Sciphy: A Bayesian phylogenetic framework using sequential genetic lineage tracing data

Sophie Seidel[1,2*†], **Antoine Zwaans**[1,2†], **Samuel Regalado**[3], **Junhong Choi**[3,4], **Jay Shendure**[3,5,6,7,8], **Tanja Stadler**[1,2]

***For correspondence:** sophie.seidel@bsse.ethz.ch, tanja.stadler@bsse.ethz.ch

[†]These authors contributed equally to this work

[1]Department of Biosystems Science and Engineering, ETH Zürich, CH-4058 Basel, Switzerland; [2]Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland; [3]Department of Genome Sciences, University of Washington, Seattle, WA, USA; [4]Developmental Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA; [5]Howard Hughes Medical Institute, Seattle, WA, USA; [6]Seattle Hub for Synthetic Biology, Seattle, WA, USA; [7]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA; [8]Allen Discovery Center for Cell Lineage Tracing, Seattle, WA, USA

Abstract

CRISPR-based lineage tracing offers a promising avenue to decipher single cell lineage trees, especially in organisms that are challenging for microscopy. A recent advancement in this domain is lineage tracing based on sequential genome editing, which not only records genetic edits but also the order in which they occur. To capitalize on this enriched data, we introduce SciPhy, a simulation and inference tool integrated within the BEAST 2 framework. SciPhy utilizes a Bayesian phylogenetic approach to estimate time-scaled phylogenies and cell population parameters. After validating SciPhy using simulations, we apply it to lineage tracing data obtained from a monoclonal culture of HEK293T cells for which we estimate time-scaled trees together with cell proliferation rates. We compare SciPhy to the lineage reconstruction based on a widely used clustering method, UPGMA, and find that the UPGMA-reconstructed lineage trees differ from SciPhy trees in some key aspects of tree structure; in particular, SciPhy trees stand out for their later estimated cell division times. In addition, SciPhy reports uncertainty as well as proliferation rates, neither of which are available within a UPGMA analysis. This study showcases the application of advanced phylogenetic and phylodynamic tools to explore and quantify cell lineage trees, laying the groundwork for enhanced and confident analyses to decode the complexities of biological development in multicellular organisms. SciPhy's codebase is publicly available at https://github.com/azwaans/SciPhy.

Introduction

The interplay between cell division, cell death and differentiation is at the core of the development of all multicellular organisms. Tracing cell populations and their lineage relationships via DNA-based molecular recording has recently emerged as a powerful approach to study this complex process (***VanHorn and Morris*** (***2021***)). Amongst the wide range of technologies that have been developed to record such lineage information into the genome, approaches that rely on CRISPR-Cas9 have stood out for their ability to generate stochastic, heritable indels in genetic barcodes during development and proliferation (***McKenna et al.*** (***2016***); ***McKenna and Gagnon*** (***2019***); ***Chan et al.*** (***2019***); ***He et al.*** (***2022***); ***Bowling et al.*** (***2020***); ***Alemany et al.*** (***2018***); ***Spanjaard et al.*** (***2018***)). However, the reliance of early versions of this approach on *unordered indels* as characters, together with the short time window for the recording process, limit the reconstruction of highly resolved cell 'lineage trees' (see Glossary, Tab. 1) (***Loveless et al.*** (***2021***); ***Choi et al.*** (***2022***)).

Recent work (***Loveless et al.*** (***2021***), ***Choi et al.*** (***2022***)) addresses some of these limitations by introducing prime-editing based recording systems resulting in *ordered insertions* that enable lineage tracing over prolonged periods of time. These methods rely on sequential insertions of nucleotide sequences at target sites mediated by the prime editor and prime editing guide RNA (pegRNA) (***Anzalone et al.*** (***2019***)), where each insertion step at a target site has a twofold effect: (1) deactivation of that site and (2) activation of a following site for a subsequent editing step. This property not only ensures the irreversible accumulation of edits to a constant number of targets, but also provides a record of the order of all editing events, encoded in the relative direction of edits within the target. These new lineage tracing technologies have the potential to enable the reconstruction of accurate, highly resolved single cell lineage trees.

However, the reconstruction of lineage trees using such data to date ***Choi et al.*** (***2022***); ***Loveless et al.*** (***2021***) has relied on UPGMA (Unweighted Pair Group Method with Arithmetic Mean, ***Sokal et al.*** (***1963***)) in tandem with custom distance metrics. This approach is agnostic to specific properties of the editing process, such as variable propensities of introducing particular inserts and variable edit insertion rates (see Glossary, Tab. 1) . Further, UPGMA, being a clustering method based on pairwise distance metrics, only leverages pairwise distances between combinations of edits per cell, ignoring all higher order information (***Felsenstein*** (***2004***)). In contrast, Bayesian phylogenetic inference frameworks developed for unordered edit lineage tracing data (***Seidel and Stadler*** (***2022***)) enable the estimation of lineage trees with cell division times (see Glossary, Tab. 1) by relying on mechanistic models of barcode generation. This approach was shown to provide accurate lineage reconstruction, attributed to the ability to explicitly incorporate *a priori* knowledge of both the editing process as well as the experimental conditions under which traced cell populations were studied.

Phylogenetic Terms	**Cell Biology Terms**
Tree topology	Lineage tree
Clock rate	Edit insertion rate by prime editor (e.g. per hour)
Bifurcation, branching point	Cell division event
Branching time	Cell division time
Time-scaled tree	Lineage tree where branches are scaled in real time
Monophyletic	Monoclonal

Table 1. Glossary Translation of phylogenetic terminology to cell biology terminology.

In this article, we derive a time-dependent mechanistic model of sequential insertion based editing (ordered inserts) for lineage tracing data. We implement this model into a new inference frame-

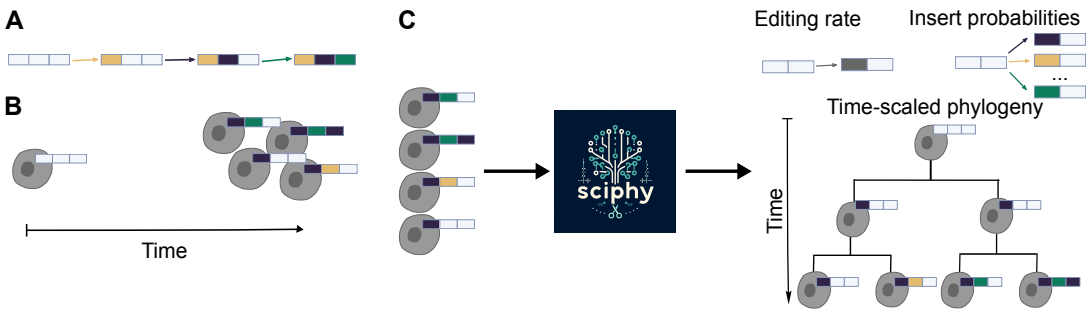


Figure 1. SciPhy is a framework for phylogenetic and phylodynamic analysis of sequential edit lineage tracing data. It implements a mechanistic model of sequential editing based recording systems (**A**), which enables their use for Bayesian phylodynamic inference in BEAST 2. These alignments are obtained following typical pipelines of genetic lineage tracing experiments: cells that harbour copies of the recording barcode accumulate insertions during growth (**B**). At the end of these experiments, single-cell sequencing generates barcode alignments (**C**), from which SciPhy can infer parameters of the recording system: the rate and probability at which different insertions are acquired.

work, coined 'SciPhy' **S**equential **C**as-9 **I**nsertion based **Phy**logenetics, made available in Bayesian inference software BEAST 2 (***Bouckaert et al.*** (***2014***)), which uses a Markov Chain Monte Carlo framework to jointly estimate time-scaled lineage trees (see Glossary, Tab. 1) and editing dynamics. Additionally, the integration of our model in BEAST 2 makes phylodynamic analyses of single cell population dynamics directly accessible to the community.

We showcase the versatility of our validated framework with a comprehensive analysis of a lineage tracing dataset derived from a monoclonal (see Glossary, Tab. 1) culture of HEK293T cells (***Choi et al.*** (***2022***)). Our analysis reveals that complex editing dynamics govern the recording process in the dataset and estimate the growth rates of the cell populations over time. We highlight significant differences in our lineage tree estimates compared with those obtained using the UPGMA clustering method. Together, our explorations of the dataset underline that the choice of lineage reconstruction method has crucial implications on both the cellular relationships and growth dynamics inferred using sequential lineage tracing data.

Results

We introduced a mechanistic editing model and likelihood calculation for prime-editing based lineage tracing data with sequential editing as the core component of SciPhy (Fig. 5). This model assumes that stably integrated barcodes accumulate edits sequentially and irreversibly at a constant rate, reflecting the Cas9 nicking rate. Subsequent pegRNA-mediated insertions at the nicked sites occur with varying probabilities. We derived a phylogenetic likelihood to calculate the probability of observing a specific barcode alignment given a phylogeny and SciPhy parameters. Details of the model, its parameterisation, and Bayesian inference implementation in BEAST 2 are in Methods and Materials.

