 1\}$ be the set of possible genotypes. The likelihood at the root r is

$$P(r|\tau, \lambda, \alpha, \beta) = \sum_{s_0 \in S} P(r = s_0) \cdot L(r = s_0), \tag{2}$$

where $P(x = s_0) = \pi_{s_0}$ is the equilibrium frequency of state s_0. The partial likelihood L for internal node x with child nodes y_1, y_2 is

$$L(x = s_0) = \sum_{s_1 \in S} P_{s_0,\, s_1}(t_1) \cdot L(y_1 = s_1) \cdot \sum_{s_2 \in S} P_{s_0,\, s_2}(t_2) \cdot L(y_2 = s_2), \tag{3}$$

where t_1 is the time between x and y_1 on tree τ, t_2 is the time between x and y_2. $P_{ij}(t) = e^{Qt}$ is the probability of going from state i to j after time t. The partial likelihood for leaf node x using the binary error model is

$$L(x = 0 \mid \alpha,\ \beta) = (1 - \alpha,\ \beta), \tag{4}$$

$$L(x = 1 \mid \alpha,\ \beta) = (\alpha,\ 1 - \beta) \tag{5}$$

For brevity, we only fully write out the likelihood of the binary model. The likelihood for the GT16 error model can be derived similarly and is available in our implementation.

Results

Evaluation on simulated datasets

First, we evaluated our implementations using a well-calibrated study (Dawid, 1982) to test the reliability of the inference when simulating directly from the model. Following the well-calibrated criterion for credible intervals, we expect 95% of the credible interval to cover the true value 95% of the time. Next, we simulated sequences with errors, and then compared the inference performance with and without modeling the error.

Simulation 1: Binary data

We performed a well-calibrated study for the binary model using binary sequences with errors. First, we generated trees using a Yule model, then binary sequences were simulated along the branches of the tree, and errors were applied at the tips. Using the sequence data, we jointly estimate the model parameters, tree topology, and branch times using the binary model. The simulation parameters are detailed in [1].

Figure S1 shows the estimates for the model parameters, tree length, and tree height compared to the true simulated values. The estimated 95% highest posterior density (HPD) intervals are shown as bars, where blue indicates the estimate covers the true value, and red indicates otherwise. Our simulations show that the true value of each parameter falls within the estimated 95% HPD interval 91-99% of the time. Figure S2 shows the estimated trees are on average 2-5 subtree prune and regraft (SPR) moves away from the true tree.

[1] Simulation 1 parameters: We generated 100 trees with 30 leaves from a Yule model, where each tree has a birthrate drawn from $Normal(\mu = 7.0, \sigma = 1.0)$. Sequences of length 400 were simulated using the binary model with $\lambda \sim Lognormal(\mu = -1)$, and error parameters $\alpha \sim Beta(1, 50)$, $\beta \sim Beta(1, 50)$.

Simulation 2: Binary data error vs. no error

To compare the effects of inference with and without error modeling, we used the data from Simulation 1, then performed inference with and without the binary error model. We compared the coverage for each parameter with and without an error model, i.e., how often the estimated 95% HPD covers the true value.

Figure S1 shows the estimated parameters with the binary error model, and Figure S3 shows the estimated parameters without using an error model. Table S1 shows the coverage of each parameter. The coverage of tree length drops from 95% when the error model is used to 39% when no error model is used. Similarly, the coverage of the substitution parameter λ drops 91% to 53%. Furthermore, the tree length tends to be overestimated when no error model is used. This suggests that both the tree length and substitution parameters are significantly biased when errors present in the data are not modeled. Figure 1 shows a comparison of the estimated tree length and tree height for these two model configurations. Other parameters such as birthrate and tree height are less biased, with a coverage of 92% and 84% respectively when no error model is used.

Simulation 3: Diploid nucleotide phased data

We performed a well-calibrated study for the GT16 model using phased sequence data with errors. First, we simulated trees using a coalescent model. Sequences were simulated down branches of the tree using the GT16 substitution model, and then errors were applied at the tips. Using these sequences as input, we estimated the tree and model parameters using the GT16 model. The simulation parameters are detailed in [2].

Figures S4-S9 show the 95% HPD estimated for each model parameter and tree branch lengths. The true value of each parameter falls within the 95% HPD interval 91-99% of the time. This shows we are able to accurately estimate the substitution, error, population parameters, and branch lengths for phased data. Next, we computed the accuracy of tree topology by comparing the estimated tree with the true tree. Figure S10 shows the average distance from the

[2] Simulation 3 parameters: We generated 100 trees with 16 leaves from a coalescent model, where the population size is drawn from $\theta \sim LogNormal(\mu = -2.0,\ \sigma = 1.0)$. Sequences of length 200 were simulated using the GT16 model with $\pi \sim Dirichlet(3, 3, \ldots, 3)$, $r \sim Dirichlet(1, 2, 1, 1, 2, 1)$, and error parameters $\epsilon \sim Beta(\alpha = 2,\ \beta = 18)$, $\delta \sim Beta(\alpha = 1.5,\ \beta = 4.5)$.

