Comparing Phylogenetic Approaches to Reconstructing Cell Lineage from Microsatellites with Missing Data

Anne-Marie Lyne, Leïla Perié

▶ **To cite this version:**

Anne-Marie Lyne, Leïla Perié. Comparing Phylogenetic Approaches to Reconstructing Cell Lineage from Microsatellites with Missing Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, In press, 10.1109/TCBB.2020.2992813 . hal-02869464v2

HAL Id: hal-02869464
https://hal.science/hal-02869464v2

Submitted on 7 May 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL
archives-ouvertes.fr

Comparing Phylogenetic Approaches to Reconstructing Cell Lineage from Microsatellites with Missing Data

Anne-Marie Lyne, Leïla Perié

▶ **To cite this version:**

Anne-Marie Lyne, Leïla Perié. Comparing Phylogenetic Approaches to Reconstructing Cell Lineage from Microsatellites with Missing Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, Institute of Electrical and Electronics Engineers, 2020, 10.1109/TCBB.2020.2992813 . hal-02869464

HAL Id: hal-02869464
https://hal.archives-ouvertes.fr/hal-02869464

Submitted on 16 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Comparing Phylogenetic Approaches to Reconstructing Cell Lineage from Microsatellites with Missing Data

Anne-Marie Lyne, Leïla Perié

Abstract—Due to the imperfect fidelity of DNA replication, somatic cells acquire DNA mutations at each division which record their lineage history. Microsatellites, tandem repeats of DNA nucleotide motifs, mutate more frequently than other genomic regions and by observing microsatellite lengths in single cells and implementing suitable inference procedures, the cell lineage tree of an organism can be reconstructed. Due to recent advances in single cell Next Generation Sequencing (NGS) and the phylogenetic methods used to infer lineage trees, this work investigates which computational approaches best exploit the lineage information found in single cell NGS data. We simulated trees representing cell division with mutating microsatellites, and tested a range of available phylogenetic algorithms to reconstruct cell lineage. We found that distance-based approaches are fast and accurate with fully observed data. However, Maximum Parsimony and the computationally intensive probabilistic methods are more robust to missing data and therefore better suited to reconstructing cell lineage from NGS datasets. We also investigated how robust reconstruction algorithms are to different tree topologies and mutation generation models. Our results show that the flexibility of Maximum Parsimony and the probabilistic approaches mean they can be adapted to allow good reconstruction across a range of biologically relevant scenarios.

Index Terms—Cell lineage, phylogenetic reconstruction, microsatellites, mutation, single cell

1 Introduction

The ability to reconstruct cell lineage relationships on a single-cell basis is central to developmental biology, to the understanding of differentiation processes such as hematopoiesis, and underpins efforts to decipher the pathology of many diseases including cancer.

Until recently, almost all fate mapping experiments proceeded via prospective labeling ([1]) i.e., by the incorporation of heritable tags in progenitor cells that are later detected in their cell progeny. With the advent of next generation sequencing technologies, new methods with single cell resolution and significantly higher throughput have been developed. For example, sparse retroviral transduction of a library of DNA oligonucleotides can be used to infect progenitor cells with unique and heritable tags, detected in downstream cells via sequencing (e.g. [2], [3], [4], [5]). The most recently developed techniques avoid *ex vivo* viral infection by using *in vivo* genetic recombination, which can be induced in a time- and tissue-specific manner, to label progenitor cells. For example, CRISPR-Cas9 genome-editing, the translocation of a transposon element or the recombination of a series of *loxP* sites, can be used to mark individual progenitor cells with unique DNA signatures which are inherited by cell progeny (e.g. [6], [7], [8], [9], [10], [11]).

These techniques have revisited and revised the differentiation hierarchy in hematopoiesis ([4], [10], [12]) and have enabled semi-automated cell lineage reconstruction on a whole-organism scale ([6], [8]). However, they all rely on genetic modification to mark progenitor cells, and are therefore not readily transferable to humans. The only notable exception is the mapping of integration sites in gene-therapy treated patients ([13]), but this is restricted to transplantable tissues in specific types of patient.

A more widely applicable method for reconstructing lineage relationships is to use retrospective lineage tracing. This class of methods reconstructs lineage *a posteriori* from information present in cells without any prospective cell labelling. It uses naturally occurring mutations, which occur during DNA replication in cell division. If frequent enough, these mutations accumulate during divisions and ideally distinguish the cell progeny within and across lineages. They then act as an intrinsic tag that reflects the familial history of cells. Reconstruction methods, usually phylogenetic algorithms, are then applied to the information of the intrinsic tags to reconstruct lineage relationships between cells.

A number of different types of mutation can be used for these purposes: single nucleotide variants, copy number variation, retrotransposons and microsatellites (e.g. [14], [15], [16], [17]). In this paper, we choose to concentrate on microsatellites for the following reasons: 1. They have the highest mutation rate ($\sim 10^{-5}$ per locus per division ([18])). 2. They are highly abundant in the human and mouse genomes (more than 1 million loci ([19])). 3. Mutations in microsatellites are largely functionally neutral as variable loci tend to be found in non-coding regions ([20]). Microsatellites are generally defined to be tandem repeats of 1-6 nucleotide motifs of at least 12 base pairs in length e.g. CACACACACACA or $(CA)_6$ ([19]). A mutation in a

• *A.-M. Lyne and L. Perié are at Laboratoire Physico-Chimie Curie, Institut Curie, PSL Research University, CNRS UMR168, 75005 Paris, France E-mail: leila.perie@curie.fr*

Manuscript received December 11, 2019; revised .

microsatellite corresponds to the insertion or deletion of one or more repeat units, e.g. $(CA)_6$ becomes $(CA)_5$ or $(CA)_7$, due to misalignment of the repetitive strands during DNA replication. By observing the number of repeats at dozens of loci, microsatellites have been used to reconstruct cell lineage in the study of stem cell dynamics in the mouse colon ([21]), relapse mechanisms in acute leukaemia ([22]), a mouse lymphoma tumour ([23]), the female germline ([24]), mouse fibroblasts ([25]) and mouse cells from various organs ([26]).

The reconstruction method, as well the number of divisions and the mutation rate, have an impact on the accuracy of the lineage relationship inference. Previous works have used both simulation studies and experimental data with known outcomes to compare the Neighbour Joining algorithm and Bayesian-based methods to reconstruct cell lineage using microsatellites ([25], [27]). However, significant improvement has been made in Bayesian-based phylogenetic software since these early works ([28], [29], [30]), and there is therefore a need to revisit which method is most appropriate, incorporating more realistic models of microsatellite mutations and state-of-the-art phylogenetic inference methods.

For tree reconstruction, sequencing of single cells is required, as the information of the status of multiple loci for each cell is necessary. Single cell data contains a larger amount of missing data than bulk data. Missing data arises because of allele drop-out during the preparation of the sample (whole genome amplification, subsequent PCR amplification) or simply due to a detection limit (sequencing or capillary detection). The impact of missing data on lineage reconstruction from multiple-loci microsatellite length data has largely not been evaluated. The only exception is the work of [31] which modelled allelic dropout in their investigation of the impact of sequencing quality and depth on the accuracy of lineage reconstruction. However, they only tested one reconstruction algorithm with two missing data scenarios, so the full impact on reconstruction accuracy and indeed on the choice of reconstruction method has not been investigated.

In this paper, we use simulation to produce trees similar to those likely to be observed in vivo, taking hematopoiesis as an example. We consider which metrics can be used to assess how well a tree has been reconstructed. We then carry out a comprehensive comparison of the available phylogenetic methods, using multiple mutation rates and numbers of microsatellites. Additionally, we investigate the impact of missing data on these reconstruction methods, the impact of trees with different structures and explore how uncertainty in the mutation model impacts reconstruction.

2 Results

2.1 Comparing reconstruction methods with complete data

We simulated trees representing cell division using the Gillespie algorithm. Each simulation starts with one cell, and at each reaction cells divide or die with a given probability, with these probabilities constant within trees, but changed across trees. The simulated trees have around 150 living cells at the final time point and an average tree depth of 10 divisions. The initial cell is assigned a number of microsatellites with a distribution of repeat numbers, and at each reaction, cells have a fixed probability to mutate, according to a symmetric multistep mutation model ([19]) unless otherwise stated. After simulation of the microsatellite mutations, we add error in the form of PCR stutter to the living cells, as microsatellite loci are prone to length changes during amplification. When comparing reconstruction methods or scenarios, we compute the effect size as well as a plausible range for the effect size. Full simulation and comparison details are provided in the Methods section 4.

Throughout the paper, unless otherwise stated, results are presented for Tree 1, depicted in Fig. 1. Results for two other trees, Trees 2 and 3 (Supplementary Figs. 1 and 2), are given in Supplementary Figs. 5 to 8 and show very similar trends. The simulations presented below model microsatellites observed on a male X chromosome, i.e., where there is only one allele observed at each locus. We also present results in the Supplementary (as described in the Methods Section 4) for simulations modelling autosomal microsatellites, where in practice, it can be difficult to assign accurately lengths observed at a given locus to the maternal or paternal allele.