In-silico validation

We validated our Bayesian method, SciPhy, using well-calibrated simulations (***Dawid*** (***1982***)), to ensure that it draws samples from the correct posterior distribution. This process involved simulating trees and a set of edited barcodes for each cell, and using them as input data to estimate the editing parameters under the SciPhy model. For each of the simulations, editing parameters were drawn from distributions chosen to be representative of a broad range of realistic editing

dynamics. These same distributions are in turn used as priors in the inference (see Methods for details).

We calculated the coverage for each editing model parameter, defined as the fraction of datasets for which SciPhy inferred a 95% highest posterior density (HPD) interval containing the true (simulation) parameter. The coverages for all parameters fall within the expected range (***Mendes et al.*** (***2024***)) (Supplementary table 5), demonstrating that our method is correctly implemented.

In addition, we plotted the simulated (true) parameters against the estimated parameters (Fig. 2) to investigate the inference power of our method across a range of parameter values. The correlation for the editing rate (Pearson's correlation coefficient on the median estimates R=0.99, 95 % confidence interval 0.995,0.998]) and the insertion probabilities (Pearson's correlation coefficient on the median estimates R=0.98, 95% confidence interval [0.978, 0.982]) confirms that SciPhy is able to extract meaningful signal from the data.

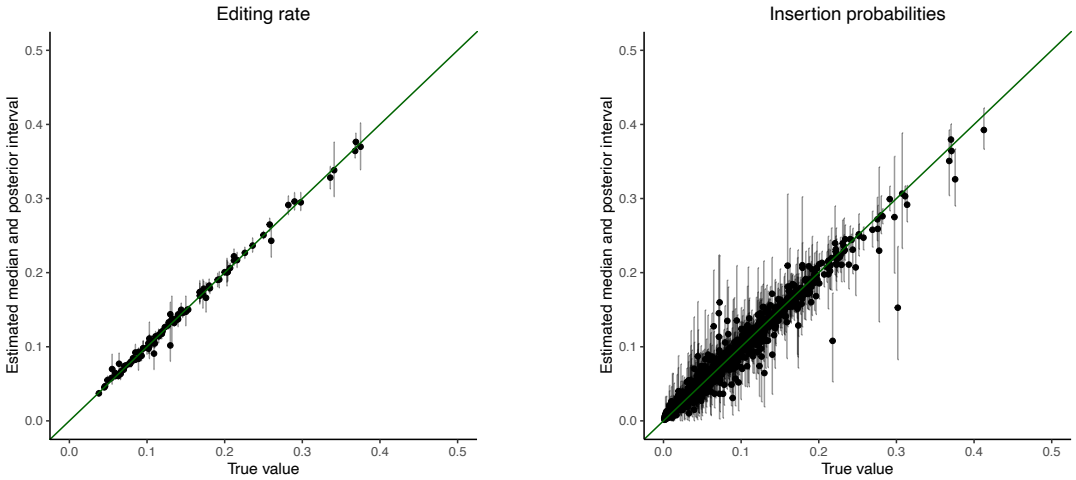


Figure 2. Validation results Here we report the results of the in-silico validation for the editing rate (left) and the insertion probabilities (right) based on 100 simulations. For each panel, we plot the estimated posterior medians against the true values as black dots. The grey intervals denote the 95% highest posterior density interval around the median. The green line represents the performance of an ideal estimator.

Phylodynamic analysis of monoclonal expansion

We analyze a dataset where DNA Typewriter was used to trace lineage in a 25-day monoclonal expansion of HEK293T cells (Fig. 3, A). Filtering of the data was performed as previously (***Choi et al.*** (***2022***), see Methods for details). Using SciPhy, we conducted a phylodynamic analysis to estimate the lineage tree, editing dynamics of the DNA Typewriter tapes as well as the growth dynamics of the sampled population of cells (see Methods for details).

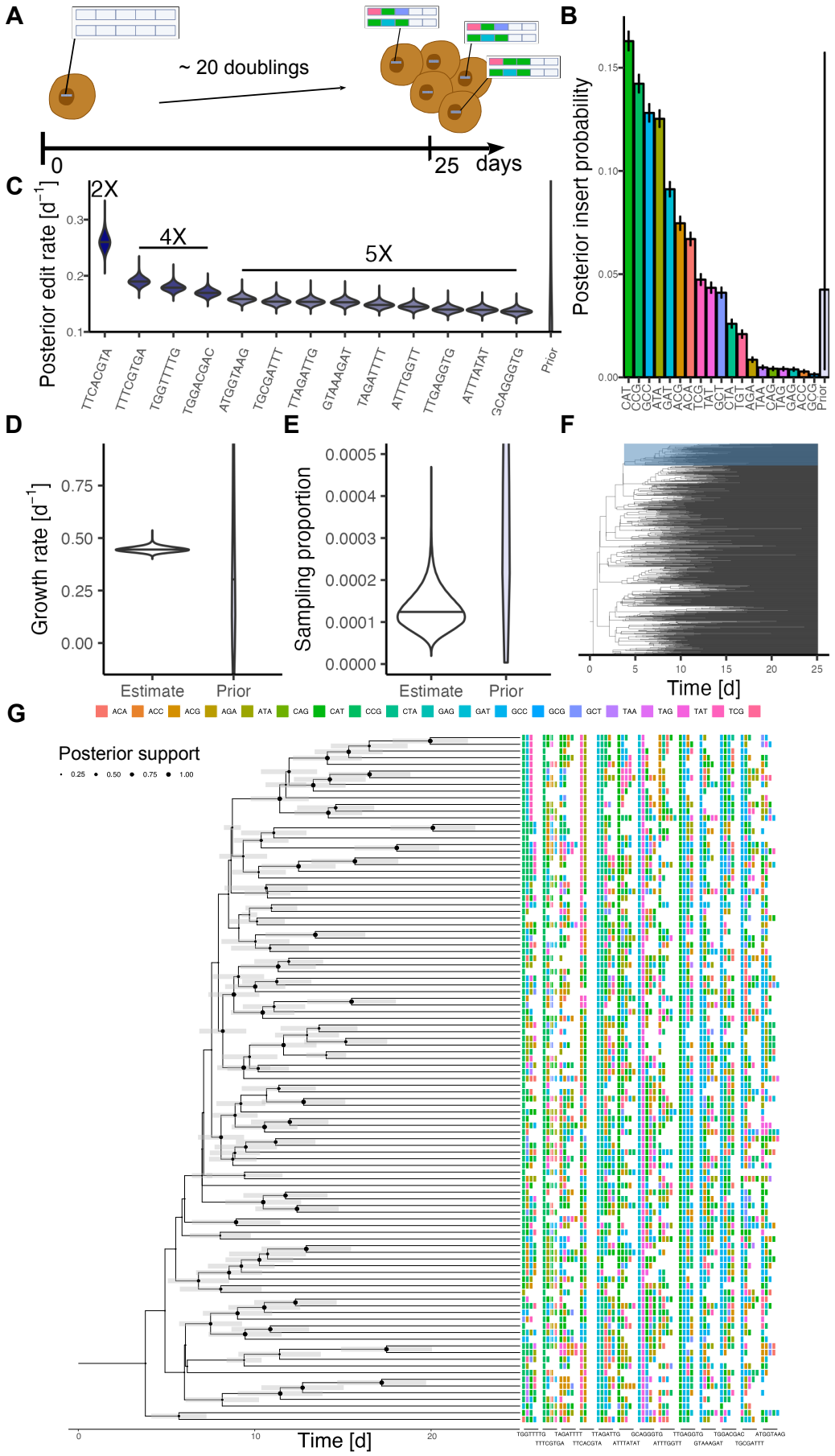

A
~ 20 doublings
0
25
days
B
Posterior insert probability
0.15
0.10
0.05
0.00
CAT
CCG
GCC
ATA
GAT
ACG
ACA
TCG
TAT
GCT
CTA
TGT
AGA
AAG
CAG
TAG
GAG
ACC
GCG
Prior
C
Posterior edit rate [d⁻¹]
2X
4X
5X
0.3
0.2
0.1
TTCACGTA
TTTCGTGA
TGGTTTTG
TGGACGAC
ATGGTAAG
TGCGATTT
TTAGATTG
GTAAAGAT
TAGATTTT
ATTTGGTT
TTGAGGTG
ATTTATAT
GCAGGGTG
Prior
D
Growth rate [d⁻¹]
0.75
0.50
0.25
0.00
Estimate
Prior
E
Sampling proportion
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
Estimate
Prior
F
0
5
10
15
20
25
Time [d]
G
ACA ACC ACG AGA ATA CAG CAT CCG CTA GAG GAT GCC GCG GCT TAA TAG TAT TCG
Posterior support
0.25 0.50 0.75 1.00
0
10
20
Time [d]
TGGTTTTG TAGATTTT TTAGATTG GCAGGGTG TTGAGGTG TGGACGAC ATGGTAAG
TTTCGTGA TTCACGTA ATTTATAT ATTTGGTT GTAAAGAT TGCGATTT