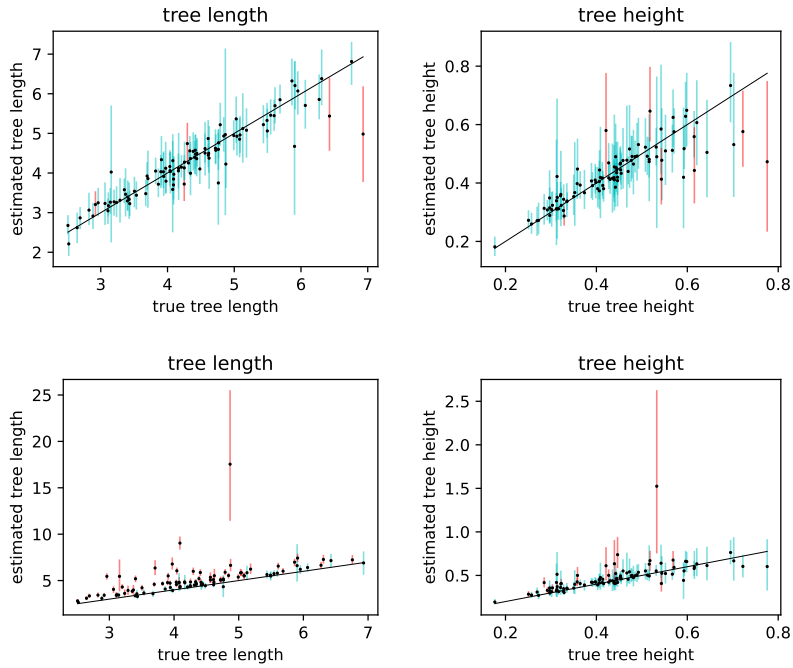

Fig 1. Comparison of branch lengths with and without the binary error model using simulated binary data. Estimated branch lengths using the binary error model (top) and without using the error model (bottom).

estimated trees to the true tree. On average, estimated trees are 2-6 SPR moves away from the true tree.

Simulation 4: Diploid nucleotide unphased data

We performed a well-calibrated study for the GT16 model using unphased sequencing data. For unphased data, we used the data generated from Simulation 3, with phasing information removed from the sequences. Phasing information was removed by mapping a heterozygous ab to both states ab and ba.

Figures S11-S15 show the estimated model parameters compared to the true simulated values. For each parameter, the estimated 95% HPD interval covers the true value 94-99% of the time, which confirms our implementation is well-calibrated. On average, the estimated trees are 2-6 SPR moves away from the true tree, Figure S16. We note that sum of paired heterozygous frequencies $(\pi_{ab} + \pi_{ba})$ are identifiable, but the individual frequencies (π_{ab}, π_{ba}) are non-identifiable as the data is unphased.

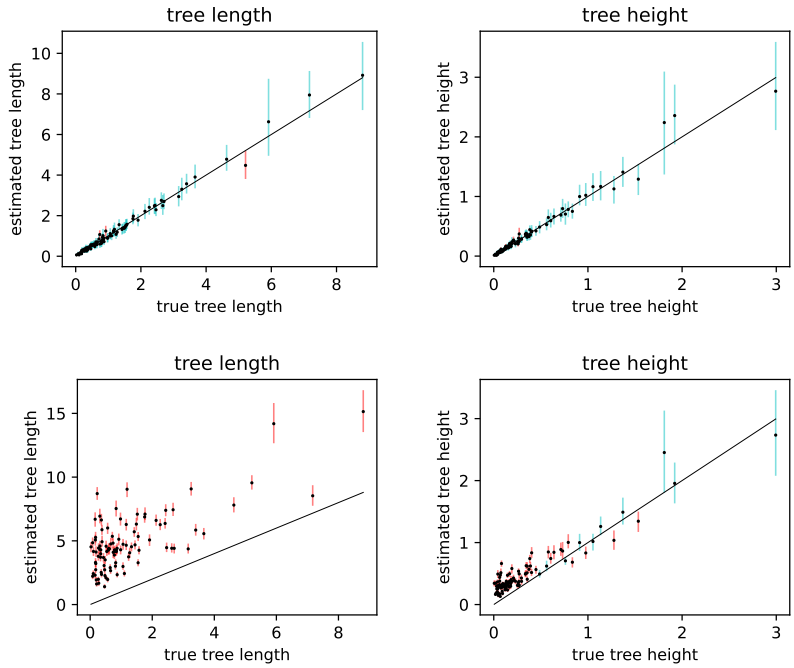

Fig 2. Comparison of branch lengths with and without the GT16 error model using simulated phased nucleotide data. Estimated branch lengths using the GT16 error model (top) and without using the error model (bottom).