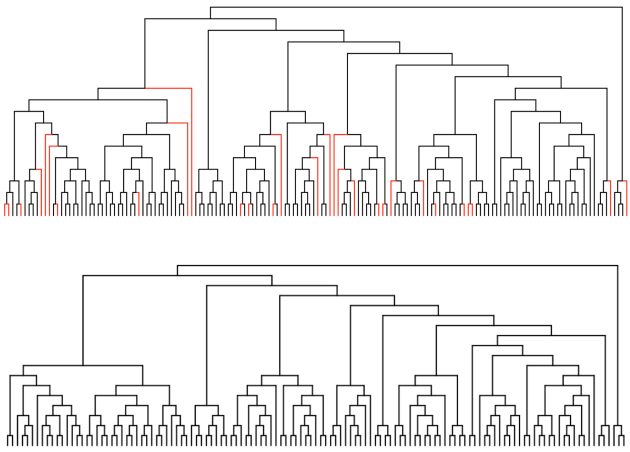

Fig. 1. **Simulated Tree 1.** Tree 1, with 125 living cells simulated by the Gillespie algorithm. At each reaction, cells have a 0.8 probability of dividing and a 0.2 probability of dying (in this realisation, 0.19 reactions were deaths). The top panel shows all cells produced, including dead cells in red. The bottom panel shows only cells which are alive at the final time point. The tree is plotted with nodes spaced evenly between the root and the final cells for clarity, although during the simulation the reactions occurred at random times as is usual with the Gillespie algorithm.

The mutation rate of microsatellites has been estimated to be between 10^{-3} and 10^{-6} per locus per division depending on the locus ([18]). [25] found, however, that the key parameter to predict the accuracy of the reconstruction is the expected number of mutations per cell division, i.e. that a lower mutation rate can be compensated for by observing a larger number of microsatellites. We observed a similar result in our simulations (Supplementary Figs. 3 and 4) where, for example, the reconstruction of a tree with 100 microsatellites and a mutation rate of 10^{-3}, is of a similar accuracy to that with 1000 microsatellites and a mutation rate of 10^{-4}. As lower mutation rates mean that more microsatellites need to be observed in order to reconstruct

the tree accurately and this results in longer computation time, particularly for the Bayesian inference, we choose to use a mutation rate of 10^{-3} to enable repeat simulations of the more computationally intensive reconstruction methods. We also simulate a maximum of 5000 microsatellites for the same reason.

We compared two distance-based methods, Neighbour Joining (NJ) and Balanced Minimum Evolution (BME), Maximum Parsimony and two probabilistic approaches, Maximum Likelihood and Bayesian inference, for phylogeny reconstruction (details in the Methods section). In the Bayesian inference, the prior distribution used for the tree is a constant-rate birth-death process which matches how the trees were simulated. To quantify how well the simulated tree was reconstructed, we defined a similarity score in which the percentage of bipartitions present in the original tree also present in the reconstructed tree is computed (described in more detail in Methods Section 4.4). If the two trees are identical, the score will be 100% as all of the bipartitions observed in the original tree will also be in the reconstructed tree.

For the two distance-based methods, Neighbour Joining ([32]) and Balanced Minimum Evolution ([33]), the L_1 (Manhattan), L_2 (Euclidean) and Cosine distances were compared. No difference was observed between the performance of Neighbour Joining and Balanced Minimum Evolution for a given distance and microsatellite number (all comparisons have <2% difference in sample mean and 0% in the plausible interval) with the exception of the L_2 distance (Fig. 2, and Supplementary Figs. 5 and 6). For the L_2 distance and 5000 microsatellites, BME performs $4.4 \pm 1.4\%$ better, however this is not informative for choosing the best method as L_2 is the worst performing distance at this setting. As there is otherwise no difference in performance between the two methods, Neighbour Joining is preferable as it is faster (see Supplementary Fig. 9). Looking just at Neighbour Joining and comparing the three distances, at 100 and 500 microsatellites the distances perform more or less the same with <2% difference in the sample means and a 0% in the plausible range for all comparisons. However, there are differences at higher microsatellite numbers, with L_1 performing $4.6 \pm 3.0\%$ and $1.8 \pm 3.1\%$ better than L_2 and Cosine respectively at 1000 microsatellites, and $7.3 \pm 1.2\%$ and $1.3 \pm 1.0\%$ better at 5000 microsatellites (Fig. 2).

Having, therefore, selected L_1 as the best overall distance for complete data, we next compared NJ to Maximum Parsimony and the probabilistic inference methods. In Fig. 3 (and Supplementary Figs 7 and 8), we see that with fully observed data, there is little difference in the performance of the methods (<2% difference in the sample means and 0% in the plausible range for all comparisons with a given microsatellite number). Distance-based methods such as Neighbour Joining are, therefore, the best choice due to their computational efficiency (as suggested previously by [27], and see run time comparison in Supplementary Fig. 9). Maximum Parsimony is also very computationally efficient, taking 10s or less to run depending on the number of microsatellites. For one tree reconstruction with a mutation rate of 0.001 and 5000 microsatellites, Bayesian inference takes almost three hours to run, whereas the distance-based approaches take less than one second. We also note that the

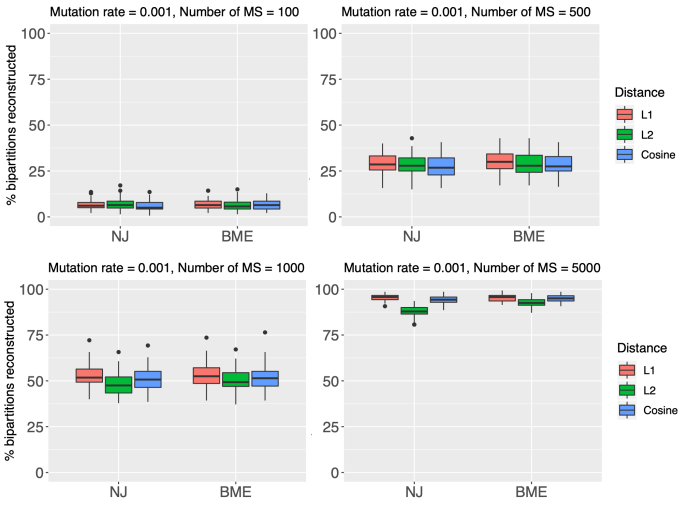

Fig. 2. **Comparison of distance-based methods, fully observed data.** Percentage of bipartitions in the true tree observed in reconstructed trees for distance-based methods, Neighbour Joining (NJ) and Balanced Minimum Evolution (BME), using L_1, L_2 and cosine distances. Boxplots are based on 60 independently simulated microsatellite mutation patterns on a fixed tree with a mutation rate of 0.001.

number of microsatellites observed is critical. Our simulations suggest that for a mutation rate of 10^{-3}, 1000s of loci need to be observed to reconstruct the tree accurately. This is higher than was previously suggested based on modelling mutations in polyguanine repeat sequences ([25]) due to the lower mutation rate use in this study (a factor of 10 lower).

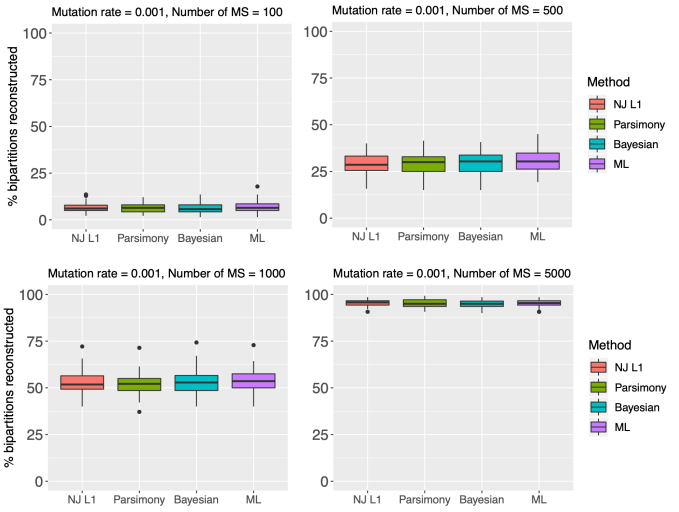

Fig. 3. **Comparison of phylogenetic methods, fully observed data.** Percentage of bipartitions from the true tree observed in reconstructed trees for Neighbour Joining (NJ), Maximum Parsimony, Maximum Likelihood and Bayesian methods with differing numbers of microsatellites. Boxplots are based on 60 independently simulated microsatellite mutation patterns on a fixed tree with a mutation rate of 0.001.

2.2 Comparing reconstruction methods with missing data

We now introduce missing data into our simulations to mimic single cell analysis of microsatellite loci. Mutations

were simulated as before and then 10, 20 and 30% of the data was assigned uniformly at random as missing data, in line with estimates for missing data suggested by [24] and [34]. For the Neighbour Joining and Balanced Minimum Evolution methods, we tested several options for computing distances when values are missing. We either impute the missing values using the mean or the modal value for each microsatellite locus or compute distances using only the microsatellites jointly observed in a given pair of cells. For the latter approach, we tested the effect of rescaling the distance to correct for the total number of loci included in the distance.

