Cellular proliferation biases clonal lineage tracing and trajectory inference

Becca Bonham-Carter and Geoffrey Schiebinger

July 20, 2023

Abstract

We identify a fundamental statistical phenomenon in single-cell time courses with clone-based lineage tracing. Through simple probabilistic arguments, we show how the relative growth rates of cells influence the probability that they will be sampled in clones observed across multiple time points. We support these arguments with a simple simulation study and a time-course of T-cell development, and we demonstrate that this bias can impact fate probability predictions from trajectory inference methods. Finally, we explore how to develop trajectory inference methods which are robust to this bias. In particular, we show how to extend *LineageOT* [1] to use data from clones observed across multiple time points.

1 Introduction

The development of high-throughput single-cell RNA-sequencing technologies has motivated a flurry of work on computational methods for inferring developmental trajectories from static snapshots [2]–[17] and, interconnected and in parallel, methods for inferring cell lineage from DNA barcodes [18]–[25]. Developmental trajectories and cell lineage contain complementary information; having both data sources is valuable for trajectory inference [1], [18], [23], [26]–[31].

In this work, we identify a fundamental statistical bias that emerges from sampling cell lineage barcodes across a time course. We consider the setting in which cell state and lineage barcodes may be measured simultaneously (i.e., for the same cells) [20], [32], and copies of the same lineage barcodes may be observed over multiple time points (e.g., by repeated sampling from the same population). While these data have been collected and analyzed for some time [23], [25], [33], this bias has not yet been reported.

We focus on the case of static lineage barcodes, which are incorporated into the genome at an early stage of the developmental process, and then inherited by daughter cells but not modified over time. Note that lineage barcodes may not be detected in all cells that are sampled, resulting in barcoded datasets containing a subset of cells where cell state is known but the lineage barcode is not. Static lineage barcodes allow the tracing of *clones*: groups of cells that share a common barcode and therefore all share a common ancestor at the time of barcoding. If a clone is observed at only a single time point we will refer to it as a *single-time clone*, otherwise we will refer to it as a *multi-time clone*. The distinction between multi-time and single-time clones is important for understanding the source of the bias we discuss in this paper, and codifies clones by the level of information that they provide on developmental trajectories. Multi-time clones are particularly valuable for trajectory inference because they contain direct observations of developmental transitions. Indeed, recent methods including *CoSpar* [31] are designed to leverage multi-time clonal barcodes to infer cell fates.

We present a mathematical analysis that proves that the relative abundance of subpopulations is changed, or biased, in multi-time clonal datasets. The source of the bias is heterogeneous growth rates; cells with more descendants are more likely to be represented in multi-time clones. Therefore, more proliferative subpopulations are over-represented in multi-time clonal datasets. We prove the existence of this effect by simple analysis of probabilities, validating our arguments with proportions obtained from simulations as well

as real data. We further show that the performance of trajectory inference methods such as *CoSpar*, which rely on this biased information, may be negatively impacted by the presence of this sampling bias.

Knowledge of this statistical phenomenon can help guide the design of robust trajectory inference methods. We introduce an extension of *LineageOT* [1], called *LineageOT-MT*, which incorporates information from multi-time clonal barcodes in addition to that from single-time clonal barcodes. We test the method on a barcoded time course of T-cell induction (see separate paper [34]), and also simulated datasets for which the ground-truth is accessible. We find that this extension offers a robust improvement over the original *LineageOT* and that it may be negatively impacted by the presence of the sampling bias. Comparing the performance of *CoSpar* and *LineageOT-MT* in the presence of the sampling bias, we provide recommendations for which method may be more robust in which settings.

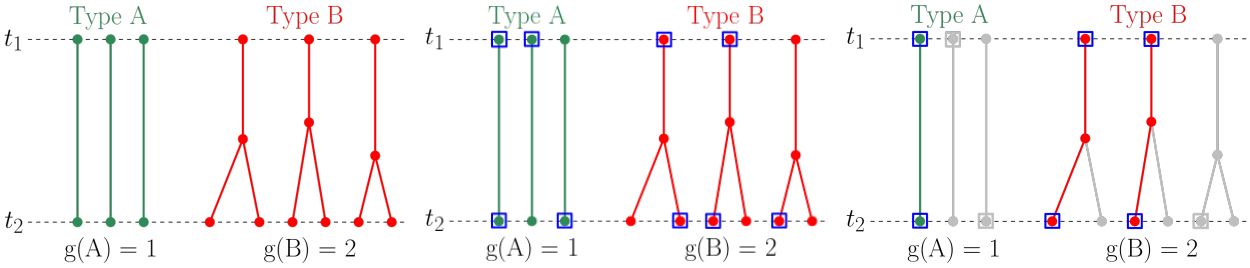
The remainder of the paper is organised as follows. In Section 2 we present a mathematical derivation of the sampling bias. In Section 3 we demonstrate that the performance of trajectory inference methods may be negatively impacted by analyzing predictions from *Cospar* in the presence of this bias. In Section 4 we introduce *LineageOT-MT*, demonstrate improved performance over the original *LineageOT*, and repeat the analysis of *Cospar* in the presence of the sampling bias for *LineageOT-MT*. We conclude in Section 5 with a Discussion, where we explore how state-of-the-art methods for lineage-informed trajectory inference may be developed to overcome the impact of this sampling bias. Given the potential for the application of trajectory inference results to biomedical technologies and treatments, understanding and improving the accuracy of these methods is of critical importance.

2 Cell growth introduces bias in multi-time clonal barcodes

In this section, we present a derivation of the statistical sampling bias in multi-time clone data.

Consider a population of cells comprised of two subpopulations denoted type A and type B, where type B cells have a higher rate of proliferation than those of type A. Each cell is given a unique barcode at some initial time t_0, and these barcodes are inherited by daughter cells with each cell division. After barcoding, the population develops over time and is then sampled at t_1 and t_2. Each sample is sequenced via scRNA-seq to determine the state (gene expression) of each cell as well as, if it is detected, the lineage barcode.

We are interested in how cell types with different growth rates are represented in multi-time clones, where the multi-time clones are constructed from sampling independently at t_1 and t_2 as illustrated by the example in Figure 1. Note that while we analyze a single pair of time points, the bias we discover may be greater for datasets with more than two time points (see Appendix A.2).



(a) Total population of cells at t_1 and t_2, illustrated as lineage trees.

(b) Step 1. Sample randomly at t_1 and t_2.

(c) Step 2. Remove cells in clones not sampled at both time points.

Figure 1: **Example demonstrating the sampling process for obtaining the set of cells in multi-time clones:** An illustration of the sampling process with two cell types, A and B, where A has growth 1, B has growth 2 and all lineages start with a single-cell at t_1. The blue squares indicate cells sampled in step 1., and the grey lineages in (c) indicate the cells/lineages removed in step 2. The set of cells remaining in (c) are those in multi-time clones, denoted MT.

