Ignoring errors causes inaccurate timing of single-cell phylogenies

Kylie Chen[1*], David Welch[1], Alexei J. Drummond[1,2]

1 School of Computer Science, University of Auckland, Auckland, New Zealand
2 School of Biological Sciences, University of Auckland, Auckland, New Zealand

[*] kche309@aucklanduni.ac.nz

Author contributions: All authors conceptualized and designed the analyses, wrote and edited the manuscript. KC developed the software, analyzed the data and performed the experiments.

Abstract

Single-cell sequencing provides a new way to explore the evolutionary history of cancers. Compared to traditional bulk sequencing, which samples multiple heterogeneous cells, single-cell sequencing isolates and amplifies genetic material from a single cell. The ability to isolate a single cell makes it ideal for evolutionary inference. However, single-cell data is more error-prone due to the limited genomic material available per cell. Previous work using single-cell data to reconstruct the evolutionary history of cancers has not been integrated with standard evolutionary models. 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. We reconstruct the phylogenetic history for a myeloproliferative cancer patient and two colorectal cancer patients. We find that the estimated times of terminal splitting events are shifted forward in time compared to models which ignore errors. Furthermore, we estimate 50% - 70% of the evolutionary distance between samples can be explained by sequencing error. Our simulation studies show that ignoring errors leads to inaccurate estimates of divergence times, mutation parameters and population parameters. Our work opens the potential for integrative Bayesian models capable of combining multiple sources of data.

Author summary

Cell heterogeneity is one of the hallmarks of cancer evolution. It is this cell-to-cell genotype variation that makes the analysis of traditional bulk samples very challenging. Recent advances in single-cell sequencing technology allows cell evolution to be traced directly, but is more error-prone than traditional methods. This motivates us to incorporate error models when inferring evolutionary history from genomic data. We used single-cell data to infer and time the evolutionary history of cancers whilst accounting for sequencing errors. Conventional methods for studying cancer evolution either do not account for errors or are limited in modeling scope. We present a mutation model that accounts for errors in a mature framework, BEAST2, which has inbuilt molecular clock models. We show errors can significantly impact the interpretation and timing of recent divergence events. Our results show sequencing errors can explain up to 50% - 70% of the genetic diversity within single-cell cancer samples.

Introduction

Cancer can be viewed as an evolutionary process where mutations accumulate along cell lines over time. Within each cell, single nucleotide variants (SNVs) act as markers for the evolutionary process. By sampling and sequencing tumour cells, we can reconstruct the possible evolutionary histories for these cell lines. 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 traditionally used in genomic studies are challenging to analyze in this context as each sample contains a population of cells. The amount of genomic signal depends on the purity of the sample, as tumour cells may be intermixed with normal cells. Additionally, tumours are often heterogeneous, containing many different genomic variants [1–3]. The detection of low-frequency variants in bulk samples can be particularly difficult. In contrast, single-cell sequencing isolates and heavily amplifies the genetic material within a single cell [4]. The isolation step alleviates the mixture problem, but errors can be magnified due to amplification of limited genetic material [4–6].

Existing models that attempt to account for sequencing error mainly use a maximum likelihood framework [7–10], providing only point estimates and are unable to quantify uncertainty in the estimated tree topology or error parameters. As single-cell data is so error-prone, quantifying uncertainty is a key challenge in its analysis [6]. A Bayesian approach provides a natural framework for quantifying uncertainty, combining multiple data sources, and incorporating prior knowledge. Bayesian methods are standard in other applications of phylogenetics, with a long history of use for studying population dynamics, speciation, and epidemiology [11–13]. However, the use of Bayesian methods in cancer phylogenetics is still limited. There is a need for cancer-specific models that are able to account for errors, sampling biases and cell biology. Our motivation is to enable inference of both the evolutionary history and model parameters specifically for single-cell cancer genomics.

Our work addresses site and error models for tip sequences in a mature Bayesian evolutionary framework, BEAST2 [14]. Our method allows easy integration with other informative models such as molecular clocks and phylogeography models. We show our implementation is well-calibrated [15] and demonstrate our 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 and timings when data is error-prone. This is a crucial first step in bringing powerful Bayesian phylogenetic tools to single-cell cancer phylogenetics.

Related work

For bulk sequencing, many tools have been developed to estimate the clonal compositions in each bulk sample [16–18] and inference of clonal history [19–21]. Clone inference tools [16–21] are most applicable to bulk sequencing samples that are mixtures of clones, where phylogenetic reconstruction is performed on the identified clones rather than directly on the sequences. However, with the recent availability of single-cell technology, variations between cells can be studied directly [22].