Figure 3. SciPhy captures complex editing dynamics and infers developmental parameters A We apply SciPhy to data from monoclonal expansion of HEK293T cells harboring the DNA Typewriter recording system. We report SciPhy's parameter estimates, namely the posterior distributions of the editing model parameters, namely the insertion probabilities (**B**) and the editing rates per day per tape (**C**). The error bars on the insertion probabilities indicate the 95% highest posterior density interval. Editing rates for truncated tapes are colored with darker shades and annotated with the corresponding number of sites (respectively 2, 4, and 5 sites). We report estimates of the cell population dynamics under the birth-death-sampling model, namely the growth rate (**D**) and the sampling proportion (**E**). In **F** we plot the time-scaled Maximum Clade Credibility (MCC) tree for 1000 cells and highlight a clade that is shown enlarged in **G**.We show a sub-clade of the MCC tree and highlight the posterior support in the node placement by the node thickness and the uncertainty, i.e. 95% HPD intervals, in the cell division time as grey shadings. We plot the alignment facing the tree tips.

We estimated the editing model parameters of SciPhy for this dataset, namely the overall insertion probabilities and an editing rate for each tape construct. Our estimates reveal that there is substantial variation in the insertion probabilities between different inserts (Fig. 3, B). Notably, the 'CAT' insert emerged as the most probable, with a probability of 16%, whereas the least probable insert, 'GCG', had a probability of less than 1%. Although this phenomenon, which could be due to differential expression of integrated pegRNAs or differential efficiencies for various hexamers, was documented in the initial publication (***Choi et al.*** (***2022***)), the UPGMA tree building method used there did not take it into account. Under the SciPhy model, we infer that when an insert is introduced, there is approximately a 50% chance that it will be one of the four most prevalent inserts (CAT, CCG, GCC, ATA). Consequently, these particular edits have a high chance of being introduced concurrently in otherwise independent lineages. In contrast with UPGMA, SciPhy can take this process into account by modelling these insert probabilities, potentially avoiding the reconstruction of false lineage relationships.

In addition to insertion probabilities, we also estimate an editing rate or clock rate for each target synthetically integrated into the genomic DNA, referred to as 'DNA Tape', or 'tape' for short, which denotes the rate at which inserts are introduced at target sites within that tape (Fig. 3, C). Our estimates reveal clock rates of around 0.15 d^{-1} for the majority of tapes. This rate corresponds to an average of 4 edits per tape over the entire experiment, a value that aligns with our dataset observations (see Appendix Fig. 1).

A point of interest is that all four of the truncated tapes, i.e. tapes with less than 5 target sites (see Methods for details), were found to possess elevated clock rates (Fig. 3, C, first 4 rates). Among these, the most truncated tape, 'TTCACGTA' —shortened to just two sites— had the highest rate. The observed pattern of editing rate ordering by tape length was unexpected under a null-model where all tapes have similar editing rates independent of tape length. We therefore asked whether early edits occur at a different rate from late edits. We compiled a dataset in which all tapes were shortened to length two and compared the estimated editing rates with those of the original data encompassing all sites (Appendix Fig. 2). Interestingly, the 2x tape "TTCACGTA" is edited the fastest in both analyses, indicating that its elevated editing rate is not due to the first two edits occurring faster across tapes of any length. We then compared the 4x and the 5x tapes. In the analysis using all sites, the three highest editing rates among the 4x and 5x tapes were all 4x tapes. This result is statistically significant under the assumption of equal rate ($p(k=3)=0.005$, see Methods for test details). However, in the analysis using only 2 sites, this order was lost, with only one 4x tape among the three tapes with fastest editing rates. This result was not statistically significant under the assumption of equal rates ($p(k=1)=0.5$). Therefore, given the editing rate estimates under SciPhy, the difference in editing rate between 4x tapes and 5x tapes may be due to a difference in editing speed across the last 2-3 sites. This difference could equally be caused by the fifth site

being edited at a slower rate or the third and fourth sites in the 4x truncated barcodes acquiring edits at a faster rate.

Along with estimating the editing model, we studied the growth dynamics of the HEK293T cell population by fitting a birth-death-sampling model to the set of sequences. These sequences are a subsample of 1000 cells from a population reported to consist of approximately $1.2 * 10^6$ cells ***Choi et al.*** (***2022***). To incorporate this knowledge, we initially fixed the sampling proportion to 0.0008, assuming a constant population growth throughout the experiment. Under this model, we estimated a cell population growth rate between 0.42 and 0.47 days^{-1} (95% highest posterior density (HPD) interval, see Appendix Fig. 10, C) "Sciphy + Fixed sampling"), corresponding to doubling times of 35 to 40 hours. This implies an average of $\approx 10^5 \pm 10^5$ (mean $\pm\sigma$) cells at the end of the experiment under a model of exponential growth, which is an order of magnitude smaller than the reported $1.2 * 10^6$ cells. After confirming that this estimate was driven by the data in tape alignments (see Appendix Sampling from prior), we hypothesized that this disagreement may stem from a model misspecification, rather than signal in the data. Therefore, we relaxed two of the main assumptions made in our model separately: first, the assumption of constant growth over the experimental period and, second, the assumption that we have accurate knowledge of the true sampling proportion.

First, we ran the same analysis with piecewise constant birth and death rates allowed to vary every two days over the timeline of the experiment. The estimated growth rates imply a pattern where the median growth rate is highest during the first 2 days of the experiment, with a median growth rate of 1.69 d^{-1} or doubling time of 9.6 hours (95% HPD of growth rate [0.57,2.76] d^{-1}), and declines for 5 days after which it stabilizes around 0.3 d^{-1} until the end of the experiment (see Appendix Fig. 9). Under this set of rates, the cell population grows to an expected size of $\approx 9 * 10^5$ cells, closer to the observed $1.2 * 10^6$ cells. However, the fast doubling time in the first days is unrealistic for HEK293T cell growth from a single cell. It should further be noted that a constant growth rate over the experimental period is not excluded under these results.

Second, we relaxed our assumption on accurate knowledge of the sampling proportion, by jointly estimating the sampling proportion along with the growth rate. In this analysis we inferred a growth rate between 0.49 and 0.58 days^{-1}, or a doubling time that ranges between 28 and 33 hours (95% HPD interval, (Fig.3, D). This doubling time is consistent with documentation for the HEK293T cell line from which these cells were derived ***Biodata*** (***2024***), and leads to an average population size contained between $2.5 * 10^5$ and $2.3 * 10^6$ (95% HPD) after 25 days, which agrees with the reported $1.2 * 10^6$ cells.

Interestingly, the inferred sampling proportion under the last model is approximately four to 16 times lower than initially assumed (95% HPD interval: $[5.2\times10^{-5}, 2.2\times10^{-4}]$, Fig. 3, E). Under the birth-death model parameterized for this analysis, a lower-than-expected sampling proportion may stem from measurement errors in estimating the total cell population size —potentially caused,e.g., by cell clumping during counting— or from non-uniform sampling, such as biases in cell selection, sequencing, or filtering. Given that the reported population size of 1.2×10^6 cells was measured using both a haemocytometer and a cell counter, which should provide reliable counts, it is more plausible that biases in the sampling process led to this reduced estimate. In such cases, the inferred sampling proportion likely reflects an 'effective' sampling proportion, accounting for these underlying biases rather than the true biological sampling proportion.

Taken together, the three analyses presented here suggest that a model of constant growth adequately describes the population dynamics under study.

Comparison to state-of-the-art

We next sought to compare the SciPhy posterior trees against current approaches that use UPGMA clustering which is a faster method but the reliance of its estimates is still unclear. To investigate whether both methods estimate similar ancestral relationships between cells, we compare the trees with respect to their topology. We employed the Robinson Foulds distance (***Robinson and Foulds*** (***1981***)) to compute pairwise tree distances and visualize them in 4 coordinates (Fig. 4A and Appendix Fig. 3, 4, and 5). Within the Robinson Foulds metric space, the UPGMA topology overlaps with SciPhy posterior tree estimates in all pairwise coordinates but the first two ((Appendix Fig. 3), implying that both methods reconstruct similar relationships between cells. Due to the Robinson Foulds metric's inherent constraints in distinguishing fine topological features, we proceeded to explore these patterns using the more discriminative Phylogenetic Information and Clustering Information distance metrics (***Smith*** (***2020***)). For the Phylogenetic Information distance metric, we find that the UPGMA tree falls separate from our inferred set of posterior tree topologies (Fig. 4 A,) in half of the pairwise coordinates (coordinates 1-2, 1-3, 1-4, Fig. 4 A and see Appendix Fig. 5). Consistently, using the Clustering Information distance metric, we find that the UPGMA tree falls clearly separate or close to the periphery of our set of inferred trees in two out of 6 pairwise coordinates (coordinates 1-2, 1-4, (Appendix Fig. 4)). Together, these results show that while overarching topological features are likely captured by both methods, SciPhy recovers distinct cellular relationships when looking at a finer resolution.