Simulation 5: Diploid nucleotide data error vs. no error

To compare the effects of inference with and without error modeling for diploid nucleotide data, we used the data from Simulation 3, then performed inference with and without the GT16 error model.

Table S2 shows the coverage of each parameter with and without an error model. Figures S17-S21 show the estimated model parameters when an error model is not used. We observe a similar trend to Simulation 2, where the tree length and substitution parameters are significantly biased without an error model. Although the tree height estimated without an error model are less biased, the tree lengths are overestimated. These differences in the tree heights and tree lengths are highlighted in Figure 2.

Evaluation on single-cell datasets

We analyzed two public datasets from previously published studies; L86, a colorectal cancer dataset (Leung et al., 2017), and E15, a healthy neurons dataset (Evrony et al., 2015). Preprocessed SNVs from CellPhy (Kozlov et al.,

2021) were used for both L86 and E15.

Colorectal cancer dataset (L86)

L86 contains 86 cells sequenced from a metastatic colorectal cancer patient. The cells were sampled from the primary tumor (colorectal), the secondary metastatic tumor (liver), and matched normal tissue. We used the GT16 model with a relaxed clock to allow for different molecular clock rates in cancer and non-cancer lineages and a coalescent skyline tree prior, which allows changes in population sizes through time. The parameters of the model are detailed in [3].

We found that the tree height is similar for both error and non-error models, but the relative ages of terminal branches are shorter for the error model. Figure 3 shows the tree length and tree height distributions under different experimental setups: with and without error modeling, and with and without an outgroup constraint. For the default setup without an outgroup, the 95% HPD estimate for tree length is $(4.79, 5.85)$ with the error model and $(7.00, 8.19)$ without the error model. The estimated tree height is $(0.12, 0.14)$ with the error model, and $(0.09, 0.12)$ without the error model. The estimated interval for the error parameters are $\delta \sim (0.62, 0.66)$ and $\epsilon \sim (7 \cdot 10^{-}6, 1 \cdot 10^{-}3)$. The error estimates are comparable to the estimates reported by CellPhy (Kozlov et al., 2021) which are $\delta \sim 0.63$ and $\epsilon \sim 0.00$.

Figure 4 summarizes the estimated tree with the error model (top), compared to without an error model (bottom). The tips of the tree are colored by cell type. The trees show most cells group together by cell type, which suggests there is signal in the data. However, there is some intermixing of metastatic tumor cells inside the primary tumor clade and missorted normal cells as previously identified by (Kozlov et al., 2021; Leung et al., 2017). We also note that the most recent common ancestor (MRCA) of the normal clade is younger than the MRCA of the two tumor clades for both analyses. This is not what we intuitively expected because we believe the normal ancestral cell should be the ancestor of both tumor and normal cells. Although surprising, this observation is in agreement with the trees estimated by ML algorithms in (Kozlov et al., 2021).

We also investigated whether constraining the tree topology to have all

[3] L86 model parameters: GT16 substitution model frequencies $\pi \sim Dirichlet(3, 3, \ldots, 3)$, relative rates $r \sim Dirichlet(1, 2, 1, 1, 2, 1)$, GT16 error model $\delta \sim Beta(1.5, 4.5), \epsilon \sim Beta(2, 18)$. Relaxed clock with Lognormal prior. Skyline coalescent tree prior with $\theta_1 \sim Lognormal(\mu = -2.3, \sigma = 1.8)$. We repeated the inference on two independent MCMC chains.