More precisely, for this derivation we will consider the set of cells in multi-time clones, denoted MT, which results from the sampling process as follows. Assume we have a population of N_{t_1} cells at t_1 and N_{t_2} cells at t_2. Each cell x at t_1 gives rise to $g(x)$ descendants at t_2, as in the example shown in **Figure 1(a)**. Denote the set of cells in the population at t_1 and t_2 by $\mathcal{M}(t_1)$ and $\mathcal{M}(t_2)$ respectively. Using this notation, the sampling process may be described as:

1. Sample cells independently at t_1 and t_2, as illustrated in **Figure 1(b)**. Denote the sets of cells sampled by $\mathcal{S}(t_1) \subset \mathcal{M}(t_1)$ and $\mathcal{S}(t_2) \subset \mathcal{M}(t_2)$.
2. Remove cells for which a lineage barcode was not detected from each sample. Denote the sets of remaining cells in each sample as $\mathcal{B}(t_1) \subset \mathcal{S}(t_1)$ and $\mathcal{B}(t_2) \subset \mathcal{S}(t_2)$. Remove cells in clones observed at only one time point (single-time clones), as in **Figure 1(c)**. We denote the remaining cells by $MT \subset (\mathcal{B}(t_1) \cup \mathcal{B}(t_2))$. These are the cells in multi-time clones.

We now analyze the proportions of cell types in MT and find that these are different from the true abundances in the population at time t_1. Similar derivations also show a bias in the abundances at t_2 (see Appendix A.1), and, for time series with more than two time points, a bias in the abundances for each sample time (see Appendix A.2).

We analyze the proportions of cells at time t_1 in MT by calculating the probability that a cell from the original population will be retained in MT. Denote the *sampling rate* (the proportion of cells sampled) by $r_{t_i} \in (0,1)$ for $i = 1, 2$. Let $b \in [0,1]$ denote the probability that a lineage barcode is detected in a cell. We will refer to b as the *barcode rate*; it represents a combined rate of barcode insertion and detection. Denote the *net growth* between t_1 and t_2 of a single-cell x at t_1 by $g(x)$. This means that $g(x)$ is the number of descendants of cell x at time t_2, where $g(x) = 0$ if x dies between t_1 and t_2. We will show that **restricting to multi-time clones can bias the proportion of cells in a way that depends on the growth $g(x)$ of the cells in the clone, sampling rates r_{t_1} and r_{t_2}, and barcode rate b.**

We now derive an expression for the probability that a cell x at time t_1 will be in MT. To simplify the expression of the probabilities, we assume that the two samples are independent and that we are sampling with replacement. Let N_{t_i} denote the number of cells in the population at time t_i for $i = 1, 2$. Define a clone membership function C by $C(x) = c$ if cell x is in clone c. Finally, let $\tilde{g}(c)$ be the sum of the growth rates of the cells at t_1 in clone c, i.e., $\tilde{g}(c) = \sum_{x\,:\,C(x)=c} g(x)$. Then, for any cell x at time t_1, the probability that x is retained in MT is

$$\begin{aligned}
\mathbb{P}(x \in MT) &= \mathbb{P}(x \in \mathcal{B}(t_1))\mathbb{P}(\text{at least one cell in } C(x) \text{ is in } \mathcal{B}(t_2)) \\
&= \mathbb{P}(x \in \mathcal{B}(t_1))(1 - \mathbb{P}(\text{no cells in } C(x) \text{ are in } \mathcal{B}(t_2))) \\
&= \mathbb{P}(x \in \mathcal{B}(t_1)) \left(1 - \prod_{i=1}^{\tilde{g}(C(x))} \mathbb{P}(i\text{-th cell in } C(x) \cap \mathcal{M}(t_2) \text{ is not in } \mathcal{B}(t_2))\right) \\
&= r_{t_1} b \left(1 - \prod_{i=1}^{\tilde{g}(C(x))} (1 - r_{t_2} b)\right),
\end{aligned}$$

where the first line follows from our independent samples assumption, and the last line follows from approximating by sampling with replacement. Simplifying, we obtain

$$\mathbb{P}(x \in MT) = r_{t_1} b \left(1 - (1 - r_{t_2} b)^{\tilde{g}(C(x))}\right). \tag{1}$$

Equation 1 shows how the probability that a cell at t_1 will be in a sampled multi-time clone depends on the growth rate of its clone relative to the other clones in the population at t_1.

One may use Equation 1 to quantify the size of the bias effect on the proportions of cells from specific clusters in MT (e.g., cell types). See Equation 2 for an example of using Equation 1 to derive predicted cell type proportions in MT.

Next, we analyze how the proportions of cell types are impacted by this sampling bias. Consider the simplified setting in which: (1) each clone contains cells of only one cell type at t_1, (2) the growth rate is homogeneous across each cell type (i.e. $g(x) = g(l)$ for all cells x of type l at t_1) and (3) all clones of type l start with the same number of cells at t_1, denoted $m(l)$. Under these simplifying assumptions, we have $\tilde{g}(C(x)) = m(l)g(l)$ for every cell x of type l. Then the bias in proportions of cell types at t_1 within MT is revealed by the following conditional probability,

$$\begin{aligned}\mathbb{P}(x \text{ is type } l \mid x \in MT) &= \frac{\mathbb{P}(x \in MT \text{ and is type } l)}{\mathbb{P}(x \in MT)} \\ &= \frac{r_{t_1} b \left(1 - (1 - r_{t_2} b)^{m(l)g(l)}\right)) \frac{N_{t_1}(l)}{N_{t_1}}}{\sum_l r_{t_1} b \left(1 - (1 - r_{t_2} b)^{m(l)g(l)}\right)) \frac{N_{t_1}(l)}{N_{t_1}}} \end{aligned} \tag{2}$$

where the sum is over all possible cell types l. It is clear from Equation 2 that the proportions of the cell types at t_1 in MT may significantly differ from the true proportions. For example, consider a population with two cell types A and B where $g(A) = 1$, $g(B) = 2$, $m(A) = 1$ and $m(B) = 1$. Suppose that the proportions of type A and B at t_1 are equal, and at each time point we sample 50% of the population and detect a lineage barcode in every cell that we sample. Then, by Equation 2, the expected proportions of the two cell types A and B at t_1 in MT are 2/5 and 3/5 respectively, in contrast to the 50:50 ratio of A:B that is present in the total sample at t_1, as well as in the cells observed in single-time clones at t_1. These proportions will depend on the growth rates, $g(l)$ and $m(l)$, of each cell type, the rates r_{t_1} and r_{t_2} of cell capture, and the barcoding efficiency b.

The derivations in Appendix A.1 show that the cell type proportions at t_2 in MT are similarly biased. However, at t_2 the probabilities depend on growth rates between t_0 and t_1 rather than between t_0 and t_2 as in Equation 2. In this work, we focused our attention on the impact of the bias in the t_1 proportions as a representative example of the effect.

To support our probabilistic reasoning, we conduct a simulation study to validate Equation 2. See Figure 2 for the results, where we plot the true simulated proportions and the proportions we predict using Equation 2. The plot shows the cell type proportions in MT for a population with two cell types, denoted A and B, with $N_{t_1} = 2000$, $m(A) = 2$ and $m(B) = 4$, $g(A) = 1$ and $g(B) = 2$, and varying sample rate $r \in [2, 100\%]$. In this example, the true proportions of type A and B at t_1 are equal. The plot clearly shows that the simulated results concentrate around the predictions and that the bias in the proportions increases at a superlinear rate as sampling rate r decreases, as expected from Equation 2. As sampling at each time is without replacement in these simulations, this provides evidence that approximating by sampling with replacement in the derivation of Equations 1 and 2 is reasonable for estimating the t_1 cell type proportions in MT. The plot demonstrates that as sampling and barcode detection rates decrease, the bias in the cell type proportions in MT may diverge significantly from their true values.