Next, we evaluated SciPhy, which is specifically designed to estimate time-scaled phylogenies, against UPGMA with respect to branch length inference. UPGMA produces trees with branches that are relative in length and not anchored to an absolute time scale. In contrast, SciPhy tree branches are calibrated in absolute time, as it incorporates the experiment's duration into its phylodynamic model, thus providing a temporal reference for the branch lengths. For fair comparison, we adapted the UPGMA tree to include temporal information by scaling it in three different ways. First, we scale its root (the most recent common ancestor (MRCA) of all cells, and representing the timing of the first division event) to 25 days to match the experiment duration (labelled "UPGMA"). However, it is likely that this first division event happened after the start of the experiment. Therefore, we also scale the UPGMA root to the median root height of SciPhy trees (labelled "UPGMA root"). Finally, we also estimate branch lengths using SciPhy on the fixed UPGMA topology (labelled "UPGMA + SciPhy").

Adopting a similar approach as before, but using the weighted Robinson Foulds distance that is sensitive to both tree topology and branch lengths, we observe that UPGMA trees diverge from trees in our posterior set in all pairwise coordinates (see Appendix Fig. 6). Given the topological similarities previously noted, this implies that UPGMA predicts branch lengths patterns distinct to those in trees from SciPhy's posterior set.

To better show how the branch lengths differ, we display the total number of lineages through time (LTT) for each reconstructed phylogeny (Fig. 4 B). Notably, SciPhy posterior trees display a slower increase in the number of lineages compared to UPGMA. For example, 500 lineages are reached after 5.5 days according to the UPGMA tree (5.7 days in the scaled UPGMA) but only after a median of 8.8 days (95% HPD interval of [8.0, 9.7]) days under SciPhy. Additionally, in the UPGMA tree, the final number of 1000 lineages is reached 1.1 (median, 95% HPD interval [0.31, 2.1]) days earlier compared to trees estimated by SciPhy. These distinctions result in the absence of overlap in LTT between the SciPhy tree posterior and the UPGMA tree for the majority of the experimental duration. Branch lengths are the key signal informing population dynamics in phylodynamic analyses. Observing the significant divergence in branch lengths estimated by UPGMA and SciPhy, we investigated the impact on growth rate estimates. We compared how the growth rate estimates

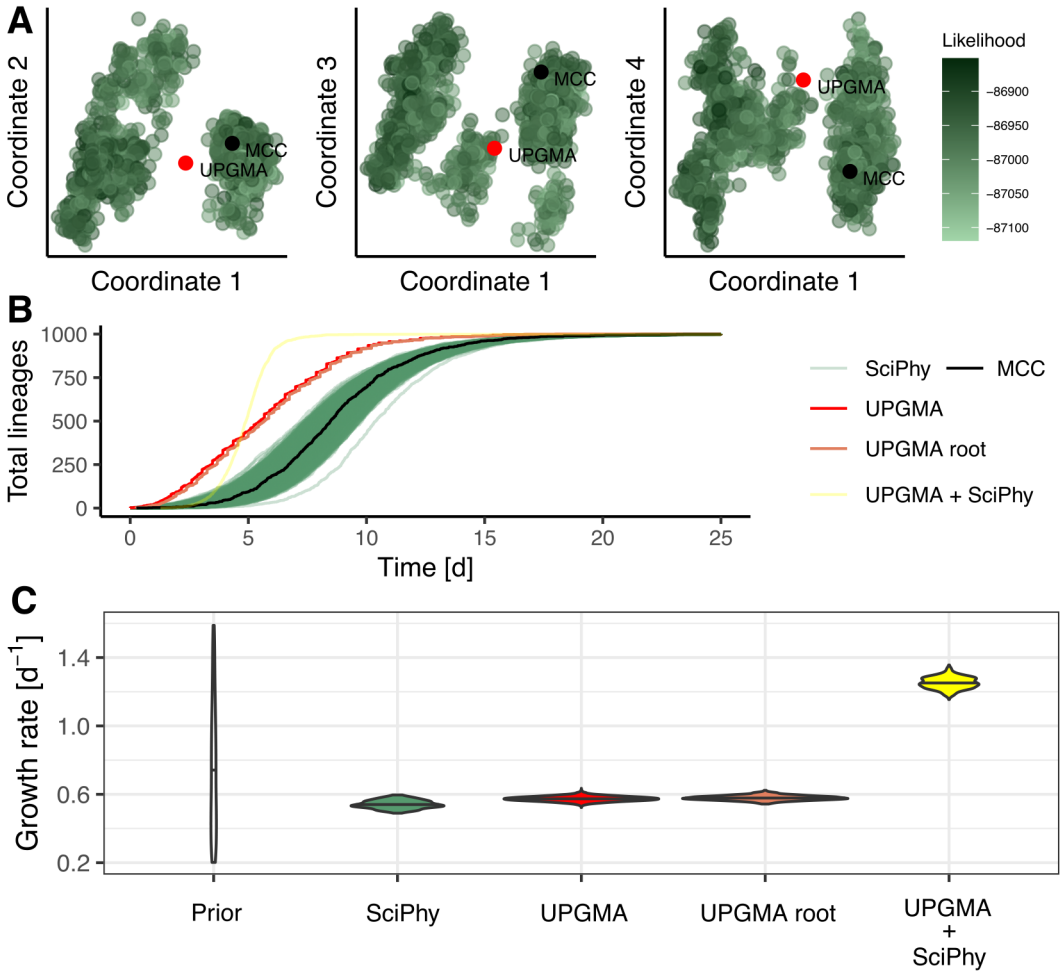


Figure 4. The SciPhy posterior trees differ from the UPGMA tree in terms of topology and branch lengths and result in different estimates of the growth dynamics. In **A**, we compare the SciPhy posterior tree set (green), including a summary tree, the MCC tree (black), to the UPGMA tree (red). We do so by calculating the Phylogenetic Information distance that captures topological distance between all pairs of trees. We use multidimensional scaling to project these distances into 4 dimensions. Here we show the results of the first coordinate against coordinates 2 to 4 from left to right. The remaining pairwise coordinate are shown in Appendix Fig. 4. In **B**), we plot the total number of lineages (y-axis) through time for each day (x-axis) for the SciPhy posterior tree set inferred under the constant birth-death-sampling model (green), the MCC trees, UPGMA (red), UPGMA with scaled root height, (orange) and UPGMA scaled using SciPhy (yellow). In **C** we compare the estimated growth rate(s) for each of these topologies.

differ based on the input trees used: (1) the posterior distribution of trees ("SciPhy") (2) the UPGMA tree scaled to the full experimental duration ("UPGMA"), (3) the UPGMA tree scaled to the median root height of SciPhy trees "UPGMA root") and (4) the UPGMA tree with SciPhy-estimated branch lengths ("UPGMA + SciPhy") ((Fig. 4 C). The median growth rate inferred on the fixed UPGMA trees (95% HPD interval of [0.54, 0.60] (2) , [0.55, 0.60] (3) and [1.17,1.30] (4)) is higher than the one inferred under SciPhy. Notably, the growth rate inferred on the UPGMA topology with branch lengths scaled with SciPhy is the highest, leading to doubling times between 18.4 and 20.5 hours, which excludes reported values for this cell line.

Overall, these results highlight that our modelling framework results in differences in the estimated tree topologies between UPGMA and SciPhy trees, although the extent of these disparities is challenging to quantify. There are notable differences in the timing of the branching events, which can lead to qualitative differences in the estimated growth rates.

Discussion

In this study, we introduced an editing model, edit sequence simulator and likelihood calculation tailored for lineage recordings based on sequential genome editing. We have packaged our model into 'SciPhy', which stands for **S**equential **C**as-9 **I**nsertion based **Phy**logenetics, and integrated it into the popular BEAST 2 Bayesian Phylogenetics framework. This integration empowers SciPhy users to specify complex editing dynamics, for example varying editing rates across DNA Tapes as done here or along tree branches (using e.g. relaxed clock models ***Drummond et al.*** (***2006***), ***Drummond and Suchard*** (***2010***)) which was not done here. Additionally, it facilitates the adoption of more nuanced models of cell development, allowing for cell type dependent division and differentiation rates using multitype models ***Vaughan et al.*** (***2014***); ***Kühnert et al.*** (***2016***) or permitting rate variations over time using piecewise-constant rates ***Drummond et al.*** (***2005***); ***Stadler et al.*** (***2013***). The BEAST 2 platform is increasingly attracting developers interested in contributing to single-cell related problems (***Seidel and Stadler*** (***2022***); ***Chen et al.*** (***2022***); ***Lewinsohn et al.*** (***2022***)) and we foresee that the support for single-cell related analyses will continue to grow in the future.