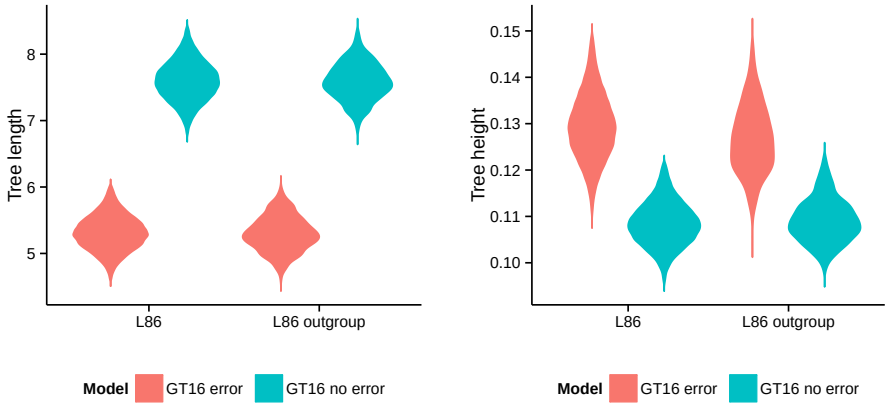

Fig 3. Estimated tree length and height distributions for the L86 dataset. The distributions of tree length (left) and tree height (right) are colored by the model used, where red plots show estimates using the GT16 error model, and blue shows estimates without using an error model. Two pairs of experiments are shown; L86, which has no tree topology constraints, and L86 outgroup, which has the tree topology constrained to normal cells in the outgroup. Outliers more than 1.5 interquartile away from the first and third quartiles are not shown on this plot.

tumor cells as the ingroup produces different estimates. We repeat the analyses with an outgroup constraint to the tree topology, setting the normal and missorted samples as the outgroup. Figure S22 shows the outgroup constrained summary tree for L86. In this outgroup constrained tree, the age of the MRCA of the normal group is also younger than the MRCA of the primary or metastatic tumors for the error model. The age of the MRCA of the normal group is indistinguishable from the MRCA of the primary or metastatic tumors for the model without error.

Healthy neurons dataset (E15)

E15 contains 15 neurons, and a blood cell taken from the heart region sequenced from a healthy patient. The model parameters are detailed in [4]. As in the L86

[4]E15 model parameters: We used a GT16 substitution and error model, with a Dirichlet prior on frequencies $\pi \sim Dirichlet(3, 3, \ldots, 3)$ and relative rates $r \sim Dirichlet(1, 2, 1, 1, 2, 1)$, Beta priors on the error parameters $\delta \sim Beta(1.5, 4.5), \epsilon \sim Beta(2, 18)$, with a relaxed clock and Skyline coalescent tree prior with $\theta_1 \sim Lognormal(\mu = -2.3, \sigma = 1.8)$. We repeat the inference on two independent MCMC chains.