Again, Neighbour Joining and Balanced Minimum Evolution methods perform similarly for a given distance and microsatellite number, with Balanced Minimum Evolution performing slightly better (Supplementary Figs. 10-12). Fig. 4 shows the reconstruction accuracy of BME with full data and with different missing data strategies for 10, 20 and 30% missing data. The picture is complex, and depends on the amount of missing data, the number of microsatellites and the missing data strategy. At low microsatellite numbers, imputing yields more accurate reconstruction than using observed values (e.g. at 500 microsatellites, imputing the mean is $1.0 \pm 2.4\%$, $4.3 \pm 2.1\%$ and $4.6 \pm 1.5\%$ better than the best alternative using observed values, with 10, 20 and 30% missing data respectively). However, with 5000 microsatellites, L_1 distance using observed values outperforms imputation at all proportions of missing data with rescaling having little impact ($6.2 \pm 1.9\%$, $7.6 \pm 2.5\%$ and $6.5 \pm 2.8\%$ better than the best imputation approach, with 10, 20 and 30% missing data respectively). This can be explained by the lineage information redundancy when large numbers of microsatellites are observed; even without observing all of the data some parts of the reconstruction can be completed accurately as there is sufficient information in the data that is present. When using imputation, however, wrongly imputed values can cause incorrect reconstruction. As with complete data, increasing the number of microsatellite loci and hence the quality of the reconstruction makes differences between the L_1 and L_2 distances more apparent, with distance L_1 performing better (Fig. 4).

Next we compared the best distance-based approaches (Neighbour Joining and BME with unscaled L_1 distance using only observed microsatellite loci) with Maximum Parsimony, and the probabilistic approaches (full details in the Methods section). Both the Maximum Parsimony and likelihood methods have inbuilt ways of handling incomplete data which do not use imputation. Maximum Parsimony in PAUP* assigns the most parsimonious state to the missing locus given the cell's placement on the tree; therefore each cell's final position on the tree is determined only by the parsimony contribution of the observed loci ([35]). The likelihood methods, both implemented in RevBayes, integrate over all possible character states for missing 'leaf' characters, i.e. they are treated the same way as unknown internal node characters. These respective approaches do not discard any data and do not use any unobserved, imputed data. In Fig. 5 (and Supplementary Figs. 13 to 14) we see that as the proportion of missing data increases accuracy drops off more quickly for some approaches than others. For all microsatellite numbers, the drop in accuracy

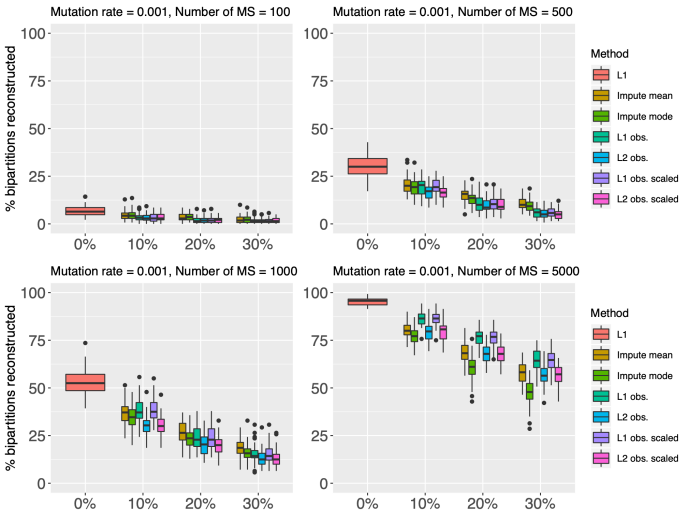

Fig. 4. **Comparison of BME with different distances, missing data.** Percentage of bipartitions from the true tree observed in reconstructed trees for BME using a variety of distances to deal with the missing data: imputing missing values using the mean/mode of observed values, L_1 or L_2 distances using observed values, unscaled or scaled. Boxplots are based on 60 independently simulated mutation patterns with a mutation rate of 0.001.

for Maximum Parsimony and the likelihood approaches is less than the drop for distance-based methods; at 1000 and 5000 microsatellites, the average reconstruction accuracy of Maximum Parsimony and the likelihood methods with 30% missing data is higher than the average accuracy of the distance methods with 20% missing data (Fig. 5). Use of either Maximum Parsimony or likelihood methods results in a 3-5%, 7-11% or 7-20% reconstruction improvement over the distance-based methods at 500, 1000 or 5000 microsatellites respectively (Fig. 5). There is little difference between the accuracy of Maximum Parsimony and the likelihood methods (<2% difference in the sample means and 0% in the plausible range for all comparisons with a given microsatellite number and missing data percentage).

As discussed above, Maximum Parsimony and the probabilistic methods have inbuilt ways of dealing with missing data, therefore performing better than the distance-based methods. However, this generally comes with the price of increased computational time (Supplementary Fig. 9). Maximum Parsimony takes considerably longer to run than the distance-based methods, and the run time increases as the proportion of missing data increases, although it always remains under two minutes. In contrast the computational time for the distance-based methods is hardly impacted. The run times for the probabilistic methods are one or two orders of magnitude higher than Maximum Parsimony, irrespective of the proportion of missing data. Therefore, if computational resources are limited, Maximum Parsimony would be recommended over the probabilistic approaches given their accuracies are similar. Note, that whilst the likelihood approaches are considerably more computationally intensive, they also yield much more information than the other options such as the posterior distributions of parameters of interest. They can also be used to compare various mutation models, and so if an understanding of

the generation process is a main goal of the study and the computational power is available, these methods will be of interest.

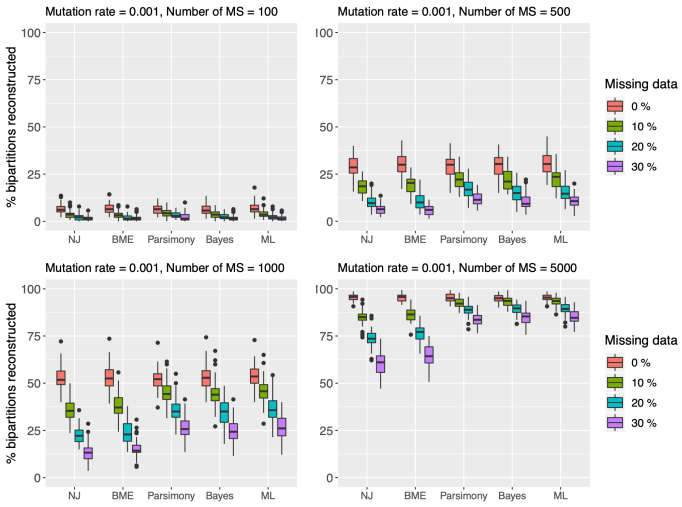

Fig. 5. **Comparison of phylogenetic methods with missing data.** Percentage of bipartitions from the true tree observed in reconstructed trees for distance-based (Neighbour Joining (NJ) and Balanced Minimum Evolution (BME)), Maximum Parsimony, Maximum Likelihood (ML) and Bayesian methods. Boxplots are based on 60 independently simulated microsatellite mutation patterns on one fixed tree with either 0, 10, 20 or 30% missing data and a mutation rate of 0.001.

Fig. 6. **Median reconstruction accuracy of Bayesian approach.** Heatmap showing the percentage of bipartitions from the true tree observed in reconstructed trees for the Bayesian method with differing numbers of microsatellites and amounts of missing data. Simulations were carried out on one fixed tree with a mutation rate of 0.001 and median percentage is shown in each box.

An alternative to leaving missing values in the input matrix is to impute values based on those observed. Unlike the distance-based methods, the input matrices to Maximum Parsimony and the probabilistic approaches require integers as the allowed transitions are between integer values, so the simplest option is to impute the modal value for each locus. As can be seen from Supplementary Figs. 15-17, it is almost always better to leave values as missing than to impute the mode, particularly for larger numbers of microsatellites and missing data percentages (with 100 microsatellites, there is very little difference in the means and a difference of 0% is always in the plausible range). Based on our results, we would therefore recommend leaving missing values rather than imputing when using Maximum Parsimony or a likelihood method.

The loss of performance as the proportion of missing data increases begs the question of whether it is better to retain or exclude microsatellite loci for which some observations are missing. Comparing, for example, the rows for 500 and 1000 microsatellites in Fig. 6, we see that the reconstruction with 1000 microsatellites and 30% missing data is worse than that with 500 fully observed microsatellites but better than that with 500 microsatellites and 10% missing. Therefore, the decision to include or exclude microsatellite loci from the analysis depends on the pattern of missing data. If certain loci have large amounts of missing data, it is better to exclude them and reduce the percentage of missing data. Conversely, if the missing data is spread evenly across loci, then removing all loci with missing values would result in a large decrease in the number of loci which would strongly adversely affect the reconstruction.