Single-cell lineage reconstruction approaches are mostly limited to parsimony [23, 24] or maximum likelihood methods [10, 25]. Parsimony aims to find a tree topology requiring the minimum number of changes but is misleading when assumptions are violated [26]. Recently, parsimony methods have been applied to single-cell DNA [27] and triple-omics sequencing [28]. Maximum likelihood methods find the evolutionary tree that maximizes the likelihood but are unable to quantify the uncertainty in the estimated tree or model parameters. Currently, most analyses of single-cell data do not inherently account for biases. As such, there is a need for models that account for systematic biases, missing data, coverage discrepancies [29], and quantify uncertainty [6]. Previous studies that deal with errors in single-cell DNA data are mostly unable to quantify uncertainty [7–10]. These include maximum likelihood approaches for infinite sites models [8], finite sites models [7] and Dollo models [9]. Bayesian clonal reconstruction methods that account for inference uncertainty have been explored previously [30]. However, since clonal inference ignores variation within clones, there is a significant loss of information when this method is applied to single-cell data which affect estimates of divergence times.

A natural question that arises from this is whether errors significantly impact the inferred evolutionary history. Previous work examined how errors in ancestral fossil ages affect evolutionary inference [31–33]. A recent paper [33] shows that when uncertainty in the age of ancestral samples are present but not modeled both the topology and timing of events become inaccurate. In this paper, we focus on modeling errors in tip sequences from extant samples. Typical applications in cancer biology are biopsies or autopsy samples. In the next section, we present the results of our Bayesian method on real and simulated single-cell data.

Results

Evaluation on single-cell datasets

We performed analyses on Myeloproliferative neoplasm [34] and colorectal cancer datasets [35] using BEAST2 [14]. Sequences were converted to binary format, where state 0 represents the normal genotype, and state 1 the mutant genotype. Binary error and binary substitution models with a Yule tree prior were used. We performed two separate Monte Carlo Markov Chain runs for each setup and selected broad uninformative priors for our model parameters.

We used the following priors for our model: $\alpha \sim Beta(1,5)$, $\beta \sim Beta(1,5)$, $\lambda \sim Lognormal(0,1)$, and $birth\ rate \sim Gamma(0.001, 1000)$, where α is the conditional false-positive probability, β is the conditional false-negative probability, λ is the relative back-mutation rate and *birth rate* is the birth rate of the Yule process. We note that the estimated conditional error probabilities for our data $\hat{\alpha}, \hat{\beta}$ are not directly comparable to unconditional error rates reported from previous experimental studies [36].

We compare the estimated model parameters and phylogenies with and without the use of an error model. We find estimates of the error and birth rate parameters were robust to prior specification. Figure 1 shows the topologies of the estimated tree posterior. Figure 2 shows the estimated posterior distributions of the tree height and tree length in units of substitutions per site. The estimated model parameters; substitution, birth rate, conditional false-positive, and conditional false-negative are shown in Supplementary Materials (Fig. S1 - S3).

Myeloproliferative neoplasms dataset

The myeloproliferative neoplasms dataset [34] contains 58 single-cell DNA cancer samples from a single Essential Thrombocythemia (ET) patient, quality filtered to 712 sites as published by the authors. We binarize the data to reflect the presence or absence of a mutant allele. Homozygous mutant sites which make up 6% of the alignment are mapped to state 1 as a preprocessing step. The highest posterior density (HPD) for the tree height is (0.18, 0.29) substitutions per site under the error model, and (0.42, 0.56) when errors are ignored. The HPD interval for the tree length is (3.34, 5.93) under the

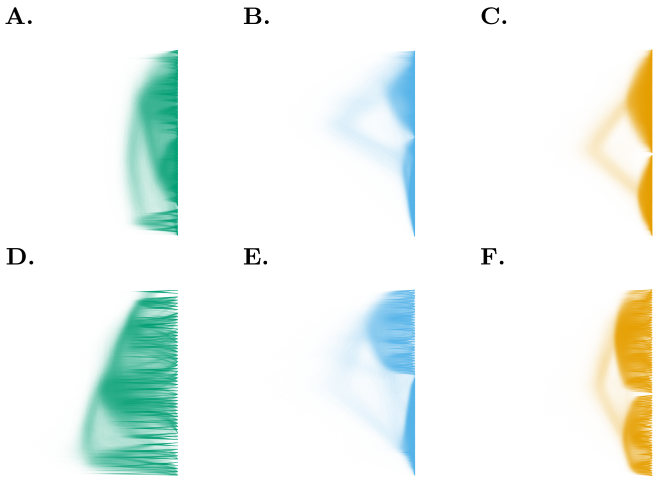

Fig 1. Comparison of tree topology with and without the binary error model. Densitree figures showing sampled trees. From left to right, (A) Myeloproliferative neoplasms dataset with error, (B) Colorectal dataset 1 with error, (C) Colorectal dataset 2 with error, (D) Myeloproliferative neoplasms dataset without error, (E) Colorectal dataset 1 without error, (F) Colorectal dataset 2 without error.