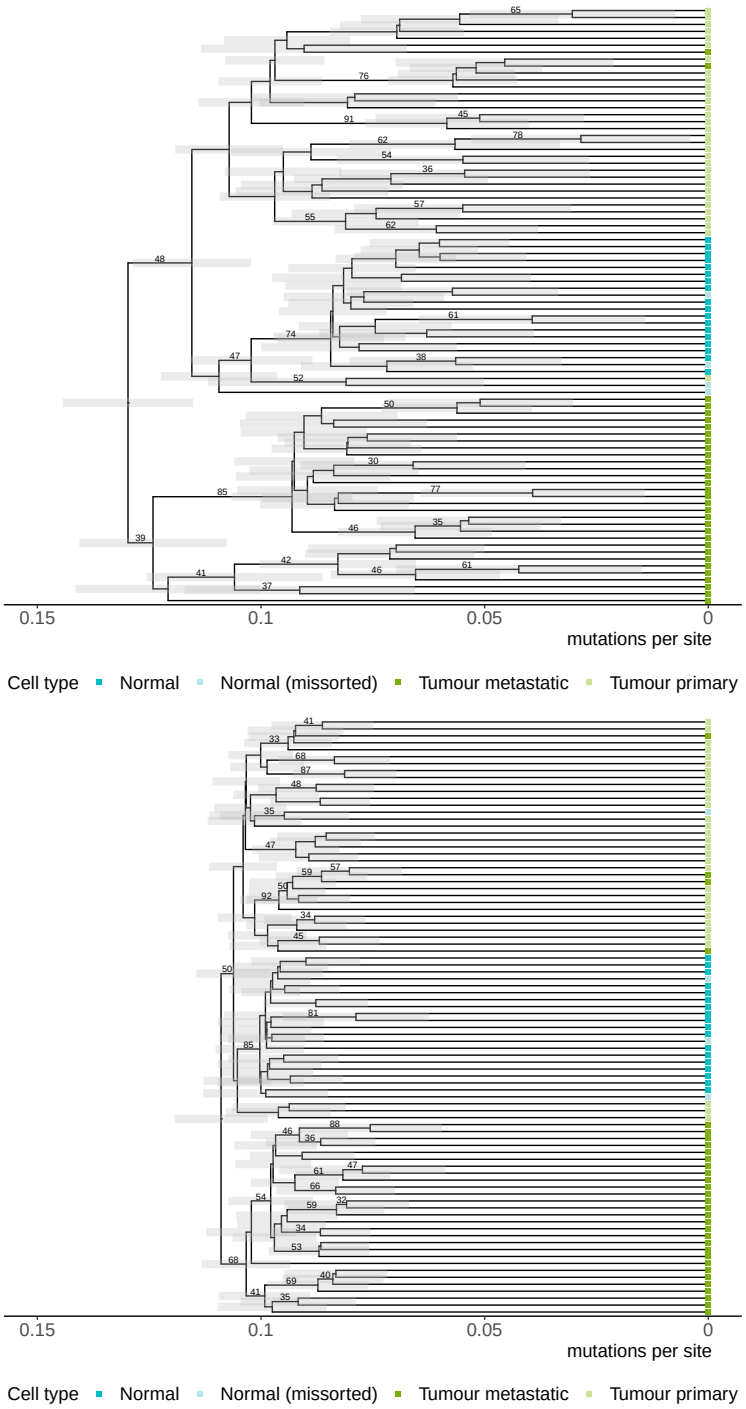

Fig 4. Maximum clade credibility trees for the L86 dataset (colorectal cancer patient) using the GT16 model with an error model (top) and without an error model (bottom). Cells are colored by their cell types: normal cells (dark blue), normal cells missorted during data collection (light blue), the metastatic tumor from the liver (dark green), and primary tumor from the colon (light green). Clades with greater than 30% are labeled.

dataset, we found that the relative ages of the tip branches are shorter for the error model (Fig 5). The 95% HPD estimates of tree length are $(1.37, 7.14)$ with the error model, and $(7.60, 13.41)$ without the error model. However the tree height for the error model $(0.20, 0.80)$ is lower than that of the non-error model $(0.54, 1.01)$. The estimated interval for the error parameters are $\delta \sim (0.86, 0.92)$ and $\epsilon \sim (0.03, 0.17)$. To test the sensitivity of the error priors, we reran our experiments with adjusted priors $\epsilon, \delta \sim Beta(1, 10)$ and $\epsilon, \delta \sim Beta(1, 20)$. We found the error estimates were similar regardless of these adjustments on the error parameter priors.

Figure 6 shows a summary of the estimated trees with the GT16 error model (top) and without an error model (bottom). The tips of the tree are colored by cell types. We expect the blood cell to be placed as an outgroup; however, the estimated trees placed the blood cell inside a clade of neuron cells. To investigate if adding an outgroup constraint to the tree topology can help direct the likelihood in the correct direction, we repeated the analyses with the blood cell as the outgroup. The estimated trees are shown in figure S23. Besides the correct placement of the outgroup enforced by the outgroup constraint, we did not observe any substantial topological discrepancies between the outgroup and non-outgroup analyses.

Discussion

We demonstrated that incorporating error parameters can affect the relative ages of single-cell datasets. We showed that models incorporating sequencing error could increase the accuracy of tree branches and model parameters inferred from noisy data. From both simulated and real single-cell data, we observed that using an error model tends to shorten the total tree length, as errors explain a portion of the genetic variability within the data. For empirical single-cell data, cells of the same type tend to be placed in the same clade; however, caution should be exercised when interpreting the outgroup and relative ages of heterogeneous data. Further work to identify better fitting molecular clock models may improve current analyses. Additionally, time sampled data would allow node ages to be converted to real time (Drummond et al., 2002, 2003).

Although the effect of filtering strategies in the context of macroevolution shows stringent filtering of sites often leads to worse phylogenetic inference (Tan et al., 2015). The effect of filtering strategies on noisy data such as single-cell

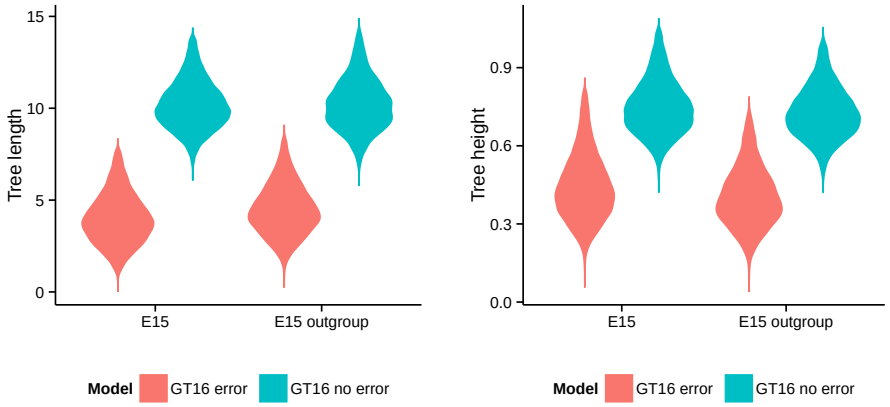

Fig 5. Estimated tree length and height distributions for the E15 dataset. The distributions of tree length (left) and tree height (right) are colored by the model used, where red plots show estimates using the GT16 error model and blue shows estimates without an error model. Two pairs of experiments are shown; E15, which has no tree topology constraints, and E15 outgroup, which has the tree topology constrained with the heart cell as the outgroup. Outliers more than 1.5 interquartile away from the first and third quartiles are not shown on this plot.