Using multiplexed PCR to genotype microsatellites, [24] estimated allelic dropout to be around 30% for one allele, corresponding to 15% missing data in this study, and [34] identified microsatellite genotypes at around 90% of their target loci using CRISPR-Cas9 fragmentation followed by sequencing. We conclude from our simulations that with these levels of missing data, it is better to include loci with missing data in the reconstruction rather than to remove them. In regimes where many microsatellites are observed e.g., the row with 5000 microsatellites, even 30% missing data has relatively little negative impact on the reconstruction (Figs 6, Supplementary Figs. 18, 19) and it therefore better to retain loci with a small percentage of missing data rather than lose many loci from the inference.

2.3 Comparing reconstruction methods with autosomal microsatellites

In order to make use of all the microsatellites in the genome, and indeed to apply the technique to females, we explore the performance of the various reconstruction methods on autosomal microsatellites, i.e., loci with two copies of each microsatellite. Reads aligning to the same genome location from two alleles can only be distinguished if there is a heterozygous polymorphism close to the microsatelilte, or if the repeat number of the two alleles is very different.

We model this scenario by grouping microsatellites into pairs (thereby keeping the number of microsatellites the same as in earlier sections) and initialising microsatellite lengths and within-locus length differences using distributions based on the data in [20] (described in full detail in the Methods, Section 4). Microsatellites mutate independently as before, PCR stutter error is added, and then, for the inference, we assume that the shorter of the two lengths in each pair would be assigned to one allele and the longer to the other. If the initial lengths of the two alleles are very different, the alleles will be correctly assigned, however if the initial lengths are the same or similar, drift in the microsatellite lengths may cause misassignment.

Supplementary Figs. 20 and 21 show that the trend and accuracy of the various methods are the same as for the X chromosome simulations (compare to Figs. 4 and 5). Whilst it is likely that misassignments negatively impact the reconstruction, they are rare in our simulations due to the relatively low tree depth and mutation rate used here.

2.4 Robustness of the inference for trees containing cells with different division rates

Thus far, we have simulated trees in which all cells have the same constant probability to divide or die. This is a realistic scenario for reconstructing lineage relations within cells of the same cell type. We would also like to compare reconstruction method for systems with hierarchical organisation, in which for example stem cells divide infrequently and can self-renew, and more differentiated cells divide more frequently. For this purpose, we next simulated trees with two types of cells: stem-like cells that have a low probability of dividing and a high probability to self-renew, and differentiated cells that divide more frequently and don't self-renew. In this section, results are presented for Tree 4, which is shown in Fig. 7 (depiction including dead cells is shown in Supplementary Fig. 22). We use the same inference methods as for Tree 1 to see if its topology impacts the accuracy of lineage reconstruction, however we now use a flat prior for the tree topology in the Bayesian inference i.e., *a priori* all trees have equal probability, so as to not have a mismatch between the assumptions of the inference and the model assumed when generating the tree.

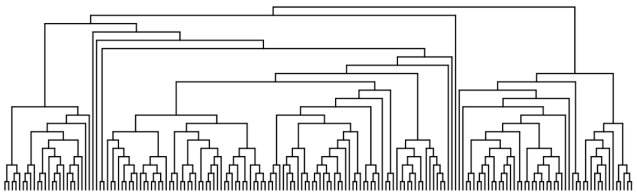

Fig. 7. **Simulated Tree 4.** Tree with 172 living cells simulated by the Gillespie algorithm. This tree contains two cell types, stem cells and differentiated cells, and after an initial self-renewal phase, at each reaction stem cells and differentiated cells had a 0.5 probability to differentiate or die respectively, otherwise the cell 'self-renews'. Differentiated cells divide at 100 times the rate of stem cells. The tree is plotted with nodes spaced evenly between the root and the final cells for clarity, although in reality division occurred at random times as is usual with the Gillespie algorithm.

When comparing distances and methods in the fully observed data scenario, a similar pattern of results were attained to those for the earlier trees (Supplementary Fig. 23 and Fig. 8), although the reconstruction works better than for the trees with one constant division rate (Trees 1-3). This is likely due to the fact that the majority of cells in Tree 4 are differentiated cells which have a greater depth than the average depth in Trees 1-3, and which have therefore accumulated more mutations aiding the reconstruction. When missing data was introduced, the overall conclusions are the same as for Trees 1-3, but the drop-off in performance was less severe at 1000 and 5000 microsatellites (compare Fig. 8 to Fig. 5, e.g., for Maximum Parsimony the mean was $25.6 \pm 3.0\%$ lower with 30% missing data for Tree 1 versus $16.8 \pm 1.8\%$ lower for Tree 4 with 1000 microsatellites), which is also likely due to the increased amount of information, and hence redundancy, from accumulated mutations. These results suggest that the inference methods tested here can be used to reconstruct lineage trees containing a variety of cell types with different division/death rates as long as the assumptions of the inference are not violated, such as those encoded in the tree prior for Bayesian inference.

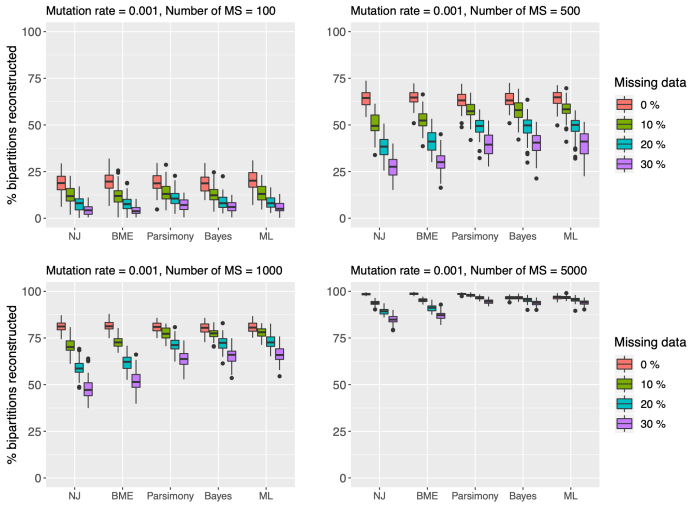

Fig. 8. **Comparison of phylogenetic methods with missing data, Tree 4.** Percentage of bipartitions from the true tree observed in reconstructed trees for distance-based (Neighbour Joining (NJ) and Balanced Minimum Evolution (BME)), Maximum Parsimony, Maximum Likelihood (ML) and Bayesian methods. Boxplots are based on 60 independently simulated microsatellite mutation patterns on one fixed tree with either 0, 10, 20 or 30% missing data and a mutation rate of 0.001.

2.5 Robustness of the inference for different mutation models

Here we will explore the impact of different mutation generation models on several methods of reconstruction. Of particular interest is how discrepancies in the mutation generation model and the models assumed in the likelihood of the probabilistic methods and the transition matrix in Maximum Parsimony impact the inference accuracy. Many models of varying complexity describing microsatellite mutation dynamics have been formulated ([36]). In the simplest and most commonly used model, the Stepwise Mutation Model (SMM, [37]), microsatellites symmetrically change length by increments of one motif. [38] found that most microsatellite variation takes place under this model, however, length changes of more than one motif have been observed ([39], [40]). An extension of the SMM, the Multistep Mutation Model (MMM), therefore allows microsatellites to change length by more than one motif but with a decreasing probability for larger jumps ([41], [42]). There are also alternative models in which microsatellites preferentially increase in length ([40], [43]), and the Asymmetric Multistep Mutation Model (AMMM) allows microsatellites to change length by one or more motif with a larger probability to increase in length than to decrease ([43]).

Thus far, the assumptions of probabilistic inference methods have matched the mutation generation model in

the simulations, with both using the MMM. Here, we took Tree 1 (Fig. 1) and a fixed overall mutation rate of 0.001, and simulated mutations using the SMM, MMM and AMMM (full details in the Methods section). We then tested the same tree reconstruction methods as in the previous sections i.e., the probabilistic approaches still assume a MMM. We also simulated mutations with the MMM and then assumed the SMM for the reconstruction with probabilistic inference.

From Fig. 9, we see that the reconstruction accuracy is not strongly impacted by the type of model used to simulate the mutations and that for a given microsatellite number and method, 0% is always in the plausible effect size range for each comparison. Surprisingly, this is also the case for the probabilistic approaches even when there is a mismatch between the generation and inference models. This is probably due to the specific models tested in which jumps of more than one repeat are considerably less likely than jumps of one repeat meaning that the data simulated under the MMM is not appreciably different than data simulated under the SMM. Based on our results, we conclude that the inference methods tested are relatively robust to slight variations in the true generation process and to the choices researchers must make when choosing a probabilistic model for inference. We cannot exclude that a very large discrepancy between generation and inference model would have a negative impact, and also note that the types of parameters inferred as well as their values will be affected by the model used in inference.