Our evaluations of SciPhy on experimental datasets demonstrate its ability to capture complex editing dynamics, such as the presence of preferential insertions and editing rate variation between tapes, patterns that are prevalent in both sequential lineage datasets we analysed and those from other lineage tracing technologies ***Chow et al.*** (***2021***); ***Seidel and Stadler*** (***2022***); ***McKenna et al.*** (***2016***); ***Feng et al.*** (***2021***). SciPhy, being a mechanistic editing model can directly account for these factors and even use the signal from tapes being edited at different rates as signal for branch length estimation.

We compared SciPhy to the phylogenetic reconstruction method used on both previously published analyses, UPGMA clustering. UPGMA clustering does not account for preferential insertions nor different edit rates for different tapes. Comparing SciPhy to one UPGMA approach (***Choi et al.*** (***2022***)), we find that some -but not all- aspects of the tree topology are captured in both SciPhy posterior trees and the UPGMA tree. Notably, the branch lengths are significantly different, which we showed impacts the downstream inference of the growth rate of the cell population. This is likely to also impact other parameters one may wish to estimate from cell phylogenies in the future, such as cell type differentiation rates. In the absence of a ground-truth lineage tree in this dataset, we cannot make conclusive statements about the correctness of either method. However, considering that SciPhy more accurately models the data generation process, we posit that it more effectively estimates the lineage tree, both with respect to the tree topology, cell division timings and downstream cell population estimates.

For the dataset analysed in our study, we used phylodynamic models to estimate the growth behaviour of cell populations based on SciPhy's derived posterior distribution of timed trees. Applied to the study of HEK293T cells, this approach showed that sequential lineage tracing data alone contains signal for the growth rate characteristic of this cell line. The use of SciPhy allowed to extracted relevant and previously inaccessible insights from the lineage tracing data.

While SciPhy offers promising avenues, it has some limitations. The size of tree space scales exponentially with the number of cells and Bayesian MCMC methods searching tree space do not converge for trees with more than around 1000 tips. Even for a dataset with 1000 cells, Bayesian phylogenetic inference requires evaluating the likelihood in the range of billions of iterations to reach convergence. Thus, despite the likelihood calculation scaling linear with the number of tips, speed-ups of the likelihood calculation can bring large benefits. We have achieved considerable improvements on the runtime of individual computations of the likelihood with caching and multi-thread parallelisation, though the empirical data analysis still required ∼15 days, and would benefit from a further reduction in run-time. For future improvements, we aim to further accelerate the likelihood calculation using software optimizations or approximations, and expedite convergence on large datasets by confining tree space.

Another limitation shared by most CRISPR-based lineage tracing data is that target or DNA Tape sequences are prone to dropout (in other words, not recovered in each cell), which SciPhy does not currently model. This dropout is a consequence of the technologies used to readout lineage recordings, such as single-cell RNA-seq, that result in loss of tape sequences for some cells. Current approaches often neglect cells with incomplete DNA Tape sets, resulting in data loss. Therefore, future work should integrate a dropout model accommodating tape absence such that data loss can be avoided.

Toward the goal of reconstructing more accurate lineage trees, many challenges remain. One of the pressing concerns is the refinement of experimental design to enhance the signal quality in lineage tracing data. Therefore, future studies could systematically compare the latest lineage tracing data generated with sequential genome editing against previous unordered methods, both with respect to the inference of the tree topology and the time estimates. As the field progresses, such comparisons could serve as essential benchmarks and guide experimental efforts.

Lineage recorders based on sequential genome editing present a notable advancement as they do not only include a number of edits but also the order in which they occur. With SciPhy, we present a model that directly captures and uses this detailed information which we expect to lead to improved tree reconstruction compared to commonly-used clustering tools. We see great promise that sequential lineage recorders which can co-track cellular signals ***Chen et al.*** (***2024***) will be developed in the future. In parallel, SciPhy can be expanded to inform about the timing of such recorded signalling events in addition to the single cell trees, leading to developmental insight in ever-greater detail.

Methods and Materials

Editing model for barcodes with ordered inserts

We present a model that captures the editing dynamics underlying the generation of sequential lineage tracing data using DNA recording systems such as the ones recently developed by ***Choi et al.*** (***2022***); ***Loveless et al.*** (***2021***). In these systems, cells carry stably integrated barcodes or "tapes", which are contiguous arrays of CRISPR-Cas9 target sites. Within these tapes, target sites

are sequentially made amenable to prime editing, resulting in the ordered accumulation of inserts. We consider an experimental setup starting with a single cell carrying several of these unedited tapes. As this cell and its offspring undergo proliferation, their tapes experience editing. At the end of the experiment, single cells are sampled and their tapes sequenced. Aligning sequenced tapes to their corresponding, unaltered tape reveals genetic modifications accumulated during the experiment. In the following, we detail the notations and assumptions we use to model this process.

Consider the following setup. Each cell carries k tapes that independently undergo editing. Each unedited tape is an array of N target sites, numbered from 1 to N. Each site can acquire a single insertion or edit, with M different inserts available.

Our model is defined as follows:

1. Initially, only a tape's first site is susceptible to editing. Subsequent sites are edited sequentially: editing at a site $i_{1<i\leq N}$ is only possible if its preceding site $i-1$ is already edited.
2. Editing is a two-step process :
   (a) Step 1: The Cas9 prime editor introduces cuts at the target sites in a time-dependent manner at rate r, called the editing rate or clock rate. This means editing follows a bounded Poisson process which stops once N edits have occurred.
   (b) Step 2: pegRNA mediated insertions following cut events are instantaneous. The set Γ of M possible insertions is predefined and denoted as:

$$\Gamma = \{\gamma_1, \gamma_2, ..., \gamma_M\}$$

   Given a cut event, the probability of adding insert γ_i is f_i (with $\eta := (f_1, f_2, ..., f_M)$), allowing for some insertions to have higher probabilities than others. We additionally assume that insertions operate independently - an insertion at a previous site does not influence the insertion probability at the currently active site.
3. Editing of any target site is irreversible.

From the model definition, if follows that the number of edits introduced in each independent tape is Poisson distributed. However, the distribution of the total number of edits per cell is not guaranteed to show features of a classic, unimodal Poisson distribution, because it is also influenced by the shared ancestry of the sampled cell population and its growth patterns (see Appendix Fig. 7 and 8).

Formally, for an array of length N, this model of editing is a continuous time Markov chain on state space Ω of all possible insertions and parameters r, η. These states are all possible finite sequences of length $n \leq N$, or n-tuples of ordered insertions, that we denote as:

$$(s_i)_{i\leq n}, s_i \in \Gamma,$$

where subscript i denotes the insertion site in the tape, i.e. its order. Note that the empty tuple $\emptyset$ represents the unedited state or starting state for all k tapes sampled per cell and the fully edited tape is an absorbing state of the Markov chain.

Let Y_t be the state of an array at time t after the start of the experiment. Given an array state $Y_t = a$ and its state after a time interval Δt, $Y_{t+\Delta t} = b$, we denote the tuple of inserts added to a to result in b as c. Specifically, c represents the elements in b that are not in a, which can be expressed as $c = b \setminus a$. The size of this tuple, denoted as $|c|$, corresponds to the number of inserts introduced between states a and b.

Transition probabilities

We denote the transition probability between states a and b under the model introduced above as $P_{a,b}(\Delta t; \eta, r) = P(Y_{t+\Delta t} = b; Y_t = a, \eta, r)$. In the following, we derive this transition probability. In step 1, we calculate the probability of introducing any $|c|$ new edits. In step 2, we calculate the probability of introducing the particular inserts in c.