phylogenies is yet to be systematically explored. We believe the error parameters in these models can provide increased flexibility, allowing key features of the sequencing and filtering process to be accounted for during evolutionary inference.

Lastly, incorporating cell biology knowledge during method development would improve the biological significance of model assumptions; and improving the interpretability of tree summarization metrics would enable single-cell phylogenies to be examined in more detail.

Code and data availability

The Phylonco v0.0.5 package is available on `www.github.com/bioDS/beast-phylonco`. The following packages were used in this paper BEAST v2.6.5 (Bouckaert et al., 2019), BeastLabs v1.9.0, LPhy v0.1.0, and LPhyBeast v0.1.0. Analysis and scripts used in this paper will be made available upon publication. The following R packages were used ggtree (Yu et al., 2017), phangorn (Schliep, 2011), tracerR (Rambaut et al.,

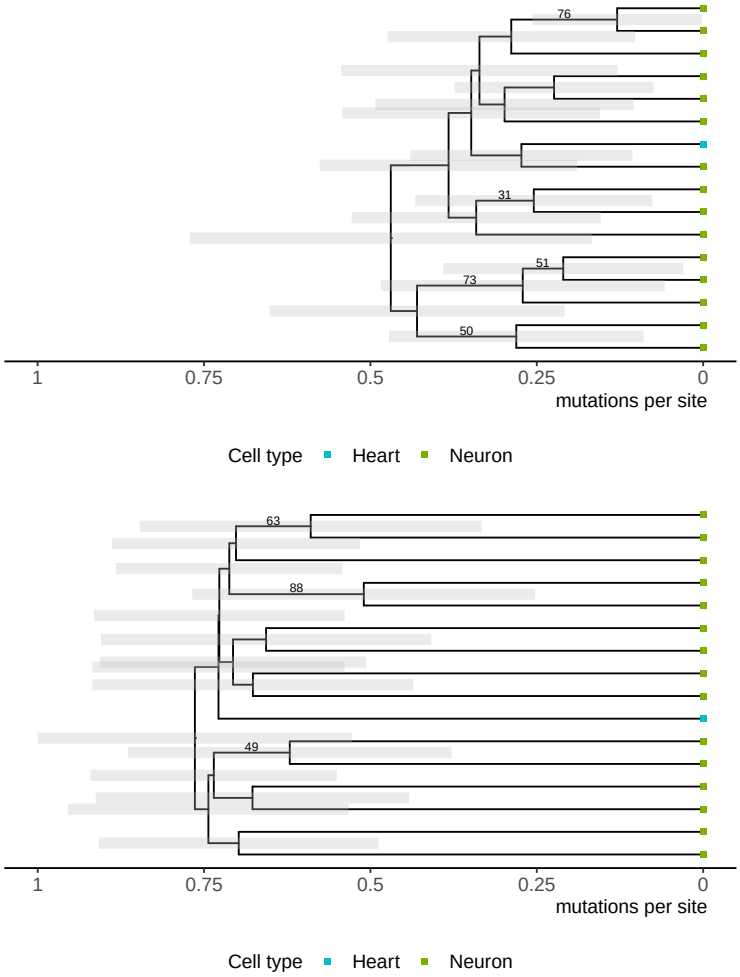

Fig 6. Maximum clade credibility trees for the E15 dataset (healthy patient) using the GT16 model with an error model (top) and without an error model (bottom). Each cell is colored by its cell type: blood cell (blue) and neuron cells (green). The posterior clade support for clades with greater than 30% support are shown on the branches.

2018), treeSim (Stadler, 2015), treeSimGM (Hagen and Stadler, 2018), ggplot2 (Wickham, 2016), and patchwork (Pedersen, 2020).

Acknowledgements

We thank Prof. David Posada and Dr. Joao Alves for generously providing us single-cell datasets and advice. We thank Dr. Remco Bouckaert for helpful discussion and Dr. Walter Xie for implementation support. We acknowledge Nectar Research Cloud and Otago Computing Services for providing the computing resources necessary for this study. AG and AJD were funded by the MBIE Endeavour Fund, ref. U00X1912; AG was supported by Rutherford Discovery Fellowship ref. RDF-UOO1702; AJD was supported by the James Cook Research Fellowship from the Royal Society of New Zealand.