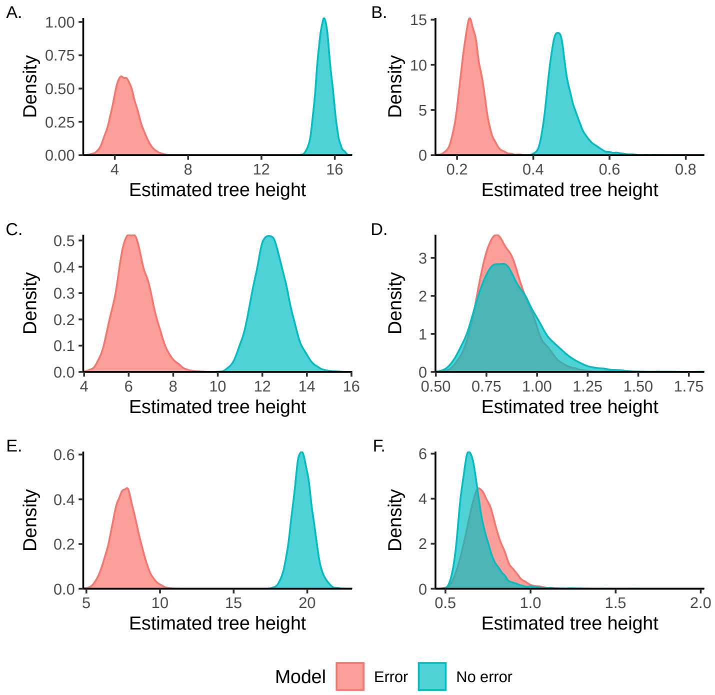

Fig 2. Estimated tree metrics for each dataset. Estimated values under the binary error model are depicted in red, estimates without an error model are depicted in blue. The tree scale is in substitutions per site, without correction for constant sites. Each row corresponds to the estimated tree length and tree height for a dataset; Myeloproliferative neoplasms (A, B), Colorectal dataset 1 (C, D), and Colorectal dataset 2 (E, F).

error model and (14.69, 16.17) when errors are ignored. The error estimates are (0.12, 0.16) for the conditional false-positive probability, and (0.14, 0.18) for the conditional false-negative probability. The maximum clade credibility tree is shown in Fig. 3.

Colorectal dataset 1

The colorectal cancer dataset (CRC0827 / K047) [35] contains 50 single-cell DNA cancer samples and 19 adenoma polyp samples from a male colorectal patient, quality filtered by [35] to 198 sites. The authors identify two cancer subclones (1CA and 2CA) and one adenoma polyp clone (XR). The evolution of cancer and polyp cells in this patient is thought to be monoclonal [35]. The HPD interval for the tree height is (0.64, 1.08) substitutions per site under the error model, and (0.60, 1.17) when errors are ignored. The HPD interval for the tree length is (4.85, 7.81) under the error model and (10.86, 13.85) when errors are ignored. The error estimates are (3.02×10^{-7}, 2.56×10^{-3}) for the conditional false-positive probability, and (0.43, 0.54) for the conditional false-negative probability. The maximum clade credibility tree is shown in Fig. 4.

Colorectal dataset 2

The colorectal cancer dataset (CRC0907 / K054) [35] contains 50 single-cell cancer samples and 25 adenoma polyp samples from a female colorectal patient, quality filtered

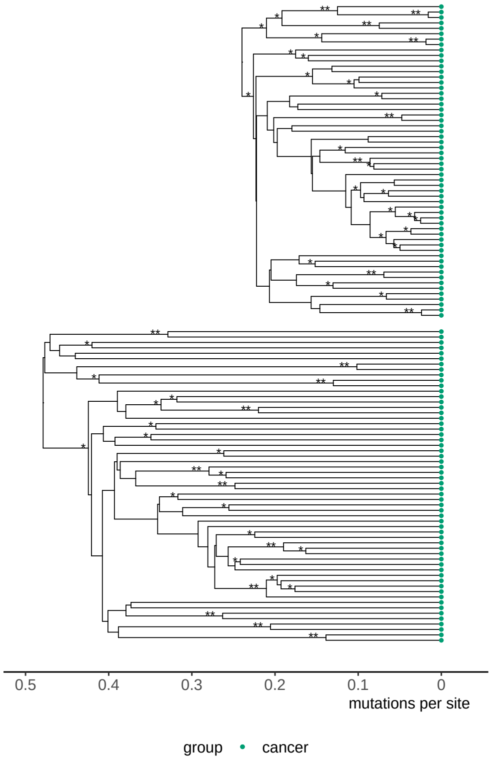

Fig 3. Summary of maximum clade credibility tree for the Myeloproliferative neoplasms dataset. Tree estimated under the binary error model (top), tree estimated without an error model (bottom). Clades with at least 10% support are annotated with an asterisk, clades with at least 50% support are annotated with two asterisks.