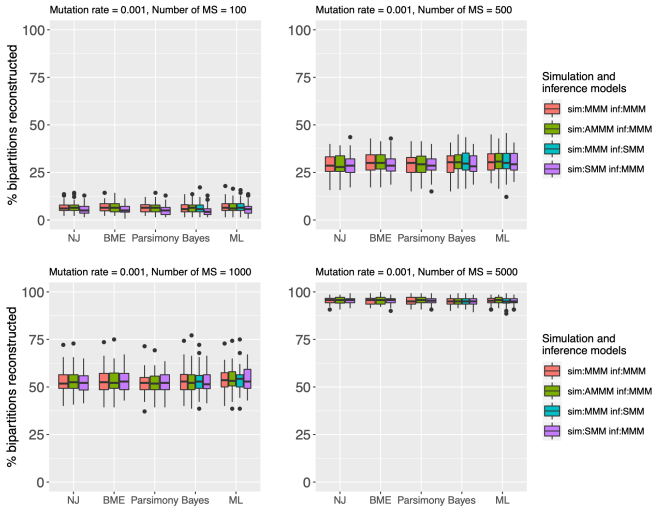

Fig. 9. **Comparison of phylogenetic methods for different microsatellite mutation models.** Percentage of clades from the true tree observed in reconstructed trees for distance-based (Neighbour Joining (NJ) and Balanced Minimum Evolution (BME)), Maximum Parsimony, Maximum Likelihood and Bayesian methods. Boxplots are based on 60 independently simulated microsatellite mutation patterns on one fixed tree with a mutation rate of 0.001. Mutations were generated using one of three models: Stepwise Mutation Model (SMM), the Multistep Mutation Model (MMM), or the Asymmetric Multistep Mutation Model (AMMM). Abbreviations: sim = mutations simulated with this model, inf = mutation model used in likelihood-based inference methods.

3 Discussion and Conclusion

There is great interest in the ability to reconstruct cell lineage without the need for prospective tagging. This can be achieved by observing mutations which occur during cell division on a single cell basis and then using phylogenetic algorithms to reconstruct relationships between cells. We focus here on mutations in microsatellites as they are abundant in the human genome and have a higher mutation rate than the other types of mutations. We note, however, that some of the methods investigated here have also been applied in other contexts, for example, Maximum Parsimony ([6]) and agglomerative clustering (similar to NJ) ([8]) have been used to reconstruct lineage trees from genetic recombination data. Recent experimental progress has led to a marked increase in the number of microsatellite loci which can be observed in parallel in individual cells. [44] have developed a pipeline in which 10,000s of pre-selected microsatellites can be genotyped in single-cell whole genome amplified DNA using molecular inversion probes. [34] used CRISPR-Cas9 targeted DNA fragmentation followed by sequencing to observe 2,000 loci in bulk DNA, but it seems plausible that this approach can be extended to single cell data. There has also been recent progress in terms of the software available to implement state-of-the-art likelihood-based phylogenetic inference. It is, therefore, a good time to reassess which reconstruction algorithms will optimise the interpretation of single cell datasets.

We tested a range of phylogenetic algorithms including computationally fast distance-based algorithms, Maximum Parsimony and computationally intensive likelihood-based approaches. In the situation where the data is fully observed, we found little difference in performance between the various reconstruction methods. However, both Maximum Parsimony and the probabilistic approaches require considerably more computational time than the distance-based approaches. For trees with even a moderate number of cells, more than around 20, it is not possible to exhaustively search the tree space, and hence heuristic searches need to be carried out which require increasingly more computation as the number of cells increases. In addition, as the tree becomes larger and/or the model more complex, the likelihood computation itself becomes more demanding. In this study, running the probabilistic inference took up the vast majority of the computational time. Therefore, in situations with low levels of missing data ($< 10\%$), distance-based approaches can be used in conjunction with distances adapted to take account of the missing data. We found the L_1 distance to work well using only the loci jointly observed.

However, recently developed single cell NGS techniques, which have already been used for lineage tracing ([45], [46]) and which have much potential for future development, result in uneven genome coverage, and hence in more sparse data matrices than those found through the more targeted approaches previously used ([24], [25]). Two questions of interest, therefore, are whether it remains possible to reconstruct lineage trees from this type of data, and whether the Maximum Parsimony and likelihood methods, which have inherent means for dealing with missing data, will perform better in this type of scenario.

Our results show that, although, as would be expected, there is some loss in accuracy as data is removed from the inference, lineage trees can still be reconstructed from observations with missing values, and that, indeed, Maximum

Parsimony and the likelihood methods lose less accuracy than the distance-based approaches for the same amount of missing data. This suggests that microsatellite data collected using new single cell experimental techniques would benefit from interpretation with the state-of-the-art likelihood algorithms used in phylogenetic studies. The price to pay for this increased accuracy comes in the form of increased computation time. This may become a problem as experimental progress leads to increases in both the number of cells and the number of microsatellites genotyped. However, with the use of parallel computing, Bayesian inference can be scaled up considerably beyond its use in this paper.

We anticipate that by harnessing the power of state-of-the-art phylogenetic likelihood methods, lineage inference from recently developed single cell sequencing technologies can be optimised to answer questions of long-standing interest about differentiation and developmental processes.

4 METHODS

4.1 Simulating trees

Trees were simulated using the Gillespie algorithm starting from a single cell. The Gillespie algorithm was developed to simulate trajectories of stochastic equations describing chemical reactions. In this paper, 'reactions' are either cell division or cell death, and these two possibilities have a constant probability given a reaction across the tree. Trees were simulated with different death probabilities to cover a range of biological situations e.g. probabilities of 0, 0.2 and 0.3 of death given a reaction.

To produce the tree with stem-like and differentiated-like cells (Tree 4), the starting stem cell was initially expanded using only self-renewal, and then differentiation of stem cells and division and death of differentiated cells was introduced. Differentiated cells divided at a rate 100 times higher than stem cells as estimated for multipotent progenitor cells versus stem cells in [47]. Table 1 shows the simulation used to produce the trees used in the paper.

The simulated trees have around 150 living cells at the final time point, which is towards the upper end of currently published experimental work ([21], [25], [44]). We expect the throughput in terms of cell numbers to increase dramatically in the near future, but in order to run reconstructions on repeat simulations using the likelihood approaches, we keep to this relatively low number of cells. We conservatively take the average tree depth to be 10 divisions, and we expect that in some experimental settings a greater number of divisions may take place, accumulating more mutation making the reconstruction easier. The figures throughout this paper show results for one specific tree (Fig. 1) to allow for comparison across situations, but results for other trees showed very similar trends and are shown in the Supporting Information.

TABLE 1
Summary of trees for which results are presented

Tree	$P(death\|reaction)$	Simulation
Tree 1	0.2	constant-rate birth-death process
Tree 2	0	constant-rate birth-death process
Tree 3	0.3	constant-rate birth-death process
Tree 4	0.5	birth-death process with two cell types

Description of Gillespie simulation of each tree for which results are presented.

4.2 Simulating mutations

To simulate changes in microsatellite lengths on the male X chromosome, the initial stem cell was assigned a set of M microsatellites which were then allowed to mutate with a given probability, p, per cell division.

The initial length distribution of the microsatellites is defined over repeat numbers 10-20, as [20] observe that there is little variation in microsatellites with fewer than 10 repeats and we assume that it will be rare to observe microsatellites of repeat number greater than 20 in short NGS reads. We model [20]'s counts of 10-20 repeat microsatellites as a truncated geometric distribution, such that

$$P(L = l) = \frac{0.7^l}{\sum_{m=10}^{20} 0.7^m} \quad \text{for } l \text{ in 10, ..., 20.}$$

The parameter of the geometric distribution was estimated from the data in [20], and averaged across di, tri, tetra and penta loci.

In the case of autosomal microsatellites, we group the microsatellites into pairs, one for each of the maternal and paternal alleles. The lengths of the maternal alleles were set as above for the X chromosome. [20] found that around half of the 10-20 repeat microsatellite loci were homozygous, so for half the loci we set the initial paternal allele length equal to the maternal length. We then modelled their distribution of repeat number difference as a geometric distribution

$$P(D = d) = (1-p)^{d-1}p \quad \text{for } d \text{ in 1, 2, ...}$$

with $p = 0.4$, again estimated from the data in [20].

For the majority of the paper, mutations were simulated using a variant of the Multistep Mutation Model (MMM). Given a mutation takes place, mutations increase or decrease the length of a microsatellite with a symmetric probability of 0.5, and can change length by one, two or three repeats. The probability distribution for length change is a truncated geometric distribution

$$P(K = k) = \frac{q(1-q)^{k-1}}{1-(1-q)^3},$$

with $q = 0.8$. We also compare the reconstruction when the mutations are simulated using two other models. The Stepwise Mutation Model (SMM) is the simplest model, with one probability of mutation, which is symmetric in direction and can only cause a change of one repeat unit. The Asymmetric Multistep Mutation Model (AMMM) relaxes the assumption that length changes are symmetric, and we take the probability that mutation increase the length of a microsatellite to be 0.6, with the same truncated geometric distribution used to model the size of the jump.