Step 1 To derive the probability of introducing |c| edits, we distinguish two cases. In case 1, the number of new inserts |c| does not lead to saturation in the array ($|a| + |c| = |b| < N$). Then, the probability of |c| = j editing events happening in time Δt is given by the Poisson distribution (by model definition 2a):

$$P(|c| = j; r, \Delta t, |b| < N) = \frac{(r * \Delta t)^j}{j!} e^{-(r\Delta t)}$$

In case 2, introducing |c| inserts leads to saturation of the array ($|a| + |c| = |b| = N$), i.e reaching position N in the tape. We calculate the probability of this event by targeting its complement, namely the probability of any fewer inserts being introduced (again, by assumption 2a):

$$P(|c| = j; r, \Delta t, |b| = N) = 1 - P(|c| < j; r, \Delta t, |b| < N) = 1 - \sum_{k=0}^{j-1} P(|c| = k; r, \Delta t, |b| < N)$$

Combining both cases, the probability of introducing j edits in an array of length N assuming an editing rate r is:

$$P(|c| = j; r, \Delta t) = \begin{cases} \frac{(r*\Delta t)^j}{j!} e^{-(r\Delta t)} & \text{if } |b| < N \\ 1 - \sum_{k=0}^{j-1} \frac{(r*\Delta t)^k}{k!} e^{-(r\Delta t)} & \text{otherwise} \end{cases} \tag{1}$$

Step 2 Let us assume that the |c| insertions having occurred from a to b are $c = (s_i)_{i=|a|+1,\ldots,|a|+j}$, $s_i \in \Gamma$. The probability of inserting c conditional on |c| is the product of the individual insertion probabilities (by independence, see Assumption 2b):

$$P(c; |c| = j, \eta) = \prod_{k=|a|+1}^{|a|+j} P(s_k) = \prod_{k=|a|+1}^{|a|+j} \prod_{l=1}^{M} f_l^{\mathbb{1}(\gamma_l, s_k)} \tag{2}$$

Step 3 Combining the steps above, the transition probability between any two states a and b in time interval Δt given model parameters r and η, is obtained as follows (and as exemplified/described in Fig. 5):

$$\begin{aligned} P_{a,b}(\Delta t; \eta, r) &= P(Y_{t+\Delta t} = b; Y_t = a, \eta, r) = P(c; \Delta t, \eta, r) \\ &= P(|c|; \Delta t, r) \times P(c; |c|, \eta) \end{aligned} \tag{3}$$

where $P(|c|; \Delta t, r)$ was derived in Step 1 and $P(c; |c|, \eta)$ in Step 2.

Likelihood calculation

The editing model defined above can be used to calculate the likelihood of the model parameters (r, η) and the cell phylogeny T given the tape data D at the tips, $Lik(r, \eta, tree; D) = P(D; r, \eta, tree)$. Let T be a tree with n tips and $n - 1$ internal nodes. Tho oldest of these nodes is the root node,

being the most recent common ancestor of all sampled cells. Additionally, we introduce a stem branch preceding the root, connecting the root node to the start of the experiment, or origin node. We number the leaf nodes from $1, ..., n$ arbitrarily, and the internal nodes from $n+1, ..., 2n$ with increasing distance from the present. Each cell may have k tapes. Since we assume that tapes are edited independently from each other, we calculate the probability density of tape i given the tree and r, η separately for $i = 1, ..., k$. The likelihood is then the product over the k probability densities. In what follows, we thus consider one particular tape.

We first obtain the sets of possible tape states at each internal node given the data at the tip nodes. The likelihood of the tree is then calculated by summing over all possible state transitions between the nodes. This summation is performed through dynamic programming, going from the leaf nodes toward the origin of the tree; the employed particular algorithm is Felsenstein's pruning algorithm ***Felsenstein*** (***1973***).

Hence, we first determine all possible states at internal nodes. We restate that editing is irreversible and ordered (model definitions 1, 3). Therefore, for any sequence at a leaf node, ancestors can only harbor either the same or fewer edits. For any leaf node m, let υ^m be its tape with elements $\upsilon^m = (s_i^m)_{|\upsilon^m|}$ where $|\upsilon^m|$ is the number of edits in υ^m as defined in the editing model section. The set of its possible ancestor tape states A_m is the set of all subsets of edits from υ^m , akin to a power set on ordered tuples:

$$A_m = \{\emptyset, (s_1), (s_1, s_2), .., (s_1, s_2, .., s_{|\upsilon^m|})\}.$$

For an internal node i of tree T, it follows that the set of possible states A_i is the intersection of possible states of its two children nodes k and j, A_k and A_j (note that for the leaves, we initialize with all possible ancestor states). This is a result of editing being irreversible (assumption 3): an edit present in one child node but not the other cannot have been present in their parent node. Conversely, an edit present in both children nodes, can either have been present in the parent node or have been acquired by both children independently, therefore it should be included in the set of possible states. Consequently, the possible state set for node i, A_i is obtained as the intersection of the possible states of its children's k, j, A_k and A_j, recursively up the tree:

$$A_i = A_k \cap A_j$$

noting that the leaves are initialize with all possible ancestor states.

Given the possible states at the internal nodes, we calculate the probability density $P(D|T, r, \eta, A_{2n})$ of the tape data given the tree and model parameters following notations in (***Seidel and Stadler*** (***2022***)). To match assumption 1, we assume the unedited state at the origin node $A_{2n} = \{\emptyset)\}$.

$$Lik(T, \eta|D) = P(D|T, \eta, A_{2n}) \tag{4}$$

The transition probabilities from every parent node π_i to its child node i along their connecting branch with length τ_i are calculated using Eq. (3). Remember that A_i is the set of all possible states at node i. Any specific state within A_i is denoted as a_i. To account for all possible transitions, we sum the probabilities over all possible states at each internal node of the tree.

$$= \sum_{a_{2n-1} \in A_{2n-1}} \cdots \sum_{a_{n+1} \in A_{n+1}} \prod_{i=1}^{2n-1} P_{a_{\pi_i}, a_i}(\tau_i, \eta) \tag{5}$$

Using Felsenstein's pruning algorithm, we 'move the summations down the tree as far as possible' (***Felsenstein*** (***2004***)), meaning we perform the summations recursively up the tree, saving computations:

$$= \prod_{i=1}^{2n-1} \sum_{a_{2n-1} \in A_{2n-1}} P_{a_{\pi_i}, a_i}^{\mathbb{1}(\pi_i, a_{2n-1})}(\tau_i, \eta) \times \cdots \times \sum_{a_{n+1} \in A_{n+1}} P_{a_{\pi_i}, a_i}^{\mathbb{1}(\pi_i, a_{n+1})}(\tau_i, \eta) \tag{6}$$

As stated above, this calculation of $P(D|T, r, \eta, A_{2n})$ is performed for each tape. Multiplying the resulting probability densities for each tape leads the targeted likelihood function.

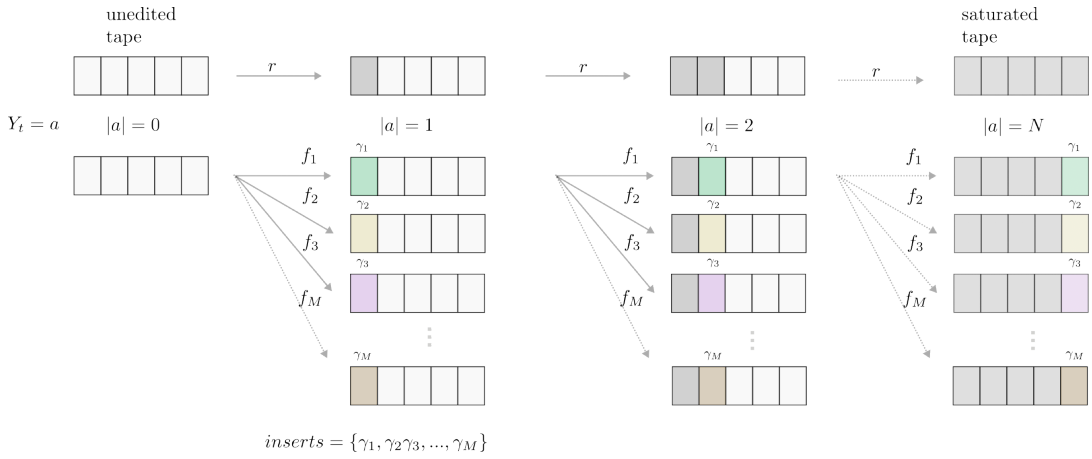
For a set of k tapes, where every tape is of length N on a tree with n tips, we perform n-1 operations to calculate the set of ancestral states at every internal node. Performing the intersection at an internal node scales with the size of the set, N. Hence, computing the ancestral states scales in O(knN).

Additionally, for each of the k tapes, we calculate the transition probabilities at every one of the n-1 internal nodes. At each node, we consider a maximum of (N+1)x(N+1) potential state transitions. Hence, our likelihood computations without optimization scales in O(knN^2) (In current technologies N is relatively small, $N \approx 5$). Note that the computation over the different k tapes is conveniently an 'embarrassingly parallel' problem.

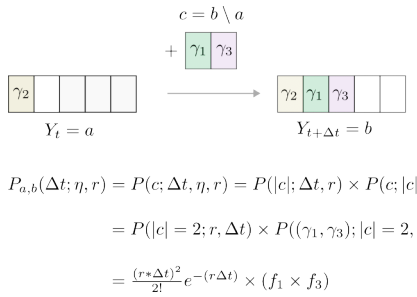
Overall, the calculation is polynomial in the size of the input, $O(knN + knN^2) = O(knN^2)$. For a fixed N and k, the runtime scales linearly with the number of tips in the tree, meaning the number of cells in the dataset.