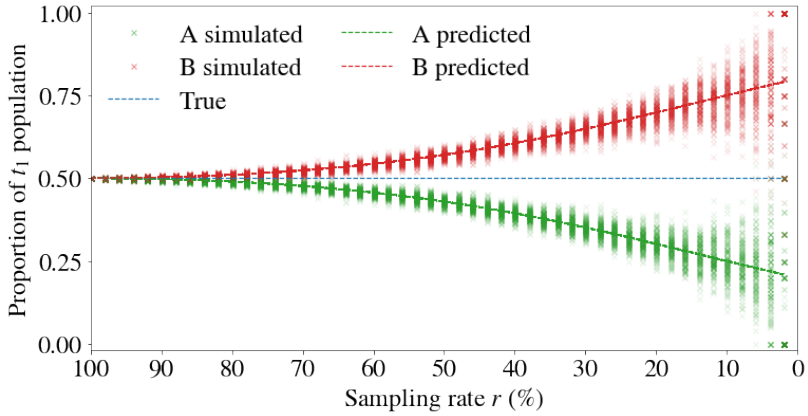


Figure 2: **Restricting to multi-time clones introduces bias in cell type proportions**. Simulated and predicted proportions of cell types A and B at t_1 are shown as a function of the barcode sampling rate (combined sampling and barcode rate). The predicted proportions are computed from Equation 2 with $N_{t_1} = 2000$ cells, $m(A) = 2$, $m(B) = 4$, $g(A) = 1$ and $g(B) = 2$. In this example, the true proportions of type A and B at t_1 are equal. The bias in the proportions increases as sampling rate r decreases, where $r = r_{t_1} b = r_{t_2} b$ is plotted for 50 values between 100-2%. 200 replicates of the sampling process were generated for each of the 50 values.

Our investigations into this bias effect were motivated by a time course of statically barcoded blood cell development published in a separate paper [34]. We include here a summary (Figure 3) of the sampled cell type proportions in different subsets of clones: all (i.e., cells where a barcode was detected), multi-cell, and multi-time clones. *Multi-cell clones* are those that contain more than a single cell.

From our derivations, we do not expect biased cell type proportions in the set of cells where a barcode was detected or those in a multi-cell clone. Indeed, in Figure 3 the difference in proportions between the total sample and these two subsets is within what one would expect from unbiased random sampling. In contrast, the difference between the proportions in the total sample and those in multi-time clones is outside the expected variance of unbiased sampling. The bias we have identified may explain this difference and be driven by a higher relative growth rate of progenitor cells. Conclusively determining that this is the cause would require growth and death rate estimates that were not available for this dataset. This rate information is needed to eliminate the alternative explanation that the lineages of cell types with decreased proportions (e.g., myeloid at $t = 5$) died out by $t = 10$ at a significantly greater frequency than those of increased proportions (e.g., progenitor at $t = 5$).

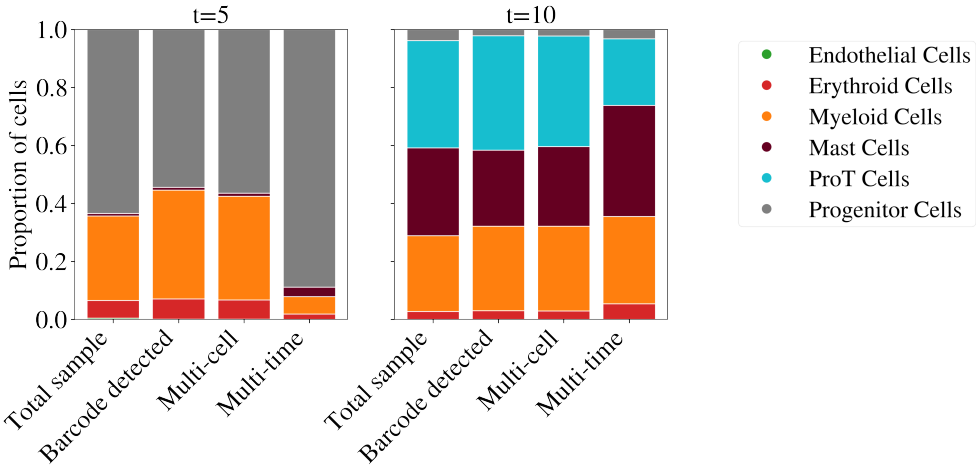


Figure 3: **Sampled cell type proportions in a time course of statically barcoded blood cell development differ when restricting to the subset of cells in multi-time clones**. The proportions of different cell types observed at two time points ($t = 5$ and $t = 10$) in different subsets of clones (barcode detected, multi-cell and multi-time) are shown as coloured areas in bar plots. The proportions at each time for the total sample, barcode detected and multi-cell clones are similar (variance within that expected from random sampling), whereas the cell type proportions in the multi-time clones are significantly different at both time points. In particular, at $t = 5$, the proportion of progenitor cells is approximately 0.9 instead of being in the range 0.5-0.7, and the myeloid proportion is approximately 0.1 instead of being in the range 0.3-0.4. The difference for the proportions at $t = 10$ is smaller, but the proportions in the multi-time clones are still the furthest from those of the total sample.

We conclude with four summary remarks related to this statistical phenomenon: (1) single-time lineage information exhibits no such bias, (2) there will be no bias when growth rates are equal across the population for all time, (3) the bias can present for clones sampled at just two time points, but it may increase with sampling at additional times (see Appendix A.2) and, (4) the cell type proportions in MT at each sample time may be biased depending on the growth rates in a similar manner as in Equation 2. These remarks and the results of this section apply to the proportions of any subdivision/clustering of cells into subpopulations that correlate with growth rates, cell types being one prominent example.

3 Bias in multi-time clonal barcodes can affect trajectory inference

Given this newly discovered sampling bias in multi-time clonal barcodes, a natural question is whether this bias impacts the current state-of-the-art methods which use this information for trajectory inference.

We investigate the answer to this question for *CoSpar*, a method that has shown promising results for datasets representing hematopoiesis, reprogramming and directed differentiation [31]. *CoSpar* (coherent, sparse optimization) features two different methods for lineage-informed trajectory inference. Both methods use cell state information, but one uses only single-time clonal data while the other uses only multi-time clonal data. We will refer to the latter as *CoSpar-MT*. The method is built on two key assumptions about trajectory couplings: (1) the coupling is 'locally coherent', and (2) the coupling is sparse. These two assumptions are built into the *CoSpar* method via a 'smoothing' and a thresholding step respectively [31].