References

Joao M Alves, Sonia Prado-Lopez, Jose Manuel Cameselle-Teijeiro, and David Posada. Rapid evolution and biogeographic spread in a colorectal cancer. Nature communications, 10(1):1–7, 2019.

Remco Bouckaert, Timothy G. Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, Michael Matschiner, Fábio K. Mendes, Nicola F. Müller, Huw A. Ogilvie, Louis du Plessis, Alex Popinga, Andrew Rambaut, David Rasmussen, Igor Siveroni, Marc A. Suchard, Chieh-Hsi Wu, Dong Xie, Chi Zhang, Tanja Stadler, and Alexei J. Drummond. Beast 2.5: An advanced software platform for bayesian evolutionary analysis. PLOS Computational Biology, 15(4):1–28, 04 2019. doi: 10.1371/journal.pcbi.1006650. URL https://doi.org/10.1371/journal.pcbi.1006650.

Colin S Cooper, Rosalind Eeles, David C Wedge, Peter Van Loo, Gunes Gundem, Ludmil B Alexandrov, Barbara Kremeyer, Adam Butler, Andrew G Lynch, Niedzica Camacho, et al. Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. Nature genetics, 47(4): 367–372, 2015.

Ibiayi Dagogo-Jack and Alice T Shaw. Tumour heterogeneity and resistance to cancer therapies. Nature reviews Clinical oncology, 15(2):81, 2018.

A Philip Dawid. The well-calibrated bayesian. Journal of the American Statistical Association, 77(379):605–610, 1982.

Elza C de Bruin, Nicholas McGranahan, Richard Mitter, Max Salm, David C Wedge, Lucy Yates, Mariam Jamal-Hanjani, Seema Shafi, Nirupa Murugaesu, Andrew J Rowan, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science, 346(6206):251–256, 2014.

Alexei J Drummond, Geoff K Nicholls, Allen G Rodrigo, and Wiremu Solomon. Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. Genetics, 161(3): 1307–1320, 07 2002. ISSN 1943-2631. doi: 10.1093/genetics/161.3.1307. URL https://doi.org/10.1093/genetics/161.3.1307.

Alexei J. Drummond, Oliver G. Pybus, Andrew Rambaut, Roald Forsberg, and Allen G. Rodrigo. Measurably evolving populations. Trends in Ecology & Evolution, 18(9):481–488, 2003. ISSN 0169-5347. doi: https://doi.org/10.1016/S0169-5347(03)00216-7.

Gilad D Evrony, Eunjung Lee, Bhaven K Mehta, Yuval Benjamini, Robert M Johnson, Xuyu Cai, Lixing Yang, Psalm Haseley, Hillel S Lehmann, Peter J Park, et al. Cell lineage analysis in human brain using endogenous retroelements. Neuron, 85(1):49–59, 2015.

Joseph Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. Journal of molecular evolution, 17(6):368–376, 1981.

Oskar Hagen and Tanja Stadler. Treesimgm: Simulating phylogenetic trees under general bellman–harris models with lineage-specific shifts of speciation and extinction in r. Methods in ecology and evolution, 9(3):754–760, 2018.

Timon Heide, Angela Maurer, Monika Eipel, Katrin Knoll, Mirja Geelvink, Juergen Veeck, Ruth Knuechel, Julius van Essen, Robert Stoehr, Arndt Hartmann, et al. Multiregion human bladder cancer sequencing reveals tumour evolution, bladder cancer phenotypes and implications for targeted therapy. The Journal of pathology, 248(2):230–242, 2019.

Katharina Jahn, Jack Kuipers, and Niko Beerenwinkel. Tree inference for single-cell data. Genome biology, 17(1):86, 2016.

Yuchao Jiang, Yu Qiu, Andy J Minn, and Nancy R Zhang. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proceedings of the National Academy of Sciences, 113(37):E5528–E5537, 2016.

Anna M Kearns, Marco Restani, Ildiko Szabo, Audun Schrøder-Nielsen, Jin Ah Kim, Hayley M Richardson, John M Marzluff, Robert C Fleischer, Arild Johnsen, and Kevin E Omland. Genomic evidence of speciation reversal in ravens. Nature Communications, 9(1):1–13, 2018.

Alexey Kozlov, Joao M. Alves, Alexandros Stamatakis, and David Posada. Cellphy: accurate and fast probabilistic inference of single-cell phylogenies from scdna-seq data. bioRxiv, 2021. doi: 10.1101/2020.07.31.230292. URL https://www.biorxiv.org/content/early/2021/04/29/2020.07.31.230292.