Implementation

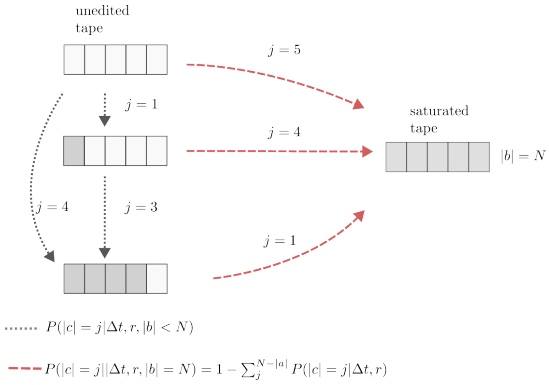
We implement the above editing model alongside an edit sequence simulator based on the feast package ***Vaughan*** (***2024b***) and a likelihood calculation in JAVA as a package for the software BEAST 2. BEAST 2 performs phylogenetic inference by using Markov Chain Monte Carlo (MCMC) to quantify the posterior distribution of the lineage tree and model parameters given a set of tape sequence alignments. During the MCMC, tree operators modify the tree to explore all possible topologies that could explain the data. As some of these changes are local, the likelihood for subtrees that are not modified remains unchanged. We make use of that property by implementing caching, where the likelihoods for subtrees are saved at each internal node to further speed up calculations (3.5-fold speed up on a personal computer running OSX on Intel(R) Core(TM) i5 Dual-Core @2.30Ghz). Additional speed-up is achieved by computing the likelihood for each tape alignment in parallel, which is easily achieved with existing components in BEAST 2 that support threading. These software optimizations lead to an overall 8-fold speed up relative to the naive implementation for an analysis involving 1000 sequences using 6 computing cores on the "Euler" scientific computing cluster .



(a) General model setup



(b) Toy example of transition probability calculation



(c) Probability of occurrence of insert events

Figure 5. General description of model components used to calculate transition probabilities in SciPhy. An overview of SciPhy's editing model for a tape with N=5 target sites and its parameters is shown in a). Target sites within tapes (rectangles in the tape) acquire inserts (colored boxes, inserts γ_1 to γ_M) with probability f_1 to f_M following Cas9 mediated nicks at rate r. Upon editing these sites are inactivated (grey boxes). b) is a toy example for the calculation of the transition probability for a tape acquiring 2 edits (γ_1 and γ_3). In c) we showcase different scenarios for the calculation of the probability of a given number of inserts to be added into a tape, distinguishing between cases where a tape gets saturated with inserts (red arrows pointing to the right) or not (grey arrows pointing down).

Validation

We employed well-calibrated simulations (***Dawid*** (***1982***)) to ensure the correct implementation of our Bayesian method SciPhy. We further used this simulation setup to assess the amount of information in the simulated data regarding parameters of our model.

We simulated 100 phylogenetic trees under a birth-death sampling model (***Stadler et al.*** (***2013***)) with parameters provided in Tab. 2. As tree inference is the primary functionality offered by the BEAST 2 platform (***Bouckaert et al.*** (***2014***)), we decided not to test tree inference here but instead use the fixed simulated trees for inference analyses.

Parameter	Value
Birth rate	0.8
Death rate	0.2
Sampling proportion	0.003
Origin or experiment duration	25

Table 2. Parameters used for simulating a phylogenetic tree under a birth-death sampling model.

We simulated alignments of tapes along each of the 100 trees. Firstly, we drew the rate of edit introduction (or clock rate) and edit probabilities to generate 13 distinct edits from a defined distribution (Tab. 3). Then we simulated the editing process of 10 independent tapes along each phylogenetic tree under these parameters. The resulting alignments of tapes were then used as input data for SciPhy to infer the rate of edit introduction and edit probabilities, using the same distributions (i.e. Tab. 3) as a prior. Finally, we compared the 100 simulated (true) parameters with the estimated parameters.

Parameter name	Symbol	Distribution	HDI
Clock rate	r	Log-normal ($\mu = -2, \sigma = 0.5$)	[0.03, 0.3]
Edit probabilities	f_k	Dirichlet ($\alpha = 1.5$)	[2e-4, 2e-1]

Table 3. Distributions for the editing model parameters used in the validation.

We chose the distributions over the editing model parameters to be representative for a broad range of editing dynamics. The clock rate distribution was selected to result in between 1 to 8 insertions for the experiment duration. We used a Dirichlet distribution with a concentration parameter of 1.5 for drawing the edit probabilities. This allows most edits to occur at similar, low probabilities while a few edits could occur more frequently.

For each MCMC inference, SciPhy was run for $5 * 10^7$ iterations to estimate the clock rate and edit probabilities, keeping all other parameters fixed to the truth (i.e. the value of the simulation). We used an ESS value of 200 as threshold to ensure convergence of the MCMC chains. We then discarded the inital 10% of the iterations as burn-in. Lastly, we calculated the coverage as the proportion of times the simulated (true) value fell within the estimated 95% HPD (define) interval.

Phylodynamic analysis of HEK293T expansion

We use SciPhy for the analysis of the publicly available dataset (i.e. taken from https://github.com/shendurelab/DNATickerTape, Supplementary_File_2_DataTableMOI19.csv) generated as described in (***Choi et al.*** (***2022***)). This study uses genetic tapes that contain 5 target sites for insertion, where each insertion step consists of the introduction of a trinucleotide character. We process the tape sequences as follows. We first retain all 3257 cells containing all 13 most frequent tapes for subsequent analysis, as done before. Out of these 13 tapes, 3 (barcoded with 'TTCCGTCA','TGGTTTTG' and 'TTCACGTA') are found to possess less than the 5 original editable target sites, respectively 4 ('TTCCGTCA','TGGTTTTG') and 2 ('TTCACGTA') sites. We refer to these tapes as "truncated". This procedure yields a dataset containing 19 different possible trinucleotide insertions. We create 3 random subsets of this data, each consisting of 1000 cells and perform phylodynamic analysis on each subset.

We use SciPhy to model the editing dynamics in this filtered dataset. We assume that each of the 13 tapes can accumulate edits at an independent rate, while insertion probabilities are shared across all tapes. We place prior distributions on the parameters of the editing model as described in table 4. As no quantitative information on the editing dynamics is available for this experimental setup, we use broad priors that capture a variety of experimentally plausible values (as in the validation part).

Being sampled from a monoclonal culture, cells are expected to undergo exponential growth and to not acquire any specific population structure. Consequently, we choose to infer the dynamics underlying the growth of this cell population using the birth-death sampling model for constant and piecewise constant rates (***Stadler et al.*** (***2013***)) . We set priors on the birth and death rate (4) that are very broad; leading for instance to between 12 and $2.5 * 10^{30}$ cells at the time of sequencing. We additionally input known experimental parameters by setting the origin time to the experiment duration of 25 days and fixing the sampling proportion to $1000/1.2e6 = 0.0008$. In the piecewise constant analysis, rates were allowed to vary every 2 days, yielding a timeline of 13 epochs where an Ornstein-Uhlenbeck (OU) prior is applied to the birth and death rates, and where we apply a Log-normal(0.5, 0) hyperprior on parameters σ and ν of the OU process.

Parameter	Prior distribution	95% HDI
Edit probabilities	Dirichlet (1.5)	[0.0038, 0.15]
Clock rate	Log-normal (-2, 0.5)	[0.06, 0.4]
Birth rate	Log-normal(-0.6, 1)	[0.1, 2.8]
Death rate	Log-normal(-2, 1)	[0.06, 0.4]
Sampling proportion	0.0008	fixed
Sampling proportion	Log-normal(-7.4, 1.2)	[0.00011,0.0045]
Origin or experiment duration	25	fixed

Table 4. Priors used for analysis of the HEK293T dataset unsing SciPhy and the birth-death sampling model.

For each subset of the data, we run 3 MCMC chains for $\sim 10^9$ iterations. We assess convergence by confirming that the three chains are sampling the same stationary distribution, with an ESS value per chain > 200 and visually in Tracer (***Bouckaert et al.*** (***2014***)). We discard 10% burn-in and combine all 3 chains using LogCombiner (***Bouckaert et al.*** (***2014***)) .

To test whether the ordering of 4x and 5x tapes is unexpected under the assumption of equal editing rates for all tapes, we compute the probability of the observed (i.e. that k of the three tapes with the fastest editing rates are 4x tapes) using the hypergeometric distribution. We model the

setup as follows: We have a total of 12 tapes (N=12), of which 3 tapes (K=3) have 4 sites and 9 tapes have 5 sites. If we draw the three tapes with the highest rates (n=3), we want to calculate $p(k)$, the probability that k of these tapes are of type 4x. We do this using the standard hypergeometric distribution:

$$p(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{7}$$

The number of cells over time, conditional on the estimated trees and phylodynamic parameters, was estimated using a stochastic mapping algorithm ***Freyman and Höhna*** (***2019***) implemented in the BDMM-Prime package ***Vaughan*** (***2024a***). Tree and log files were subsampled to retain every 10^7th state using LogCombiner, resulting in 350 states from which the cell population size was estimated.