We focus on the impact of the bias on the inferred fate probabilities of the cells, a key prediction derived from trajectory couplings. Our results demonstrate that it is possible for the bias in the multi-time clone data to negatively impact the performance of *CoSpar-MT* on the task of fate probability prediction (see

Figure 4). The results were gathered for a simple simulated population with two cell types at t_1 and t_2 for which the ground truth fate probabilities are known.

Two variations, *baseline* and *effect*, of the population were simulated to separate the effect of the bias in the multi-time clones from the effect of the sampling rate. In the *baseline*, all growth rates are equal resulting in no bias effect. In the *effect* case, the growth rates vary between two and eight descendants across the population, resulting in a bias in the t_1 cell type proportions in the multi-time clones. The growth rate distribution includes a mixture of higher and lower growth cells in one of the cell types, which could represent cells with a distinct chromatin accessibility profile (between t_1 and t_2) to the rest of the cluster.

To control for the effect of the number of cells sampled and manage the computational resources required for trajectory inference, 1000 cells were sampled at t_1 and t_2 for every sampling rate. Sampling a fixed number of cells rather than sampling at a fixed rate also more realistically models experimental procedures, in which only a fixed budget of cells can be sequenced. Each specified t_1 sampling rate, r_{t_1}, is obtained by simulating a population of N_{t_1} cells where $N_{t_1} = \lceil 1000/r_{t_1} \rceil$. For the *effect* and *baseline* cases each, the results were generated from 30 simulated populations (one for each t_1 sampling rate tested), each containing between 2500 and 60 000 cells with population size scaling inversely to sampling rate. The true trajectory coupling was computed for each population, and then 1000 cells were sampled at each t_1 and t_2. Next, *CoSpar-MT* was used to estimate the trajectory coupling between the t_1 and t_2 samples.

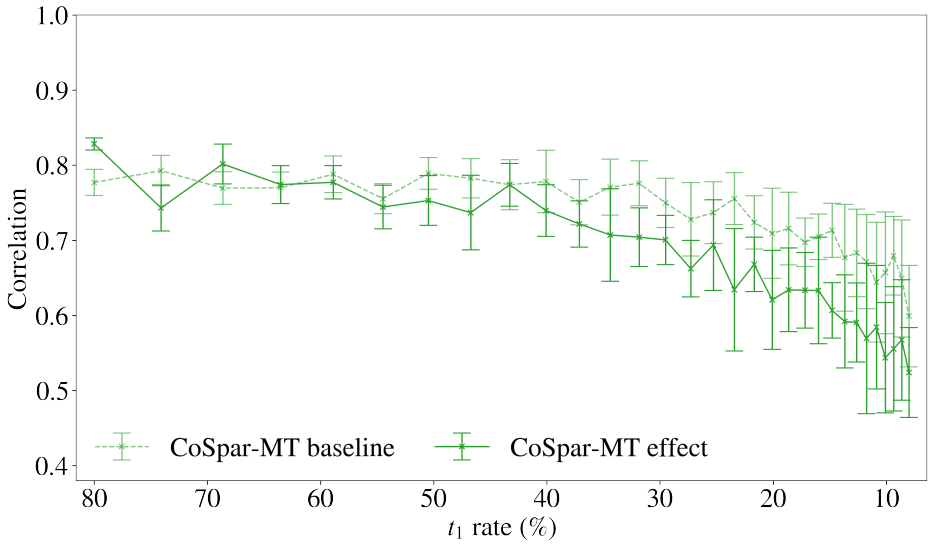


Figure 4: **It is possible for the bias to impact fate probability predictions from CoSpar-MT:** Pearson correlation coefficient between the true fate probability and the fate probability estimated from *CoSpar-MT* (the probability that each cell is fated to be type A at t_2) as a function of t_1 rate (combined sampling and barcode rate $r_{t_1}b$) for $b = 0.8$ fixed for the *baseline* and *effect* cases. Values are summarized as a mean across ten replicates with error bars given by the standard deviation. The plot shows that the mean correlation decreases as the sampling rate decreases and that once the bias effect is sufficiently large (around $r_{t_1}b$=40%) the mean correlation is lower in the *effect* case than the *baseline* case with a difference of approximately 0.1. This demonstrates that it is possible for the bias in multi-time clones to negatively impact fate probability prediction. The results for fate B are identical and hence are omitted.

The results in Figure 4 show that the mean correlation decreases as the sampling rate decreases, and furthermore, once the bias effect is sufficiently large (around $r_{t_1}b$=40%) the mean correlation is lower in the *effect* case than the *baseline* case, with a difference of approximately 0.1. Since the error bars of the *baseline* and *effect* results overlap, we do not claim that correlation is lower in the *effect* case with high probability, but the trend of the mean is clear. This demonstrates that it is possible for the bias in multi-time clones to negatively impact fate probability predictions from state-of-the-art methods.

4 *LineageOT-MT*: Extension of LineageOT to use multi-time clonal data

LineageOT [1] is an extension of the state-based *WaddingtonOT* framework [12] to incorporate lineage information from single-time lineage trees (trees where all observed cells are from a single time point). The current *LineageOT* method neglects to use valuable lineage information encoded in multi-time clonal barcodes, as is available in static lineage barcoding with population subsampling across time points. This limitation motivated our extension of *LineageOT* to use *both* multi-time and single-time clonal barcode information. We refer to this extension as *LineageOT-MT*[1]. Our approach in *LineageOT-MT* differs from the approach in *CoSpar* where either multi-time (as in *CoSpar-MT*) *or* single-time clonal barcodes are used but *not both*.

The extension requires no new mathematical theory and only a relatively small change in the implementation of the algorithms. We simply condition on the state of all observed cells in each clone in the ancestor estimation step in *LineageOT*, rather than just those at the later time point. This estimation is performed by modeling the lineage trees as Gaussian graphical models relating the cell states [1]. In the extension, we condition in the Gaussian model on all cells observed in the lineage tree (over all time points), not just those sampled from the later time point being coupled. The Gaussian estimation now benefits from using the observed states of ancestral and present-time clonal relatives in addition to the descendant relatives.

Trajectory inference results from simulated datasets show that multi-time clonal information improves the performance of *LineageOT*. Figure 5 demonstrates the improvement possible for the task of fate probability prediction. The results from this figure describe prediction performance on a simulation scenario similar to those used in the analysis of *CoSpar* in Section 3. Additional such comparative results were gathered for growth rate variations of this scenario and the simulation scenarios described in Section 3, using over 1200 simulated datasets in total (ten replicates per each of the 120 variations). *LineageOT-MT* achieved an equal or higher mean correlation with the true fate probability than *LineageOT* on all of these datasets.

The results in Figure 5 show that at the highest rate tested of 80% our extension of the method increases correlation with the true fate probability from approximately 0.6 to over 0.9. The performance of *LineageOT-MT* over *LineageOT* was assessed on simulated datasets across a range of sampling rates, as shown in Figure 5, to illustrate the effect of decreasing the available multi-time clonal barcode information on prediction performance. As the sampling rate decreases, the decline in performance of *LineageOT-MT* is approximately linear and converges towards the performance of *LineageOT*. This is expected since the amount of multi-time clonal information is decreasing to zero with the sampling rate.

LineageOT-MT incorporates additional information about the developmental process compared to *LineageOT*, so these results are not surprising. However, the size of the boost in performance is encouraging for further development in this direction.