Jack Kuipers, Katharina Jahn, and Niko Beerenwinkel. Advances in understanding tumour evolution through single-cell sequencing. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 1867(2):127–138, 2017.

David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. Genome biology, 21(1):1–35, 2020.

Jeongwoo Lee, Daehee Hwang, et al. Single-cell multiomics: technologies and data analysis methods. Experimental & Molecular Medicine, 52(9): 1428–1442, 2020.

Marco L Leung, Alexander Davis, Ruli Gao, Anna Casasent, Yong Wang, Emi Sei, Eduardo Vilar, Dipen Maru, Scott Kopetz, and Nicholas E Navin. Single-cell dna sequencing reveals a late-dissemination model in metastatic colorectal cancer. Genome research, 27(8):1287–1299, 2017.

Jinping Liu, Hien Dang, and Xin Wei Wang. The significance of intertumor and intratumor heterogeneity in liver cancer. Experimental & molecular medicine, 50(1):e416–e416, 2018.

Martin Malmstrøm, Michael Matschiner, Ole K Tørresen, Bastiaan Star, Lars G Snipen, Thomas F Hansen, Helle T Baalsrud, Alexander J Nederbragt, Reinhold Hanel, Walter Salzburger, et al. Evolution of the immune system influences speciation rates in teleost fishes. Nature Genetics, 48(10): 1204–1210, 2016.

Adam Meijer, Sabine van der Sanden, Bianca EP Snijders, Giovanna Jaramillo-Gutierrez, Louis Bont, Cornelis K van der Ent, Pieter Overduin, Shireen L Jenny, Edin Jusic, Harrie GAM van der Avoort, et al. Emergence and epidemic occurrence of enterovirus 68 respiratory infections in the netherlands in 2010. Virology, 423(1):49–57, 2012.

Sayaka Miura, Karen Gomez, Oscar Murillo, Louise A Huuki, Tracy Vu, Tiffany Buturla, and Sudhir Kumar. Predicting clone genotypes from tumor bulk sequencing of multiple samples. Bioinformatics, 34(23):4017–4026, 06 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty469. URL https://doi.org/10.1093/bioinformatics/bty469.

Thomas Lin Pedersen. patchwork: The Composer of Plots, 2020. URL https://CRAN.R-project.org/package=patchwork. R package version 1.1.1.

Victoria Popic, Raheleh Salari, Iman Hajirasouliha, Dorna Kashef-Haghighi, Robert B West, and Serafim Batzoglou. Fast and scalable inference of multi-sample cancer lineages. Genome biology, 16(1):91, 2015.

Andrew Rambaut, Alexei J Drummond, Dong Xie, Guy Baele, and Marc A Suchard. Posterior summarization in bayesian phylogenetics using tracer 1.7. Systematic biology, 67(5):901, 2018.

Edith M Ross and Florian Markowetz. Onconem: inferring tumor evolution from single-cell sequencing data. Genome biology, 17(1):1–14, 2016.

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

Klaus Peter Schliep. phangorn: phylogenetic analysis in r. Bioinformatics, 27(4): 592–593, 2011.

Russell Schwartz and Alejandro A Schäffer. The evolution of tumour phylogenetics: principles and practice. Nature Reviews Genetics, 18(4): 213–229, 2017.

Tanja Stadler. Treesim: simulating phylogenetic trees. R package version, 2(2), 2015.

Ge Tan, Matthieu Muffato, Christian Ledergerber, Javier Herrero, Nick Goldman, Manuel Gil, and Christophe Dessimoz. Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference. Systematic biology, 64(5):778–791, 2015.

Simon Tavaré et al. Some probabilistic and statistical problems in the analysis of dna sequences. Lectures on mathematics in the life sciences, 17(2):57–86, 1986.

Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Mollie B Woodworth, Kelly M Girskis, and Christopher A Walsh. Building a lineage from single cells: genetic techniques for cell lineage tracking. Nature Reviews Genetics, 18(4):230, 2017.

Guangchuang Yu, David K Smith, Huachen Zhu, Yi Guan, and Tommy Tsan-Yuk Lam. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution, 8(1):28–36, 2017.

Hamim Zafar, Anthony Tzen, Nicholas Navin, Ken Chen, and Luay Nakhleh. Sifit: inferring tumor trees from single-cell sequencing data under finite-sites models. Genome biology, 18(1):178, 2017.

Hamim Zafar, Nicholas Navin, Ken Chen, and Luay Nakhleh. Siclonefit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. Genome research, 29(11): 1847–1859, 2019.