4.3 Simulating PCR stutter error and allele ambiguity

The high mutability which makes microsatellite loci useful for reconstructing cell lineage also results in the acquisition

of noise during PCR amplification. Known as PCR stutter, it primarily results in signal at N-1 repeats for a true allele with N repeats [48]. In general, amplification of normal amounts of DNA results in stutter signal which is less than 15% of the allele peak, however when PCR amplification starts from a very small amount of initial template, slippage in early cycles can lead to larger stutter peaks and even result in the wrong allele being called ([49]). To model this, based on the low template simulations in [49], we assume a prior distribution for the stutter ratio, r, of uniform on $[0.1, 0.3]$. Each microsatellite locus will therefore yield a pool of NGS reads in which the ratio of reads with N repeats to those with (N-1) repeats is (1-r)/r. We then make the simple assumption that 10 reads are sampled from this pool at each locus and compute the probability that we incorrectly call the length as N-1 rather than N, i.e. the probability that the number of reads with the stutter length, N_{stutter}, is greater than the number of reads with the true length.

$$\begin{aligned} P(\text{calling } N-1) &= P(N_{\text{stutter}} > 5) \\ &= \int P(N_{\text{stutter}} > 5|r)P(r)\mathrm{d}r \\ &= \int_{0.1}^{0.3} \sum_{k=6}^{10} \binom{10}{k} r^k (1-r)^{10-k} \mathrm{d}r \\ &\approx 0.0024, \end{aligned}$$

with the integration done numerically using the trapezium rule. To add this source of noise in our simulations, we simply output N-1 rather than N with a probability 0.0024 and do the reconstruction using these values rather than the true values.

We also model the error due to the ambiguity of allelic origin for autosomal microsatellites in NGS data. To understand why autosomal microsatellite are more problematic than X chromosome microsateliltes, it is helpful to consider the form of the input to each reconstruction method, which for autosomal data is a matrix of dimension 'Number of cells' × '2 × Microsatellite number'. For each locus, the two observed alleles need to be assigned to the appropriate column in the matrix, and errors can occur here if the two alleles have similar lengths and if there is no nearby heterozygous polymorphism to distinguish them. We assume that practitioners would take a simple approach and assign the shorter length at each locus to the maternal allele and the longer to the paternal allele. We replicate this experimental situation by simulating microsatellite mutations and stutter error independently for each allele, and then, before inference, allocating the smaller length of each pair to the maternal allele and the larger to the paternal.

4.4 Measuring tree similarity

The most commonly used distance between trees is the Robinson-Foulds (RF) distance ([50]), which is based on similarity of tree bipartitions. Cutting an internal tree branch partitions the leaves into two sets, and the bipartition resulting from cutting each internal branch is unique. The RF distance counts the number of bipartitions observed in one tree but not the other. The drawbacks to this type of distance are that moving one leaf can change the distance considerably if many clades are changed, and that the maximum value depends on the size of the tree making interpretation difficult.

To aid interpretation and enable comparisons across trees, we defined a similarity score based on the idea of the RF distance in which the percentage of bipartitions present in the original tree also present in the reconstructed tree is computed. If the two trees are identical, the score will be 100% as all of the clades observed in the original tree will also be in the reconstructed tree. Most of the reconstruction methods output unrooted trees (other than the distance-based clustering methods) and therefore both the reconstructed trees and the true tree are unrooted (using the 'unroot' function from the ape package in R ([51])) before the percentage of common bipartitions is computed.

Other tree comparison metrics, such as the tree alignment score ([52]) or the Maximum Agreement Subtree ([53]), are available and were also computed, but as the trends were similar across approaches, and the percentage approach is the easiest to interpret, we only show the percentage of shared bipartitions score.

4.5 Comparing scenarios

In order to compare the reconstruction accuracy of various scenarios, we concentrate primarily on the effect size rather than on p-values. This reflects our primary interest in how much each method or setting impacts our reconstruction, but also the fact that, unlike an experimental setting where money and time can be prohibitive, in a simulation study one can always achieve a desired significance by increasing the number of repeats ([54], [55], [56]).

For each setting, generally one box-and-whisker in a figure, we compute the sample mean, $\bar{x}$, of the 60 independent repeats and also a 95% confidence interval (CI) of the mean assuming the data is normally distributed, using $\bar{x} \pm 1.96\sigma_m$, where σ_m is the standard error of the mean. When we compare two scenarios, we compute the effect size by subtracting the means, and then we find a plausible range for the effect size by computing the largest and smallest possible difference using the boundary values of the confidence intervals, as suggested by [57].

4.6 Tree reconstruction

We use directly, or adapt for our purpose, several tools from evolutionary phylogenetics designed to reconstruct relationships between species. Neighbour joining (NJ) is a distance-based clustering method which builds a tree using a matrix of pairwise evolutionary distances, which can be defined in a number of ways ([32]). NJ iteratively selects a taxon pair, builds a new subtree and agglomerates the pair, greedily minimising a weighted tree length ([58]). Distance-based methods are quick and therefore suitable for problems such as ours where we hope to reconstruct trees with large numbers of cells.

Balanced minimum evolution (BME) is another distance-based method for phylogeny reconstruction based on NJ ([33]). BME minimises the same weighted tree length as NJ, but it isn't a greedy algorithm. Instead, the BME method

rearranges the tree multiple times, computes the weighted tree length, and outputs the tree with the minimum.

NJ ([32]) and BME ([33]) were implemented in the R software suite ([59]) using the Phangorn ([60]) and Ape ([51]) packages respectively. In the 'fastme.bal' function used to implement BME, all rearrangement options were set to TRUE. Both methods take as input a distance matrix which was computed in advance using either Manhattan (L_1), Euclidean (L_2) or cosine distance. The cosine distance between two vectors $\mathbf{a}$ and $\mathbf{b}$ was computed

$$1 - \frac{\mathbf{a}.\mathbf{b}}{||\mathbf{a}||\,||\mathbf{b}||},$$

where $||\mathbf{a}||$ is the Euclidean norm of the vector $\mathbf{a}$.

Maximum Parsimony selects the tree which minimises the number of mutations required to explain the data. For a given tree, a parsimony score which counts the number of mutational changes required is computed (computationally fast) and then the space of possible trees is searched to find the tree with the lowest score (computationally slow). The tree space can only be exhaustively searched for very small trees, and therefore, the normal strategy is to input a 'sensible' starting tree e.g. the output of a NJ algorithm, and then search for trees close by.

Maximum Parsimony was implemented in PAUP* ([35]). Ordered characters were used which means that when the parsimony score is computed for any given tree, it is assumed that microsatellite length changes must proceed progressively through the numbered repeats (similar to the assumptions of the Stepwise Mutation Model). The following parameters were set when running the inference: 'condense collapse = NO' (branches with zero length not collapsed so as to output a binary tree), 'addseq = random' (the starting tree was build using a random sequence of addition of cells), 'nreps = 5' (5 runs with different random starting trees), 'swap = tbr' (Tree Bisection and Reconnection used to rearrange the tree each iteration) and 'rearrlimit = 10^6' (10^6 tree rearrangements tried).

Maximum Likelihood and Bayesian methods require a probabilistic model of the process which produced the pattern of microsatellite lengths; the tree itself is a parameter in the model. These 'state-of-the-art' methods allow knowledge of the mutation process to be incorporated into the inference process, and are very flexible as many different models can be formulated. However, as with Maximum Parsimony, these methods require a lot of computation as it is generally not possible to maximise the likelihood (or posterior in the Bayesian case) analytically across all the possible trees, and so a heuristic search in tree space is required.

The likelihood approaches were implemented in the RevBayes software ([28], [61]) for Bayesian phylogenetic inference. Microsatellite mutation is modelled as a continuous time Markov model characterised by its instantaneous-rate matrix. We assume that the rate matrix, Q, is symmetric and that the transition rates, α, β and γ, between numbers of repeats depend only on the magnitude of the change

$$Q = \begin{bmatrix} * & \alpha & \beta & \gamma \\ \alpha & * & \alpha & \beta \\ \beta & \alpha & * & \alpha \\ \gamma & \beta & \alpha & * \end{bmatrix}.$$

We set priors on each of the transition rates of an exponential distribution with a mean of 1. For Trees 1-3, the prior for the tree is a constant-rate birth-death process with the turnover and diversification given lognormal priors as described in Chapter 12 of the full RevBayes manual (Basic Diversification Rate Estimation http://revbayes.github.io/tutorials.html). Following these suggestions, the prior mean for the diversification and turnover rates was set as the number of observed cells and the standard deviation was set to cover two orders of magnitude.

Markov chain Monte Carlo was used to sample trees and parameter values from the posterior distributions. To enable repeat simulations to be carried out, a relatively low number of MCMC iterations were used in simulations: 2000 burn-in iterations during which the proposal distributions were tuned followed by 8000 samples. This low number of iterations was chosen to allow the inference to be repeated 60 times for each setting. The tree with the highest likelihood was used as the estimate of the Maximum Likelihood tree, and the tree with the maximum posterior probability was used as the Bayesian tree.