[1]The implementation of *LineageOT-MT* is publicly available on a fork of the *LineageOT* package at: https://github.com/rbonhamcarter/LineageOT/tree/multi-time-clones.

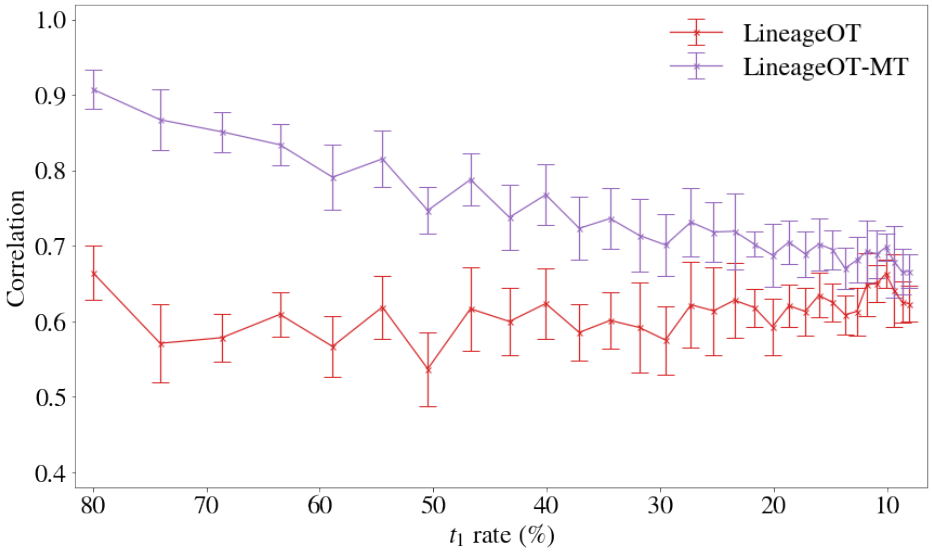


Figure 5: **Multi-time information improves the fate probability predictions from LineageOT:** Pearson correlation between the true fate probability and the fate probability estimated from *LineageOT* and *LineageOT-MT* for fate A at t_2 as a function of t_1 rate (combined sampling and barcoding rate $r_{t_1}b$) for $b = 0.8$. Values are summarized as a mean across ten replicates with error bars given by the standard deviation. The results for fate B are identical and hence are omitted.

4.1 Effect of bias in multi-time clonal barcodes on *LineageOT-MT*

In this section, we explore the impact of multi-time clonal barcode bias on *LineageOT-MT*. This exploration is analogous to the study of the effect of the bias on *CoSpar-MT* in Section 3.

The results were gathered using *baseline* and *effect* variations (as defined in Section 3) of the simulation scenario that was used to benchmark the performance of *LineageOT-MT* above. The performance of *LineageOT-MT* over decreasing sample rates for these two variations is shown in Figure 6. This figure demonstrates that the mean correlation decreases as the sampling rate decreases. Furthermore, for a relatively high range of sampling rates (from t_1b =0.8 to around t_1b =0.3), the mean correlation is lower in the *effect* case than in the *baseline* case, with a difference of approximately 0.1. These results demonstrate that it is possible for the bias in multi-time clones to negatively impact fate probability predictions from *LineageOT-MT*.

The lack of impact on *LineageOT-MT* at rates less than 0.3, where the size of the bias effect is larger, may be explained by the methodological 'convergence' of *LineageOT-MT* to *LineageOT* as the sampling rate decreases. This convergence corresponds to a shift from using *multi-time* and *single-time* lineage information to only using *single-time* lineage information. The number of cells observed in multi-time clones decreases with the sampling rate at a faster rate than the number in single-time clones, resulting in this shift occurring before the sampling rate nears zero. Recalling that the single-time information is not biased, this convergence may counter the impact of the bias on cell fate predictions from *LineageOT-MT*.

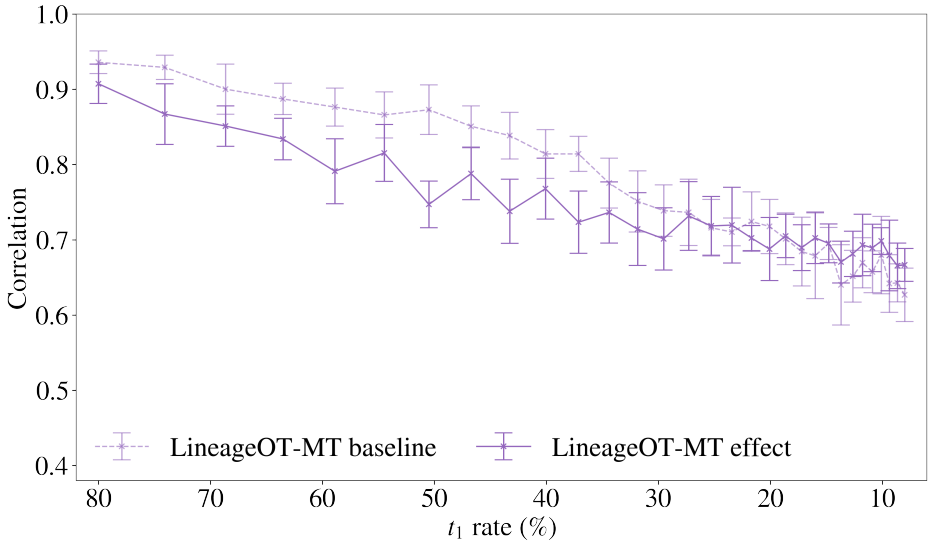


Figure 6: **It is possible for the bias to impact fate probability predictions from LineageOT-MT:** Pearson correlation between the true fate probability and the fate probability estimated from *LineageOT-MT* (the probability each cell is fated to be type A at t_2) as a function of t_1 rate (combined sampling and barcode rate $r_{t_1}b$) for $b = 0.8$ fixed for the *baseline* and *effect* cases. Values are summarized as a mean across 10 replicates with error bars given by the standard deviation. The plot shows that the mean correlation decreases as the sampling rate decreases, and that for sufficiently *high* sampling rate (greater than approximately $r_{t_1}b$=30%) the mean correlation is lower in the *effect* case than the *baseline* case, with a difference of approximately 0.1. This demonstrates that it is possible for the bias in multi-time clones to negatively impact fate probability prediction. The results for fate B are identical and hence are omitted.

5 Discussion

In this work, we consider the problem of developmental trajectory inference using lineage information derived from multi-time clonal barcodes. We have shown that in certain experimental settings, there may exist a bias in the proportions of cell types represented in the multi-time clonal barcode data. This sampling bias emerges for clones sampled at just two time points and may increase with sampling at additional times. The experimental settings where this bias emerges are those with heterogeneous growth rates and a combined sampling and barcode detection rate of less than 100% for at least one sample in the time course. These are of course very general conditions, and we suspect this bias to be present in most multi-time clonal barcode datasets.

In addition, we showed that it is possible for this bias to negatively impact predictions made from inferred developmental cell trajectories. This was demonstrated on a state-of-the-art trajectory inference method, *CoSpar* [31], using simulated datasets for which the ground truth is known.