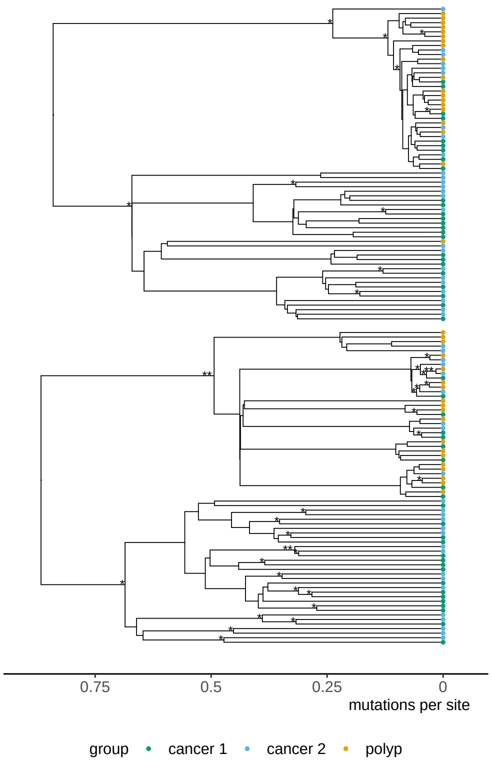

Fig 4. Summary of maximum clade credibility tree for Colorectal dataset 1. Tree estimated under the binary error model (top), tree estimated without an error model (bottom). Clades with at least 10% support are annotated with an asterisk, clades with at least 50% support are annotated with two asterisks.

by [35] to 452 sites. The authors identify two cancer subclones (1CA and 2CA) and one adenoma polyp clone (XR). Similarly to Colorectal dataset 1 above, the authors also claim monoclonal evolution occurred in this female patient [35]. The HPD interval for the tree height is (0.56, 0.93) substitutions per site under the error model, and (0.54, 0.85) when errors are ignored. The HPD interval for the tree length is (5.86, 9.36) under the error model and (18.44, 21.02) when errors are ignored. The error estimates are $(1.60 \times 10^{-6}, 2.0 \times 10^{-2})$ for the conditional false-positive probability, and (0.58, 0.63) for the conditional false-negative probability. The maximum clade credibility tree is shown in Fig. 5.

Evaluation on simulated datasets

We validated our implementation by simulating directly from the model and attempting to recover the substitution, error and population dynamic parameters. We generated sequences of length 400 and ran 200 trials per setup, using the following simulation parameters: $\lambda \sim Lognormal(\mu = -1)$, $\alpha \sim Beta(1, 50)$, and $\beta \sim Beta(1, 50)$. Using a well-calibrated study, we quantify our parameter estimates using the 95% HPD coverage and the tree space using the subtree prune and regraft (SPR) distance to the true tree. See Supplementary Materials for detailed results.

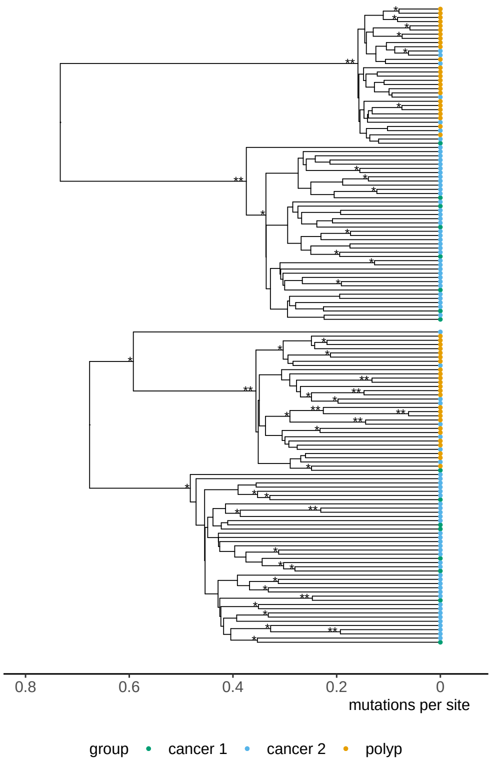

Fig 5. Summary of maximum clade credibility tree for Colorectal dataset 2. Tree estimated under the binary error model (top), tree estimated without an error model (bottom). Clades with at least 10% support are annotated with an asterisk, clades with at least 50% support are annotated with two asterisks.