ACKNOWLEDGMENTS

The authors would like to thank Jason Cosgrove and Ken Duffy for discussions and critical feedback on the manuscript.

REFERENCES

[1] B. Spanjaard and J. P. Junker, "Methods for lineage tracing on the organism-wide level," *Current Opinion in Cell Biology*, vol. 49, pp. 16–21, 2017.

[2] K. Schepers, E. Swart, J. W. van Heijst, C. Gerlach, M. Castrucci, D. Sie, M. Heimerikx, A. Velds, R. M. Kerkhoven, R. Arens, and T. N. Schumacher, "Dissecting T cell lineage relationships by cellular barcoding," *The Journal of Experimental Medicine*, vol. 205, no. 10, pp. 2309–2318, Sep. 2008.

[3] A. Gerrits, B. Dykstra, O. J. Kalmykowa, K. Klauke, E. Verovskaya, M. J. Broekhuis, G. De Haan, and L. V. Bystrykh, "Cellular barcoding tool for clonal analysis in the hematopoietic system," *Blood*, vol. 115, no. 13, pp. 2610–2618, Apr. 2010.

[4] S. H. Naik, L. Perié, E. Swart, C. Gerlach, N. Van Rooij, R. J. De Boer, and T. N. Schumacher, "Diverse and heritable lineage imprinting of early haematopoietic progenitors," *Nature*, vol. 496, no. 7444, pp. 229–232, Apr. 2013.

[5] R. Lu, N. F. Neff, S. R. Quake, and I. L. Weissman, "Tracking single hematopoietic stem cells *in vivo* using high-throughput sequencing in conjunction with viral genetic barcoding," *Nature Biotechnology*, vol. 29, no. 10, pp. 928–933, Oct. 2011.

[6] A. McKenna, G. M. Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier, and J. Shendure, "Whole-organism lineage tracing by combinatorial and cumulative genome editing," *Science*, vol. 353, no. 6298, pp. aaf7907–aaf7907, 2016.

[7] R. Kalhor, P. Mali, and G. M. Church, "Rapidly evolving homing CRISPR barcodes," *Nature Methods*, vol. 14, no. 2, pp. 195–200, Dec. 2017.

[8] J. P. Junker, B. Spanjaard, J. Peterson-Maduro, A. Alemany, B. Hu, M. Florescu, and A. van Oudenaarden, "Massively parallel whole-organism lineage tracing using CRISPR/Cas9 induced genetic scars," Tech. Rep. 10.1101/056499, 2016.

[9] W. Pei, T. B. Feyerabend, J. Rössler, X. Wang, D. Postrach, K. Busch, I. Rode, K. Klapproth, N. Dietlein, C. Quedenau, W. Chen, S. Sauer, S. Wolf, T. Höfer, and H.-R. Rodewald, "*Polylox* barcoding reveals haematopoietic stem cell fates realized *in vivo*," *Nature*, vol. 548, no. 7668, pp. 456–460, Aug. 2017.

[10] A. E. Rodriguez-Fraticelli, S. L. Wolock, C. S. Weinreb, R. Panero, S. H. Patel, M. Jankovic, J. Sun, R. A. Calogero, A. M. Klein, and F. D. Camargo, "Clonal analysis of lineage fate in native haematopoiesis," *Nature*, vol. 553, no. 7687, pp. 212–216, Jan. 2018.

[11] J. Sun, A. Ramos, B. Chapman, J. B. Johnnidis, L. Le, Y.-J. Ho, A. Klein, O. Hofmann, and F. D. Camargo, "Clonal dynamics of native haematopoiesis," *Nature*, vol. 514, no. 7522, pp. 322–327, Oct. 2014.

[12] L. Perié, K. R. Duffy, L. Kok, R. J. de Boer, and T. N. Schumacher, "The Branching Point in Erythro-Myeloid Differentiation," *Cell*, vol. 163, no. 7, pp. 1655–1662, Dec. 2015.

[13] L. Biasco, D. Pellin, S. Scala, F. Dionisio, L. Basso-Ricci, L. Leonardelli, S. Scaramuzza, C. Baricordi, F. Ferrua, M. P. Cicalese, S. Giannelli, V. Neduva, D. J. Dow, M. Schmidt, C. Von Kalle, M. G. Roncarolo, F. Ciceri, P. Vicard, E. Wit, C. Di Serio, L. Naldini, and A. Aiuti, "In Vivo Tracking of Human Hematopoiesis Reveals Patterns of Clonal Dynamics during Early and Steady-State Reconstitution Phases," *Cell Stem Cell*, vol. 19, no. 1, pp. 107–119, Jul. 2016.

[14] G. D. Evrony, X. Cai, E. Lee, L. B. Hills, P. C. Elhosary, H. S. Lehmann, J. J. Parker, K. D. Atabay, E. C. Gilmore, A. Poduri, P. J. Park, and C. A. Walsh, "Single-Neuron Sequencing Analysis of L1 Retrotransposition and Somatic Mutation in the Human Brain," *Cell*, vol. 151, no. 3, pp. 483–496, Oct. 2012.

[15] N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks, and M. Wigler, "Tumour evolution inferred by single-cell sequencing," *Nature*, vol. 472, no. 7341, pp. 90–94, Apr. 2011.

[16] S. Behjati, M. Huch, R. van Boxtel, W. Karthaus, D. C. Wedge, A. U. Tamuri, I. Martincorena, M. Petljak, L. B. Alexandrov, G. Gundem, P. S. Tarpey, S. Roerink, J. Blokker, M. Maddison, L. Mudie, B. Robinson, S. Nik-Zainal, P. Campbell, N. Goldman, M. van de Wetering, E. Cuppen, H. Clevers, and M. R. Stratton, "Genome sequencing of normal cells reveals developmental lineages and mutational processes," *Nature*, vol. 513, no. 7518, pp. 422–425, Sep. 2014.

[17] D. Frumkin, A. Wasserstrom, S. Kaplan, U. Feige, and E. Shapiro, "Genomic variability within an organism exposes its cell lineage tree," *PLoS Computional Biology*, vol. 1, no. 5, 2005.

[18] J. X. Sun, A. Helgason, G. Masson, S. S. Ebenesersdóttir, H. Li, S. Mallick, S. Gnerre, N. Patterson, A. Kong, D. Reich, and K. Stefansson, "A direct characterization of human mutation based on microsatellites," *Nature Genetics*, vol. 44, no. 10, pp. 1161–1165, Oct. 2012.

[19] H. Ellegren, "Microsatellites: Simple sequences with complex evolution," *Nature Reviews*, vol. 5, pp. 435–445, 2004.

[20] B. A. Payseur, P. Jing, and R. J. Haasl, "A Genomic Portrait of Human Microsatellite Variation," *Molecular Biology and Evolution*, vol. 28, no. 1, pp. 303–312, Jan. 2011.

[21] Y. Reizel, N. Chapal-Ilani, R. Adar, S. Itzkovitz, J. Elbaz, Y. E. Maruvka, E. Segev, L. I. Shlush, N. Dekel, and E. Shapiro, "Colon stem cell and crypt dynamics exposed by cell lineage reconstruction," *PLoS Genetics*, vol. 7, no. 7, 2011.

[22] L. I. Shlush, N. Chapal-Ilani, R. Adar, N. Pery, Y. Maruvka, A. Spiro, R. Shouval, J. M. Rowe, M. Tzukerman, D. Bercovich, S. Izraeli, G. Marcucci, C. D. Bloomfield, T. Zuckerman, K. Skorecki, and E. Shapiro, "Cell lineage analysis of acute leukemia relapse uncovers the role of replication-rate heterogeneity and microsatellite instability," *Blood*, vol. 120, no. 3, pp. 603–612, 2012.

[23] D. Frumkin, A. Wasserstrom, S. Itzkovitz, D. Frumkin, A. Wasserstrom, S. Itzkovitz, and T. Stern, "Cell Lineage Analysis of a Mouse Tumor Cell Lineage Analysis of a Mouse Tumor," *Cancer Research*, pp. 5924–5931, 2008.

[24] Y. Reizel, S. Itzkovitz, R. Adar, J. Elbaz, A. Jinich, N. Chapal-ilani, D. Benayahu, K. Skorecki, E. Segal, N. Dekel, and E. Shapiro, "Cell Lineage Analysis of the Mammalian Female Germline," *PLoS*, vol. 8, no. 2, 2012.

[25] S. J. Salipante, J. M. Thompson, and M. S. Horwitz, "Phylogenetic fate mapping: Theoretical and experimental studies applied to the development of mouse fibroblasts," *Genetics*, vol. 178, no. 2, pp. 967–977, Feb. 2008.

[26] S. J. Salipante and M. S. Horwitz, "Phylogenetic fate mapping." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 14, pp. 5448–5453, Apr. 2006.