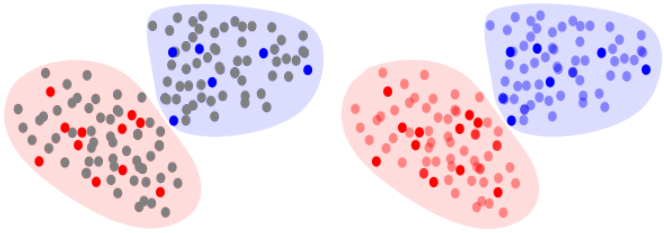
Our third contribution is an extension of the *LineageOT* trajectory inference method [1] to additionally use multi-time clonal barcode information. This extension, *LineageOT-MT*, robustly improves the performance of the method. We then showed that it was possible for the bias to impact predictions made from trajectories inferred using *LineageOT-MT*. We note that there may be more sophisticated ways to integrate the multi-time clone data into the *LineageOT* method beyond our present approach that could yield further performance improvements and robustness.

Our studies on the impact of the sampling bias on fate probability prediction with *CoSpar-MT* and *LineageOT-MT* show that each method is sensitive to the bias for distinct scenarios. Certain features of each method mediate the robustness or sensitivity to the bias effect. Qualitatively, the smoothing step

in *CoSpar-MT* (which is strongly dependent on the state space geometry of the sampled cells) seems to improve the robustness of the method to the bias effect in some settings. For *LineageOT-MT*, using cell state and single-time clonal barcode information in addition to multi-time appears to provide robustness to the bias effect. This conclusion is supported by these first two data sources being unbiased and the prediction accuracy of the method being unaffected in the low sampling rate regime. Purely considering the trends observed with respect to sampling rate, *LineageOT-MT* may offer robustness to the bias effect in the lower sampling rate regime while *CoSpar* may offer robustness at higher rates.

Multi-time clonal data, even if known to be biased by this sampling effect, is too valuable not to leverage in the study of single-cell development. This point is evidenced in this paper by the robust improvement of *LineageOT-MT* over the original *LineageOT*. What is needed is a solution for correcting the bias. Our work suggests that the most promising avenue is to leverage the unbiased single-time clonal barcode and cell state data.

We observe that the smoothing step of the *CoSpar* algorithm could be considered a kind of label propagation or 'pseudo-labeling' [35]–[37] that leverages the unbiased cell state data. The smoothing effectively propagates the multi-time clone label information from cells to nearby cells (in state space) that were not observed in a multi-time clone. See Figure 7 for an illustration of the effect of label propagation on a dataset. Label propagation is one form of semi-supervised learning that can be applied to barcoded single-cell datasets. *Semi-supervised learning* refers to inference in settings where only a subset of the training data is labeled. This is the case in our application, where some cells are labeled (by their clone) and some are unlabeled as a lineage barcode was not detected. For any label propagation approach, it will be important to account for growth rates in the propagation to generate labels that correct the bias in the multi-time clones.



(a) Raw data (b) Post label propagation.

Figure 7: **An illustration of label propagation:** The left image illustrates a raw partially labeled dataset where the dots are data points coloured by their label. The grey dots are the unlabeled points and the background colour shows the true distribution of each cluster. The right image shows the same dataset after performing successful label propagation to share labels from the labeled points to the unlabeled points, resulting in a fully labeled dataset.

Motivated by the concept of pseudo-labeling and the unbiased nature of single-time clonal data, one idea to improve the robustness of *CoSpar-MT* to the bias is to correct the multi-time clonal matrix in the algorithm [31] using iterative proportional fitting (IPF). IPF is also known as Sinkhorn's algorithm in the optimal transport literature. It is a procedure for fitting data to a set of marginals which are typically known to be more accurate than the those of the data itself. In the case of *CoSpar-MT*, the two marginals of the multi-time clonal matrix correspond to the proportion of clones present in each subpopulation at $t1$ and t_2. The idea is to use IPF to project the matrix to the subspace of matrices with the "correct" marginals, i.e. those given by the unbiased single-time clonal data. Further investigation is required to determine whether this correction would indeed eliminate the impact of the bias in any predictions made using *CoSpar-MT*.

The equations in Section 2, and studies on *CoSpar-MT* and *LineageOT-MT* in the subsequent sections, provide some indication of when the bias effect will impact trajectory inference. In the future, it would be useful to have a statistical test that quantifies the expected impact of the bias on trajectory inference. A practical statistical test should use only information available in standard datasets of lineage barcoded sequencing time courses, which importantly may not include precise estimates of cell growth rates. An-

other important step is the definitive identification of the bias effect and its impact on trajectory inference methods for a real dataset. We leave such developments to future work, having focused in this work on the mathematical details and validations only possible using simulated data.

We hope that our identification of this statistical phenomenon serves to inform practitioners using multi-time clonal barcodes of possible biases in the data, and guide future method development. Beyond their use in proving the existence of the bias, the probabilities derived in Section 2 can also be used as a tool by practitioners for approximating the size of the bias effect in a dataset. Our contributions beyond these points are our extension to *LineageOT*, and the above discussion of how methods might be developed to overcome the impact of this sampling bias. Such development is an important endeavour to allow practitioners to make use of the valuable information in multi-time clonal barcodes, without these concerns of biased results and conclusions.

Acknowledgements

Funding: This work was supported by a New Frontiers in Research Fund Exploration Grant, and a Natural Sciences and Engineering Research Council Discovery Grant.

Conflicts of Interest: None to declare.

References

[1] A. Forrow and G. Schiebinger, "LineageOT is a unified framework for lineage tracing and trajectory inference," *Nature communications*, vol. 12, no. 1, pp. 1–10, 2021.

[2] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys, "A comparison of single-cell trajectory inference methods," *Nature biotechnology*, vol. 37, no. 5, pp. 547–554, 2019.

[3] S. Zhang, A. Afanassiev, L. Greenstreet, T. Matsumoto, and G. Schiebinger, "Optimal transport analysis reveals trajectories in steady-state systems," *PLoS computational biology*, vol. 17, no. 12, e1009466, 2021.

[4] C. Lin and Z. Bar-Joseph, "Continuous-state hmms for modeling time-series single-cell rna-seq data," *Bioinformatics*, vol. 35, no. 22, pp. 4707–4715, 2019.

[5] H. Chen, L. Albergante, J. Y. Hsu, *et al.*, "Single-cell trajectories reconstruction, exploration and mapping of omics data with stream," *Nature Communications*, vol. 10, no. 1, p. 1903, 2019.

[6] T. N. Tran and G. D. Bader, "Tempora: Cell trajectory inference using time-series single-cell rna sequencing data," *PLoS Computational Biology*, vol. 16, no. 9, e1008205, 2020.

[7] M. Setty, V. Kiseliovas, J. Levine, A. Gayoso, L. Mazutis, and D. Pe'er, "Characterization of cell fate probabilities in single-cell data with palantir," *Nature Biotechnology*, vol. 37, no. 4, pp. 451–460, 2019.

[8] A. Tong, J. Huang, G. Wolf, D. Van Dijk, and S. Krishnaswamy, "Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics," in *International conference on machine learning*, PMLR, 2020, pp. 9526–9536.