Estimating model parameters on an ultrametric tree

We first examined the accuracy of parameter estimation on trees with data sampled at a single point in time, given the true tree. We use a fixed 10 leaf tree generated by a Yule process with the tree height set to 0.1. The tree height is set to allow for sufficient phylogenetic signal. Our experiments show that all parameters, error probabilities, and substitution rate, can be estimated accurately at the same time (Fig. S4). We find that not using an error model significantly biases the estimation of the substitution rate λ. Without an error model, the coverage of λ is 71% when data is simulated with errors compared to a baseline coverage of 94% where data is simulated without errors. Supplementary Figure S4 shows the coverage of λ on noisy data increases to 96% when an error model is incorporated.

Estimating model parameters on a time-sampled tree

Next, we examined the estimation on trees with time sampled data, assuming the true tree is known. We use a fixed 10 leaf tree generated from a birth-death process with tips sequentially sampled through time and tree height scaled to 0.1. In Supplementary Figure S5, we show that all the model parameters can also be accurately estimated in this scenario.

Joint estimation of model parameters and evolutionary tree

Lastly, we jointly estimate the model parameters, topology and timing of the evolutionary tree. For our data, we generated 200 Yule trees with 30 leaves, using the following parameters: $birth\ rate \sim Normal(\mu = 7.0, \sigma = 1.0)$, $\lambda \sim Lognormal(\mu = -1)$, $\alpha \sim Beta(1, 50)$, and $\beta \sim Beta(1, 50)$. We selected a Normal distribution on *birth rate* to generate trees with phylogenetic signal. We are able to accurately estimate the tree topology and model parameters simultaneously (Fig. S6 - S7). We also observed that the conditional false-positive probability α is more difficult to estimate than the conditional false-negative probability β.

Discussion

Errors can affect age estimates

Compared to models without error, incorporating an error model significantly reduces the estimated tree length. The evolutionary distance between samples is on average reduced by 50% for colorectal dataset 1, 61% for colorectal dataset 2, and 70% for the myeloproliferative dataset. This discrepancy in the distance is jointly explained by conditional false-positive and conditional false-negative errors in our model. A similar effect from ancestral age errors was discussed in [33], which shows that not accounting for uncertainty in fossil ages causes inaccuracies in topology and divergence time estimates.

Error model supports a recent rapid expansion

In Figure 4, the polyp group undergoes a recent rapid expansion (top) rather than parallel evolution followed by expansion when errors are not modeled (bottom). In Figure 5, the error model has a more recent expansion event for the polyp group (top) compared to the model without error (bottom). Our evolutionary reconstruction supports the placement of the polyp group as a sister group to the tumour groups identified by the authors [35]. We note that the lab sample labels do not necessarily reflect the natural *normal*, *tumour*, *polyp* groupings. For example, normal cells present within the tumour, such as T-cells, may be sampled and incorrectly labelled as tumour cells.

Effect of quality filtering

In our analysis, we used the full set of quality filtered sites by [35]. In contrast to [7, 35] which used a smaller manually curated subset of sites; there is no clear separation between the 1CA and 2CA samples for our models. Although we currently use filters proposed by the original authors, we believe using more data could improve the amount of phylogenetic signal, allowing more accurate tree reconstruction. The effect of filtering methods on phylogenetic reconstruction has been examined in previous work [31–33]. Evidence from [31, 32] show the choice of filters can affect the tree topology.

Interpretation of parameters

In genomic studies it is common practice to only report sites (or genes) that show genetic variation. Data is filtered to select sites which are mutated in at least one sample [34, 35]. As a consequence of this, the conditional false-positive probability estimated in our paper, $\hat{\alpha}$, is not equivalent to the per-base false-positive rate. Here, $\hat{\alpha}$ is the false-positive rate conditioned on observing a mutation in at least one sample, or an "effective false-positive rate". Similarly for the estimated false-negative rate $\hat{\beta}$. These biases need to be taken into consideration when specifying the priors. The substitution parameter, λ, represents the back-mutation rate relative to a forward-mutation rate of 1. The birth rate parameter of the Yule process represents the rate of cell division per mutation.

Limits to phylogenetic error modeling

From our simulation experiments, we find the conditional false-positive probability may be more difficult to recover than the conditional false-negative probability. We showed that we are able to estimate parameters in our Bayesian framework accurately, but we have not explored the theoretical limits of our approach. In addition to this, we observed that the tree space could be sparse for some datasets. We believe future work on tree summarization could enable a more detailed analysis of tree topology distributions.

Future directions

In our paper, we demonstrated that incorporating error parameters can affect the scale and evolutionary interpretation of single-cell cancer datasets. Modelling errors can have significant consequences for evolutionary interpretation; for example, our error model supports a recent expansion in colorectal polyp cells rather than parallel evolution which is supported by a standard model. Our error models are formulated within a Bayesian framework, allowing data filtering strategies to be accounted for during evolutionary inference. Hence, these parameter estimates reflect the sequencing technology and pre-processing steps used for different datasets, and their priors can be chosen based on such considerations.