[27] N. Chapal-Ilani, Y. E. Maruvka, A. Spiro, Y. Reizel, R. Adar, L. I. Shlush, and E. Shapiro, "Comparing Algorithms That Reconstruct Cell Lineage Trees Utilizing Information on Microsatellite Mutations," *PLoS Computational Biology*, vol. 9, no. 11, 2013.

[28] S. Hohna, M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist, "RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language," *Systematic Biology*, vol. 65, no. 4, pp. 726–736, Jul. 2016.

[29] A. Csordas and R. Bouckaert, "Bayesian analysis of normal mouse cell lineage trees allowing intra-individual, cell population specific mutation rates," Tech. Rep., 2016.

[30] C.-H. Wu and A. J. Drummond, "Joint Inference of Microsatellite Mutation Models, Population History and Genealogies Using Transdimensional Markov Chain Monte Carlo," *Genetics*, vol. 188, no. 1, pp. 151–164, May 2011.

[31] A. Spiro and E. Shapiro, "Accuracy of Answers to Cell Lineage Questions Depends on Single-Cell Genomics Data Quality and Quantity," *PLOS Computational Biology*, vol. 12, no. 6, p. e1004983, Jun. 2016.

[32] M. Nei and N. Saitou, "The neighbor-joining method: A new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, Jul. 1987.

[33] R. Desper and O. Gascuel, "Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle," *Journal of Computational Biology*, vol. 9, no. 5, pp. 687–705, Oct. 2002.

[34] G. Shin, S. M. Grimes, H. Lee, B. T. Lau, L. C. Xia, and H. P. Ji, "CRISPR–Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis," *Nature Communications*, vol. 8, pp. 14 291–14 291, Feb. 2017.

[35] D. Swofford, "PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)." Sinauer Associates, Sunderland, Massachusetts, 2003.

[36] A. Bhargava and F. F. Fuentes, "Mutational Dynamics of Microsatellites," *Molecular Biotechnology*, vol. 44, no. 3, pp. 250–266, Mar. 2010.

[37] T. Ohta and M. Kimura, "A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population*," *Genetics Research*, vol. 22, no. 2, pp. 201–204, Oct. 1973.

[38] J. R. Dettman and J. W. Taylor, "Mutation and Evolution of Microsatellite Loci in Neurospora," *Genetics*, vol. 168, no. 3, pp. 1231–1248, Nov. 2004.

[39] X. Xu, M. Peng, Z. Fang, and X. Xu, "The direction of microsatellite mutations is dependent upon allele length," *Nature Genetics*, vol. 24, no. 4, pp. 396–399, Apr. 2000.

[40] Q.-Y. Huang, F.-H. Xu, H. Shen, H.-Y. Deng, Y.-J. Liu, Y.-Z. Liu, J.-L. Li, R. R. Recker, and H.-W. Deng, "Mutation Patterns at Dinucleotide Microsatellite Loci in Humans," *The American Journal of Human Genetics*, vol. 70, no. 3, pp. 625–634, Mar. 2002.

[41] A. D. Rienzo, A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin, and N. B. Freimer, "Mutational processes of simple-sequence repeat loci in human populations." *Proceedings of the National Academy of Sciences*, vol. 91, no. 8, pp. 3166–3170, Apr. 1994.

[42] R. Nielsen and P. J. Palsbøll, "Single-Locus Tests of Microsatellite Evolution: Multi-Step Mutations and Constraints on Allele Size," *Molecular Phylogenetics and Evolution*, vol. 11, no. 3, pp. 477–484, Apr. 1999.

[43] G. Cooper, N. J. Burroughs, D. A. Rand, D. C. Rubinsztein, and W. Amos, "Markov Chain Monte Carlo analysis of human Y-chromosome microsatellites provides evidence of biased mutation," *Proceedings of the National Academy of Sciences*, vol. 96, no. 21, pp. 11 916–11 921, Oct. 1999.

[44] L. Tao, O. Raz, Z. Marx, T. Biezuner, S. Amir, L. Milo, R. Adar, A. Onn, N. Chapal-Ilani, V. Berman, R. Levy, B. Oron, and E. Shapiro, "A duplex MIPs-based biological-computational cell lineage discovery platform," *bioRxiv*, pp. 191 296–191 296, Oct. 2017.

[45] G. D. Evrony, E. Lee, B. K. Mehta, Y. Benjamini, R. M. Johnson, X. Cai, L. Yang, P. Haseley, H. S. Lehmann, P. J. Park, and C. A. Walsh, "Cell Lineage Analysis in Human Brain Using Endogenous Retroelements," *Neuron*, vol. 85, no. 1, pp. 49–59, Jan. 2015.

[46] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas, "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, vol. 8, p. 14049, Jan. 2017.
[47] K. Busch, K. Klapproth, M. Barile, M. Flossdorf, T. Holland-Letz, S. M. Schlenner, M. Reth, T. Höfer, and H.-R. Rodewald, "Fundamental properties of unperturbed haematopoiesis from stem cells *in vivo*," *Nature*, vol. 518, no. 7540, pp. 542–546, Feb. 2015.
[48] P. S. Walsh, N. J. Fildes, and R. Reynolds, "Sequence Analysis and Characterization of Stutter Products at the Tetranucleotide Repeat Locus VWA," *Nucleic Acids Research*, vol. 24, no. 14, pp. 2807–2812, Jul. 1996.
[49] K. R. Duffy, N. Gurram, K. C. Peters, G. Wellner, and C. M. Grgicak, "Exploring STR signal in the single- and multicopy number regimes: Deductions from an in silico model of the entire DNA laboratory process: Nucleic Acids," *ELECTROPHORESIS*, vol. 38, no. 6, pp. 855–868, Mar. 2017.
[50] D. F. Robinson and L. R. Foulds, "Comparison of phylogenetic trees," *Mathematical Biosciences*, vol. 53, no. 1-2, pp. 131–147, 1981.
[51] E. Paradis, J. Claude, and K. Strimmer, "APE: Analyses of phylogenetics and evolution in R language," *Bioinformatics*, vol. 20, no. 2, pp. 289–290, Jan. 2004.
[52] T. Nye, P. Lio, and W. Gilks, "A novel algorithm and web-based tool for comparing two alternative phylogenetic trees," *Bioinformatics*, vol. 22, no. 1, pp. 117–119, 2006.
[53] A. D. Gordon, "On the assessment and comparison of classifications," in *Analyse de Données et Informatique*, 1980, pp. 149–160.
[54] J. W. White, A. Rassweiler, J. F. Samhouri, A. C. Stier, and C. White, "Ecologists should not use statistical significance tests to interpret simulation model results," *Oikos*, vol. 123, no. 4, pp. 385–388, Apr. 2014.
[55] M. A. Hofmann, "Null hypothesis significance testing in simulation," in *2016 Winter Simulation Conference (WSC)*. Washington, DC, USA: IEEE, Dec. 2016, pp. 522–533.
[56] M. A. Hofmann, S. Meyer-Nieberg, and T. Uhlig, "INFERENTIAL STATISTICS AND SIMULATION GENERATED SAMPLES: A CRITICAL REFLECTION," in *2018 Winter Simulation Conference (WSC)*. Gothenburg, Sweden: IEEE, Dec. 2018, pp. 479–490.
[57] L. G. Halsey, "The reign of the p -value is over: What alternative analyses could we employ to fill the power vacuum?" *Biology Letters*, vol. 15, no. 5, p. 20190174, May 2019.
[58] O. Gascuel and M. Steel, "Neighbor-Joining Revealed," *Molecular Biology and Evolution*, vol. 23, no. 11, pp. 1997–2000, Aug. 2006.
[59] R. D. C. Team, "Computational Many-Particle Physics," Vienna, Austria, 2008.
[60] K. P. Schliep, "Phangorn: Phylogenetic analysis in R," *Bioinformatics*, vol. 27, no. 4, pp. 592–593, 2011.
[61] S. Höhna, T. A. Heath, B. Boussau, M. J. Landis, F. Ronquist, and J. P. Huelsenbeck, "Probabilistic graphical model representation in phylogenetics," *Systematic Biology*, vol. 63, no. 5, pp. 753–771, Sep. 2014.

Anne-Marie Lyne received her PhD degree in Systems Biology from University College London in 2015 and currently works as a post-doctoral researcher in Quantitative Immunology at Institut Curie, Paris. Her research interests include the application of bioinformatics, mathematical modelling and statistical inference to problems in Biology.

Leïla Perié received her PhD degree in Immunology from AgroParistech Doctoral School in 2009 and now leads the Quantitative Immunology group at Institut Curie, Paris. Her research interests include a variety of experimental and analytical approaches to the study of haematopoiesis.

Funding: 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 758170). The study was supported by an ATIPAvenir grant from CNRS and Bettencourt-Schueller Foundation (to L.P.) and two grants from the Labex CelTisPhyBio (ANR-10-LBX-0038) and Idex Paris-Science-Lettres Program (ANR-10-IDEX-0001-02 PSL) (to L.P.).