[9] G. H. T. Yeo, S. D. Saksena, and D. K. Gifford, "Generative modeling of single-cell time series with prescient enables prediction of cell trajectories with interventions," *Nature Communications*, vol. 12, no. 1, p. 3222, 2021.

[10] C. Weinreb, S. Wolock, B. K. Tusi, M. Socolovsky, and A. M. Klein, "Fundamental limits on dynamic inference from single-cell snapshots," *Proceedings of the National Academy of Sciences*, vol. 115, no. 10, E2467–E2476, 2018.

[11] S. Tritschler, M. Büttner, D. S. Fischer, *et al.*, "Concepts and limitations for learning developmental trajectories from single cell genomics," *Development*, vol. 146, no. 12, dev170506, 2019.

[12] G. Schiebinger, J. Shu, M. Tabaka, *et al.*, "Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming," *Cell*, vol. 176, no. 4, 928–943.e22, 2019, ISSN: 0092-8674.

[13] J. Cao, M. Spielmann, X. Qiu, *et al.*, "The single-cell transcriptional landscape of mammalian organogenesis," *Nature*, vol. 566, no. 7745, pp. 496–502, 2019.

[14] K. Street, D. Risso, R. B. Fletcher, *et al.*, "Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics," *BMC Genomics*, vol. 19, pp. 1–16, 2018.

[15] J. A. Farrell, Y. Wang, S. J. Riesenfeld, K. Shekhar, A. Regev, and A. F. Schier, "Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis," *Science*, vol. 360, no. 6392, eaar3131, 2018.

[16] F. A. Wolf, F. K. Hamey, M. Plass, *et al.*, "PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells," *Genome Biology*, vol. 20, pp. 1–9, 2019.

[17] M. Setty, M. D. Tadmor, S. Reich-Zeliger, *et al.*, "Wishbone identifies bifurcating developmental trajectories from single-cell data," *Nature Biotechnology*, vol. 34, no. 6, pp. 637–645, 2016.

[18] L. Kester and A. van Oudenaarden, "Single-cell transcriptomics meets lineage tracing," *Cell Stem Cell*, vol. 23, no. 2, pp. 166–179, 2018, ISSN: 1934-5909.

[19] A. McKenna and J. A. Gagnon, "Recording development with single cell dynamic lineage tracing," *Development*, vol. 146, no. 12, dev169730, 2019, ISSN: 0950-1991.

[20] B. Raj, D. E. Wagner, A. McKenna, *et al.*, "Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain," *Nature Biotechnology*, vol. 36, no. 5, pp. 442–450, Mar. 2018.

[21] C. S. Baron and A. van Oudenaarden, "Unravelling cellular relationships during development and regeneration using genetic lineage tracing," *Nature Reviews Molecular Cell Biology*, vol. 20, no. 12, pp. 753–765, 2019.

[22] R. C. Sakata, S. Ishiguro, H. Mori, *et al.*, "Base editors for simultaneous introduction of C-to-T and A-to-G mutations," *Nature Biotechnology*, vol. 38, no. 7, pp. 865–869, Jun. 2020.

[23] C. Weinreb, A. Rodriguez-Fraticelli, F. D. Camargo, and A. M. Klein, "Lineage tracing on transcriptional landscapes links state to fate during differentiation," *Science*, vol. 367, no. 6479, eaaw3381, 2020.

[24] S. Bowling, D. Sritharan, F. G. Osorio, *et al.*, "An engineered crispr-cas9 mouse line for simultaneous readout of lineage histories and gene expression profiles in single cells," *Cell*, vol. 181, no. 6, pp. 1410–1422, 2020.

[25] B. A. Biddy, W. Kong, K. Kamimoto, *et al.*, "Single-cell mapping of lineage and identity in direct reprogramming," *Nature*, vol. 564, no. 7735, pp. 219–224, 2018.

[26] R. B. Fletcher, D. Das, and J. Ngai, "Creating lineage trajectory maps via integration of single-cell RNA-sequencing and lineage tracing: Integrating transgenic lineage tracing and single-cell RNA-sequencing is a robust approach for mapping developmental lineage trajectories and cell fate changes," *BioEssays*, vol. 40, no. 8, p. 1 800 056, 2018.

[27] H. Zafar, C. Lin, and Z. Bar-Joseph, "Single-cell lineage tracing by integrating CRISPR-Cas9 mutations with transcriptomic data," *Nature Communications*, vol. 11, no. 1, p. 3055, Jun. 2020.

[28] D. E. Wagner and A. M. Klein, "Lineage tracing meets single-cell omics: Opportunities and challenges," *Nature Reviews Genetics*, vol. 21, no. 7, pp. 410–427, 2020.

[29] N. Prasad, "Optimal transport using GANs for lineage tracing," *Proceedings of Machine Learning Research*, vol. 119, 2020.

[30] M. Lange, Z. Piran, M. Klein, F. J. Theis, and M. Nitzan, "Mapping lineage-traced single-cells across time-points," in *Learning Meaningful Representations of Life (NeurIPS 2021)*, 2021.

[31] S.-W. Wang, M. J. Herriges, K. Hurley, D. N. Kotton, and A. M. Klein, "CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information," *Nature Biotechnology*, vol. 40, no. 7, pp. 1066–1074, 2022.

[32] B. Spanjaard, B. Hu, N. Mitic, *et al.*, "Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars," *Nature Biotechnology*, vol. 36, no. 5, pp. 469–473, 2018.

[33] K. Hurley, J. Ding, C. Villacorta-Martin, *et al.*, "Reconstructed single-cell fate trajectories define lineage plasticity windows during differentiation of human psc-derived distal lung progenitors," *Cell Stem Cell*, vol. 26, no. 4, pp. 593–608, 2020.

[34] Y. S. Michaels, M. C. Major, B. Bonham-Carter, *et al.*, "Time- and lineage-resolved transcriptional profiling uncovers gene expression programs and clonal relationships that underlie human T lineage specification," *Preprint in submission*, 2023.

[35] J. Kim, Y. Hur, S. Park, E. Yang, S. J. Hwang, and J. Shin, "Distribution aligning refinery of pseudo-label for imbalanced semi-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 567–14 579, 2020.

[36] R. He, J. Yang, and X. Qi, "Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6930–6940.

[37] F. Taherkhani, A. Dabouei, S. Soleymani, J. Dawson, and N. M. Nasrabadi, "Self-supervised Wasserstein pseudo-labeling for semi-supervised image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 267–12 277.

A Appendix

A.1 Derivation of biased proportions in MT at $t2$

We include below the derivation of probabilities of subpopulation proportions at t_2, which is analogous to the derivation included in the paper for t_1 (Equations 1 and 2). Derivations for time points beyond t_1 and t_2 would follow similar arguments.