To compare the UPGMA tree to SciPhy's estimated trees, we use the R package TreeDist (***Smith*** (***2022***)) to provide mappings of tree sets for four dimensions, pairwise. For this, we extract a random subsample of 650 trees inferred under SciPhy to calculate distances to the UPGMA topology. We do so using three metrics for topological comparisons, namely the Robinson Foulds distance, the Phylogenetc Information distance, and the Clustering Information distance. Further we employ one metric that also incorporates differences in branch lengths, namely the weighted Robinson Foulds metric. For each metric, the distances are mapped in lower dimensions using multidimensional scaling (***Gower*** (***1966***)). For each of these mappings we choose the number of dimensions required to faithfully visualize tree distances based on the product of the trustworthiness and continuity metrics as described by ***Smith*** (***2022***). For both mappings we find that 4 dimensions provide acceptable levels of distortion (trustworthiness x continuity > 8.5).

Code availability

All code to reproduce the analyses is publicly available from https://github.com/seidels/typewriter_analysis.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used the ChatGPT language model by OpenAI to refine the manuscript's language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement no. 101001077 (to T.S.) and the Paul G. Allen Family Foundation (to J.S.). The authors would like to thank François Bienvenu, Timothy G. Vaughan, Louis A. Du Plessis, Cecilia Valenzuela Agüi, Marcus Overwater, Nicola Mulberry and Ugnė Stolz for helpful discussions and feedback.

Authors' contributions

S.S.: conceptualisation, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft, writing—review and editing, visualisation, project administration. A.Z.: conceptualisation, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft, writing—review and editing, visualisation S.R.: data curation, writing—review and editing. J.C.: data curation, writing—review and editing. J.S.: writing—review and editing, supervision, funding acquisition. T.S.: conceptualisation, methodology, resources, writing—review and editing, supervision, funding acquisition.

Competing interests

The University of Washington has filed a patent application based on the DNA Typewriter system for sequential genome editing, in which J.C. and J.S. are listed as inventors. J.S. is on the scientific advisory board, a consultant and/or a cofounder of Adaptive Biotechnologies, Cajal Neuroscience, Camp4 Therapeutics, Guardant Health, Maze Therapeutics, Pacific Biosciences, Phase Genomics, Prime Medicine, Scale Biosciences, Somite Therapeutics and Sixth Street Capital. The remaining authors declare no competing interests.

References

Alemany A, Florescu M, Baron CS, Peterson-Maduro J, Van Oudenaarden A. Whole-organism clone tracing using single-cell sequencing. Nature. 2018; 556(7699):108–112.

Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, Chen PJ, Wilson C, Newby GA, Raguram A, Liu DR. Search-and-replace genome editing without double-strand breaks or donor DNA. Nature 2019 576:7785. 2019 oct; 576(7785):149–157.

Biodata S, HEK293T. cellosaurus; 2024. https://www.cellosaurus.org/CVCL_0063.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS computational biology. 2014; 10(4):e1003537.

Bowling S, Sritharan D, Osorio FG, Nguyen M, Cheung P, Rodriguez-Fraticelli A, Patel S, Yuan WC, Fujiwara Y, Li BE, et al. An engineered CRISPR-Cas9 mouse line for simultaneous readout of lineage histories and gene expression profiles in single cells. Cell. 2020; 181(6):1410–1422.

Chan MM, Smith ZD, Grosswendt S, Kretzmer H, Norman TM, Adamson B, Jost M, Quinn JJ, Yang D, Jones MG, et al. Molecular recording of mammalian embryogenesis. Nature. 2019; 570(7759):77–82.

Chen K, Moravec JC, Gavryushkin A, Welch D, Drummond AJ. Accounting for errors in data improves timing in single-cell cancer evolution. bioRxiv. 2022 may; p. 2021.03.17.435906.

Chen W, Choi J, Li X, Nathans JF, Martin B, Yang W, Hamazaki N, Qiu C, Lalanne JB, Regalado S, et al. Symbolic recording of signalling and cis-regulatory element activity to DNA. Nature. 2024; p. 1–9.

Choi J, Chen W, Minkina A, Chardon FM, Suiter CC, Regalado SG, Domcke S, Hamazaki N, Lee C, Martin B, et al. A time-resolved, multi-symbol molecular recorder via sequential genome editing. Nature. 2022; 608(7921):98–107.

Chow KHK, Budde MW, Granados AA, Cabrera M, Yoon S, Cho S, Huang TH, Koulena N, Frieda KL, Cai L, et al. Imaging cell lineage with a synthetic digital recording system. Science. 2021; 372(6538):eabb3099.

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

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. Molecular Biology and Evolution. 2005 may; 22(5):1185–1192.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS biology. 2006; 4(5):e88.

Drummond AJ, Suchard MA. Bayesian random local clocks, or one rate to rule them all. BMC biology. 2010; 8(1):1–12.

Felsenstein J. Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. Systematic Biology. 1973 sep; 22(3):240–249.

Felsenstein J. Inferring phylogenies. . 2004; p. 664.

Feng J, DEWITT III WS, McKenna A, Simon N, Willis AD, MATSEN IV FA. Estimation of cell lineage trees by maximum-likelihood phylogenetics. The annals of applied statistics. 2021; 15(1):343.

Freyman WA, Höhna S. Stochastic Character Mapping of State-Dependent Diversification Reveals the Tempo of Evolutionary Decline in Self-Compatible Onagraceae Lineages. Systematic Biology. 2019 may; 68(3):505–519. https://dx.doi.org/10.1093/sysbio/syy078, doi: 10.1093/SYSBIO/SYY078.

Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika. 1966; 53(3-4):325–338.

He Z, Maynard A, Jain A, Gerber T, Petri R, Lin HC, Santel M, Ly K, Dupré JS, Sidow L, et al. Lineage recording in human cerebral organoids. Nature methods. 2022; 19(1):90–99.

Kühnert D, Stadler T, Vaughan TG, Drummond AJ. Phylodynamics with Migration: A Computational Framework to Quantify Population Structure from Genomic Data. Molecular Biology and Evolution. 2016; 33(8):2102–2116.

Lewinsohn MA, Bedford T, Müller NF, Feder AF. State-dependent evolutionary models reveal modes of solid tumor growth. bioRxiv. 2022 aug; p. 2022.08.05.502978.

Loveless TB, Carlson CK, Hu VJ, Dentzel Helmy CA, Liang G, Ficht M, Singhai A, Liu CC. Molecular recording of sequential cellular events into DNA. bioRxiv. 2021; p. 2021–11.

McKenna A, Findlay GM, Gagnon JA, Horwitz MS, Schier AF, Shendure J. Whole-organism lineage tracing by combinatorial and cumulative genome editing. Science. 2016; 353(6298):aaf7907.

McKenna A, Gagnon JA. Recording development with single cell dynamic lineage tracing. Development (Cambridge). 2019 jun; 146(12).

Mendes FK, Bouckaert R, Carvalho LM, Drummond AJ. How to validate a Bayesian evolutionary model. bioRxiv. 2024 2; p. 2024.02.11.579856. https://www.biorxiv.org/content/10.1101/2024.02.11.579856v1https://www.biorxiv.org/content/10.1101/2024.02.11.579856v1.abstract, doi: 10.1101/2024.02.11.579856.

Robinson DF, Foulds LR. Comparison of phylogenetic trees. Mathematical biosciences. 1981; 53(1-2):131–147.

Seidel S, Stadler T. TiDeTree: a Bayesian phylogenetic framework to estimate single-cell trees and population dynamic parameters from genetic lineage tracing data. Proceedings of the Royal Society B. 2022; 289(1986):20221844.

Smith MR. Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. Bioinformatics. 2020; 36(20):5007–5013.

Smith MR. Robust analysis of phylogenetic tree space. Systematic Biology. 2022; 71(5):1255–1270.

Sokal RR, Sneath PH, et al. Principles of numerical taxonomy. Principles of numerical taxonomy. 1963; .

Spanjaard B, Hu B, Mitic N, Olivares-Chauvet P, Janjuha S, Ninov N, Junker JP. Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. Nature biotechnology. 2018; 36(5):469–473.

Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proceedings of the National Academy of Sciences. 2013; 110(1):228–233.

VanHorn S, Morris SA. Next-Generation Lineage Tracing and Fate Mapping to Interrogate Development. Developmental Cell. 2021 jan; 56(1):7–21.

Vaughan T, BDMM-Prime Package. GitHub; 2024. https://github.com/tgvaughan/BDMM-Prime.

Vaughan T, Feast Package. GitHub; 2024. https://github.com/tgvaughan/feast.git.

Vaughan TG, Kühnert D, Popinga A, Welch D, Drummond AJ. Efficient Bayesian inference under the structured coalescent. Bioinformatics. 2014; 30(16):2272–2279.