Currently, the effect of filtering strategies on phylogenetic reconstruction is yet to be systematically explored. However, we believe less stringent filtering strategies, and ascertainment correction will provide better estimates of important parameters including mutation rate and divergence times, while more accurately accounting for uncertainty. In addition to this, translating advances in oncology and cell biology knowledge into model priors is also a crucial step to building biologically informative models. Lastly, methods to effectively summarize and visualize diffuse tree distributions will enable results from single-cell phylogenetics to be examined in more detail.

Materials and Methods

We model evolution as a continuous-time Markov chain (CTMC) on a timed tree τ where the rate of base substitutions are defined by a substitution matrix Q. The error model on tip sequences is parametrized by the conditional false-positive probability α and conditional false-negative probability β.

Binary substitution model

Our binary model has two states $S = \{1, 0\}$, representing the presence or absence of mutation in a diploid cell. The substitution matrix Q has a single parameter λ the back-mutation rate, and a forward-mutation rate of 1. The substitution matrix is

$$Q = \begin{array}{c|cc} & 0 & 1 \\ \hline 0 & -1 & 1 \\ 1 & \lambda & -\lambda \end{array} \tag{1}$$

with equilibrium frequencies

$$\pi_0 = \lambda/(\lambda + 1) \tag{2}$$
$$\pi_1 = 1/(\lambda + 1) \tag{3}$$

Ternary substitution model

We implement the ternary substitution model from [7,37] with genotypes

$$\begin{aligned} 0 &= (0\,/\,0) \text{ or } (-\,/\,0), \\ 1 &= (0\,/\,1), \\ 2 &= (1\,/\,1) \text{ or } (-\,/\,1), \end{aligned}$$

where and 0 denotes the normal allele, 1 denotes a mutated allele and - denotes the loss of an allele. The substitution matrix Q has two parameters, λ_l the rate of loss of heterozygosity (LOH), and λ_d the rate of back-mutations.

$$Q = \begin{array}{cc} & \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \end{array} & \begin{pmatrix} -1 & 1 & 0 \\ \frac{\lambda_d+\lambda_l}{2} & -\lambda_d - \lambda_l & \frac{\lambda_d+\lambda_l}{2} \\ 0 & \lambda_d & -\lambda_d \end{pmatrix} \end{array} \tag{4}$$

with equilibrium frequencies

$$\pi_0 = 1/x \tag{5}$$
$$\pi_1 = 2/(x \cdot (\lambda_d + \lambda_l)) \tag{6}$$
$$\pi_2 = 1/(x \cdot \lambda_d), \tag{7}$$

where $x = 1 + 2/(\lambda_d + \lambda_l) + (1/\lambda_d)$. We also derive a modified version of the rate matrix to explicitly account for back-mutations. Our modified rate matrix, Q is

$$Q = \begin{array}{cc} & \begin{array}{ccc} 0 & 1 & 2 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \end{array} & \begin{pmatrix} -1 & 1 & 0 \\ \frac{1}{6} + \lambda_l & -\frac{2}{3} - 2\lambda_l & \frac{1}{2} + \lambda_l \\ 0 & \frac{1}{3} & -\frac{1}{3} \end{pmatrix} \end{array} \tag{8}$$

with equilibrium frequencies

$$\pi_0 = x/(x + y + 1) \tag{9}$$
$$\pi_1 = 1/(x + y + 1) \tag{10}$$
$$\pi_2 = y/(x + y + 1), \tag{11}$$

where $x = (1/6 + \lambda_l)$ and $y = 3 \cdot (1/2 + \lambda_l)$.

Error models

We follow the conditional error parameterization presented in [7,8,10]. For binary data, the conditional error probabilities are

$$\begin{array}{cc} & \text{Genotype (truth)} \\ & \begin{array}{cc} 0 & 1 \end{array} \\ \text{Data (obs)} \begin{array}{c} 0 \\ 1 \end{array} & \begin{pmatrix} 1 - \alpha & \beta \\ \alpha & 1 - \beta \end{pmatrix} \end{array} \tag{12}$$

For ternary data, the conditional error probabilities are

$$\begin{array}{ccccc} & & & \text{Genotype (truth)} & \\ & & 0 & 1 & 2 \\ \text{Data (obs)} & 0 & 1-\alpha-\frac{\alpha\beta}{2} & \frac{\beta}{2} & 0 \\ & 1 & \alpha & 1-\beta & 0 \\ & 2 & \frac{\alpha\beta}{2} & \frac{\beta}{2} & 1 \end{array} \tag{13}$$