As we derive below, the cell type proportions in MT (as defined in Section 2) at t_2 are biased by the relative cell growth between t_0 and t_1. Let $m(c)$ denote the number of cells in clone c at t_1, or equivalently, the net growth rate between t_0 and t_1 of the cell at t_0 which was barcoded as clone c. Then we have:

$$\begin{aligned}
\mathbb{P}(y \in MT \cap \mathcal{M}(t_2)) &= \mathbb{P}(y \in \mathcal{B}(t_2))\mathbb{P}(\text{at least one cell in } C(y) \text{ is in } \mathcal{B}(t_1)) \\
&= \mathbb{P}(y \in \mathcal{B}(t_2))(1 - \mathbb{P}(\text{no cells in } C(y) \text{ are in } \mathcal{B}(t_1))) \\
&= r_{t_2} b \left(1 - \prod_{i=1}^{m(C(y))} (1 - r_{t_1} b)\right) \\
&= r_{t_2} b \left(1 - (1 - r_{t_1} b)^{m(C(y))}\right) \qquad (3)
\end{aligned}$$

and, if we simplify the expression by assuming that all clones of type l have the same value of $m(\cdot)$, which we denote $m(l)$ as in the derivation of Equation 2, then we have

$$\begin{aligned}
\mathbb{P}(y \in \mathcal{M}(t_2) \cap MT \text{ and is type } l) &= \mathbb{P}(y \in \mathcal{M}(t_2) \cap MT \mid y \in \mathcal{M}(t_2) \text{ is type } l)\mathbb{P}(y \in \mathcal{M}(t_2) \text{ is type } l) \\
&= r_{t_2} b \left(1 - (1 - r_{t_1} b)^{m(l)}\right) \frac{N_{t_2}(l)}{N_{t_2}},
\end{aligned}$$

and therefore we obtain,

$$\begin{aligned}
\mathbb{P}(y \in \mathcal{M}(t_2) \text{ is type } l \mid y \in \mathcal{M}(t_2) \cap MT) &= \frac{\mathbb{P}(y \in \mathcal{M}(t_2) \cap MT \text{ and is type } l)}{\mathbb{P}(y \in \mathcal{M}(t_2) \cap MT)} \\
&= \frac{r_{t_2} b \left(1 - (1 - r_{t_1} b)^{m(l)}\right)) \frac{N_{t_2}(l)}{N_{t_2}}}{\sum_l r_{t_2} b \left(1 - (1 - r_{t_1} b)^{m(l)}\right)) \frac{N_{t_2}(l)}{N_{t_2}}}. \qquad (4)
\end{aligned}$$

These equations are very similar to those for t_1 (Equations 1 and 2) except here the probabilities depend on growth rates between t_0 and t_1 rather than between t_0 and t_2. Comparing Equation 4 to Equation 2, if the true cell type proportions at t_1 and t_2 are the equal then the bias in the proportions may be lesser or greater at t_1 than at t_2 (depending if $g(l) > 1.0$ or < 1.0 respectively) due to the exponent being $m(l)g(l)$ rather than $m(l)$.

A.2 Derivation of biased proportions in MT for an arbitrary number of time points

We include below the derivation of the probabilities of subpopulation proportions at any sample time $t_k \in \{t_1, t_2, \ldots, t_T\}$ in a time course with a finite but arbitrary number of time points $T \geq 2$. This derivation follows a similar argument to that for t_1 and t_2 for $T = 2$ in Section 2 and Appendix A.1 respectively.

Let x denote a cell alive at time t_k and let r_t denote the sampling rate at time $t \in \{t_1, t_2, \ldots, t_T\}$. Then, by the law of total probability and De Morgan's law, we have:

$$\begin{aligned}
\mathbb{P}(x \in MT \cap \mathcal{M}(t_k)) &= \mathbb{P}(x \in \mathcal{B}(t_k))\mathbb{P}(\cup_{j=1, j\neq k}^{T} \{\text{at least one cell in } C(x) \text{ is in } \mathcal{B}(t_j)\}) \\
&= \mathbb{P}(x \in \mathcal{B}(t_k)) \left(1 - \mathbb{P}(\cap_{j=1, j\neq k}^{T} \{\text{no cells in } C(x) \text{ are in } \mathcal{B}(t_j)\})\right).
\end{aligned}$$

Now let $\tilde{g}(C(x), t)$ denote the sum of the growth rates of all cells alive in clone $C(x)$ at time t. Then, by the same independence assumption made in Section 2, the probability of the intersection can be broken down as the following product,

$$\begin{aligned}\mathbb{P}(x \in MT \cap \mathcal{M}(t_k)) &= \mathbb{P}(x \in \mathcal{B}(t_k)) \left(1 - \prod_{j=1, j\neq k}^{T} \mathbb{P}(\text{no cells in } C(x) \text{ are in } \mathcal{B}(t_j))\right) \\ &= r_{t_k} b \left(1 - \prod_{j=1, j\neq k}^{T} \prod_{i=1}^{\tilde{g}(C(x), t_{j-1})} (1 - r_{t_j} b)\right). \end{aligned} \tag{5}$$

By Equation 5, since each factor in bounded in $[0, 1]$ (as they represent probabilities), as T increases the product over the time points will be non-increasing, and hence the size of $\mathbb{P}(x \in MT \cap \mathcal{M}(t_k))$ will be non-decreasing. To reason how this effects the biased proportions of subpopulations we consider the following probability for some cell type (or any subpopulation correlated with growth) l:

$$\mathbb{P}(x \in \mathcal{M}(t_k) \text{ is type } l \,|\, x \in \mathcal{M}(t_k) \cap MT) = \frac{\mathbb{P}(x \in \mathcal{M}(t_k) \cap MT \;\mid\; x \text{ is type } l) \frac{N_{t_k}(l)}{N_{t_k}}}{\sum_l \mathbb{P}(y \in \mathcal{M}(t_k) \cap MT \;\mid\; x \text{ is type } l) \frac{N_{t_k}(l)}{N_{t_k}}}, \tag{6}$$

where $\mathbb{P}(y \in \mathcal{M}(t_k) \cap MT \;\mid\; x \text{ is type } l)$ takes an analogous form to in Section 2 based off of Equation 5. From Equation 6 and the observation that $\mathbb{P}(x \in MT \cap \mathcal{M}(t_k))$ will be non-decreasing as T increases, the behaviour of the probability representing the biased proportion $\mathbb{P}(x \in \mathcal{M}(t_k) \text{ is type } l \,|\, x \in \mathcal{M}(t_k) \cap MT)$ as T increases will depend on how the denominator changes. More specifically, since $\frac{N_{t_k}(l)}{N_{t_k}}$ is constant for all l as T increases, it will depend on how the values of $\mathbb{P}(x \in \mathcal{M}(t_k) \cap MT \;\mid\; x \text{ is type } l)$ behave for all types l.

Let us consider the case in which $\mathbb{P}(x \in \mathcal{M}(t_k) \cap MT \;\mid\; x \text{ is type } l)$ only increases for type l (for all other types the probability stays constant as T increases). Then it is easy to prove with simple algebraic manipulation that the new expression (after the increase in T) on the right hand side of Equation 6 is greater than the previous expression. This case demonstrates that the size of the bias effect may increase as T increases beyond $T = 2$. In general, whether or not the bias effect will increase depends on complex interactions between the growth rates in the new periods of time in the time course, as summarized with Equation 6. This equation may be used to estimate the size of the bias effect as demonstrated using Equation 2 in Section 2.