Tree likelihood

The likelihood of observing an alignment D given a tree τ, substitution matrix Q and model parameters λ, α, β can be calculated using Felsenstein's peeling algorithm [38]

$$P(D|\tau, Q, \lambda, \alpha, \beta) = \prod_{i=1}^{n} P(D_i|\tau, Q, \lambda, \alpha, \beta), \tag{14}$$

The likelihood of the data D_i at each site i in the alignment is calculated independently. Let S be the set of possible genotypes, e.g. $S = \{0, 1\}$ for binary data. The likelihood at the root x is

$$P(x|\tau, Q, \lambda, \alpha, \beta) = \sum_{s_0 \in S} P(x = s_0) \cdot L(x = s_0), \tag{15}$$

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

$$L(x = s_0) = \sum_{s_1 \in S} P_{s_0,\, s_1}(t_1) \cdot L(y_1 = s_1) \cdot \sum_{s_2 \in S} P_{s_0,\, s_2}(t_2) \cdot L(y_2 = s_2), \tag{16}$$

where t_1 is the time between x and y_1 on tree τ, t_2 is the time between x and y_2, and $P_{ij}(t) = e^{Qt}$ 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

$$L(x = 0 \mid \alpha,\ \beta) = (1 - \alpha,\ \beta), \tag{17}$$

$$L(x = 1 \mid \alpha,\ \beta) = (\alpha,\ 1 - \beta) \tag{18}$$

The partial likelihoods for leaf nodes using the ternary error model is given by

$$L(x = 0 \mid \alpha,\ \beta) = (1 - \alpha - \frac{\alpha\beta}{2},\ \frac{\beta}{2},\ 0), \tag{19}$$

$$L(x = 1 \mid \alpha,\ \beta) = (\alpha,\ 1 - \beta,\ 0), \tag{20}$$

$$L(x = 2 \mid \alpha,\ \beta) = (\frac{\alpha\beta}{2},\ \frac{\beta}{2},\ 1) \tag{21}$$

Code and data availability

All code and data will be made publicly available. Our BEAST2 package, Phylonco v.0.0.2 is available on GitHub `www.github.com/kche309/beast-phylonco`. The following R packages were used in this paper: ggtree [39], phangorn [40], treeSim [41], and treeSimGM [42].

Acknowledgements

We thank Prof. Bong-Gun Ju and Hongjin Wu for providing us with their colorectal datasets. We thank Dr. Alex Gavryushkin, Dr. Jiří C. Moravec, Dr. Remco Bouckaert, and Dr. Hamim Zafar for helpful discussion. We acknowledge Nectar Research Cloud for providing the computing resources necessary for this study. AJD was funded by the MBIE Endeavour Fund, ref. U00X1912.

References

1. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nature reviews Clinical oncology. 2018;15(2):81.
2. de Bruin EC, McGranahan N, Mitter R, Salm M, Wedge DC, Yates L, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science. 2014;346(6206):251–256.
3. Liu J, Dang H, Wang XW. The significance of intertumor and intratumor heterogeneity in liver cancer. Experimental & molecular medicine. 2018;50(1):e416–e416.
4. Kuipers J, Jahn K, Beerenwinkel N. Advances in understanding tumour evolution through single-cell sequencing. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer. 2017;1867(2):127–138.
5. Woodworth MB, Girskis KM, Walsh CA. Building a lineage from single cells: genetic techniques for cell lineage tracking. Nature Reviews Genetics. 2017;18(4):230.
6. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome biology. 2020;21(1):1–35.
7. Zafar H, Tzen A, Navin N, Chen K, Nakhleh L. SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. Genome biology. 2017;18(1):178.
8. Jahn K, Kuipers J, Beerenwinkel N. Tree inference for single-cell data. Genome biology. 2016;17(1):86.
9. El-Kebir M. SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. Bioinformatics. 2018;34(17):i671–i679.
10. Ross EM, Markowetz F. OncoNEM: inferring tumor evolution from single-cell sequencing data. Genome biology. 2016;17(1):1–14.
11. Meijer A, van der Sanden S, Snijders BE, Jaramillo-Gutierrez G, Bont L, van der Ent CK, et al. Emergence and epidemic occurrence of enterovirus 68 respiratory infections in The Netherlands in 2010. Virology. 2012;423(1):49–57.
12. Malmstrøm M, Matschiner M, Tørresen OK, Star B, Snipen LG, Hansen TF, et al. Evolution of the immune system influences speciation rates in teleost fishes. Nature Genetics. 2016;48(10):1204–1210.
13. Kearns AM, Restani M, Szabo I, Schrøder-Nielsen A, Kim JA, Richardson HM, et al. Genomic evidence of speciation reversal in ravens. Nature Communications. 2018;9(1):1–13.

14. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLOS Computational Biology. 2019;15(4):1–28. doi:10.1371/journal.pcbi.1006650.

15. Dawid AP. The well-calibrated Bayesian. Journal of the American Statistical Association. 1982;77(379):605–610.

16. Popic V, Salari R, Hajirasouliha I, Kashef-Haghighi D, West RB, Batzoglou S. Fast and scalable inference of multi-sample cancer lineages. Genome biology. 2015;16(1):91.

17. Miura S, Gomez K, Murillo O, Huuki LA, Vu T, Buturla T, et al. Predicting clone genotypes from tumor bulk sequencing of multiple samples. Bioinformatics. 2018;34(23):4017–4026. doi:10.1093/bioinformatics/bty469.

18. Jiang Y, Qiu Y, Minn AJ, Zhang NR. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proceedings of the National Academy of Sciences. 2016;113(37):E5528–E5537.

19. Heide T, Maurer A, Eipel M, Knoll K, Geelvink M, Veeck J, et al. Multiregion human bladder cancer sequencing reveals tumour evolution, bladder cancer phenotypes and implications for targeted therapy. The Journal of pathology. 2019;248(2):230–242.

20. Cooper CS, Eeles R, Wedge DC, Van Loo P, Gundem G, Alexandrov LB, 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. 2015;47(4):367–372.

21. Alves JM, Prado-Lopez S, Cameselle-Teijeiro JM, Posada D. Rapid evolution and biogeographic spread in a colorectal cancer. Nature communications. 2019;10(1):1–7.

22. Schwartz R, Schäffer AA. The evolution of tumour phylogenetics: principles and practice. Nature Reviews Genetics. 2017;18(4):213–229.

23. Gertz EM, Chowdhury SA, Lee WJ, Wangsa D, Heselmeyer-Haddad K, Ried T, et al. FISHtrees 3.0: tumor phylogenetics using a ploidy probe. PLoS One. 2016;11(6):e0158569.

24. Potter NE, Ermini L, Papaemmanuil E, Cazzaniga G, Vijayaraghavan G, Titley I, et al. Single-cell mutational profiling and clonal phylogeny in cancer. Genome research. 2013;23(12):2115–2125.

25. Pennington G, Smith CA, Shackney S, Schwartz R. Reconstructing tumor phylogenies from heterogeneous single-cell data. Journal of Bioinformatics and Computational Biology. 2007;5(02a):407–427.

26. Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. Systematic zoology. 1978;27(4):401–410.

27. Potter N, Miraki-Moud F, Ermini L, Titley I, Vijayaraghavan G, Papaemmanuil E, et al. Single cell analysis of clonal architecture in acute myeloid leukaemia. Leukemia. 2019;33(5):1113–1123.

28. Roerink SF, Sasaki N, Lee-Six H, Young MD, Alexandrov LB, Behjati S, et al. Intra-tumour diversification in colorectal cancer at the single-cell level. Nature. 2018;556(7702):457–462.

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

30. Zafar H, Navin N, Chen K, Nakhleh L. SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. Genome research. 2019;29(11):1847–1859.

31. Axelsson E, Willerslev E, Gilbert MTP, Nielsen R. The effect of ancient DNA damage on inferences of demographic histories. Molecular Biology and Evolution. 2008;25(10):2181–2187.

32. Rambaut A, Ho SY, Drummond AJ, Shapiro B. Accommodating the effect of ancient DNA damage on inferences of demographic histories. Molecular Biology and Evolution. 2009;26(2):245–248.

33. Barido-Sottani J, van Tiel NMA, Hopkins MJ, Wright DF, Stadler T, Warnock RCM. Ignoring Fossil Age Uncertainty Leads to Inaccurate Topology and Divergence Time Estimates in Time Calibrated Tree Inference. Frontiers in Ecology and Evolution. 2020;8:183. doi:10.3389/fevo.2020.00183.

34. Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, et al. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell. 2012;148(5):873–885.

35. Wu H, Zhang X, Hu Z, Hou Q, Zhang H, Li Y, et al. Evolution and heterogeneity of non-hereditary colorectal cancer revealed by single-cell exome sequencing. Oncogene. 2017;36(20):2857–2867.

36. Chen C, Xing D, Tan L, Li H, Zhou G, Huang L, et al. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). Science. 2017;356(6334):189–194.

37. Zafar H, Tzen A, Navin N, Chen K, Nakhleh L. Comments on the model parameters in “SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models”. Genome biology. 2019;20(1):95.

38. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution. 1981;17(6):368–376.

39. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution. 2017;8(1):28–36.

40. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27(4):592–593.

41. Stadler T. TreeSim: simulating phylogenetic trees. R package version. 2015;2(2).

42. Hagen O, Stadler T. TreeSimGM: Simulating phylogenetic trees under general Bellman–Harris models with lineage-specific shifts of speciation and extinction in R. Methods in ecology and evolution. 2018;9(3):754–760.