

A statistical method for quantifying progenitor cells reveals incipient cell fate commitments

Received: 4 March 2023

Accepted: 19 January 2024

Published online: 20 February 2024

 Check for updatesShanjun Deng^{1,2}, Han Gong^{1,2}, Di Zhang¹, Mengdong Zhang¹ & Xionglei He¹✉

Quantifying the number of progenitor cells that found an organ, tissue or cell population is of fundamental importance for understanding the development and homeostasis of a multicellular organism. Previous efforts rely on marker genes that are specifically expressed in progenitors. This strategy is, however, often hindered by the lack of ideal markers. Here we propose a general statistical method to quantify the progenitors of any tissues or cell populations in an organism, even in the absence of progenitor-specific markers, by exploring the cell phylogenetic tree that records the cell division history during development. The method, termed targeting coalescent analysis (TarCA), computes the probability that two randomly sampled cells of a tissue coalesce within the tissue-specific monophyletic clades. The inverse of this probability then serves as a measure of the progenitor number of the tissue. Both mathematic modeling and computer simulations demonstrated the high accuracy of TarCA, which was then validated using real data from nematode, fruit fly and mouse, all with related cell phylogenetic trees. We further showed that TarCA can be used to identify lineage-specific upregulated genes during embryogenesis, revealing incipient cell fate commitments in mouse embryos.

Genetic diversity has long been a central concept for understanding the evolutionary dynamics in populations of organisms^{1,2}. Within a multicellular organism, the population of cells that execute a specific function is often derived from a set of progenitors, and the number of progenitors determines the lineage diversity of the cell population^{3,4}. As in organismal populations in which clonal expansions to fixation are expected to be common in a small asexual population⁵, a low lineage diversity of the cell population may increase the susceptibility of associated tissues to deleterious somatic mutations^{6–8}. A salient example is the human vascular system that is contributed to by ~20,000–200,000 hematopoietic stem cells or multipotent progenitors. This population of hematopoietic stem cells or multipotent progenitors is essential for preventing clonal expansions in blood at a young age and explains the often-abrupt reduction of blood lineage diversity after the age of

70 years⁹. Such age-dependent clonal expansions are also observed in a variety of solid organs or tissues^{10–13}. Hence, a good method of quantifying the progenitor number of cell populations is valuable for understanding tissue homeostasis^{4,14}. It would also be useful for studying embryonic development, which involves complex cell population expansions and differentiations.

The state-of-the-art strategy for tracking the progenitors of a cell population relies on progenitor-specific expression markers that can be engineered to label the descendant cell population^{15–17}. Despite successful cases^{18–20}, the strategy is limited by the dependence on an ideal marker that is specifically expressed in all progenitor lineages of the focal cell population^{15,21}. Additionally, this method is confirmatory rather than exploratory as it requires prior knowledge of the progenitors and prospective as the revealed progenitors may not all

¹MOE Key Laboratory of Gene Function and Regulation, State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-Sen University, Guangzhou, China. ²These authors contributed equally: Shanjun Deng, Han Gong. ✉e-mail: hexiongli@mail.sysu.edu.cn

effectively contribute to the descendant cell population²². In recent years, single-cell RNA sequencing (scRNA-seq) has been applied to track the expression dynamics during development^{16,23}. Although pseudo-time developmental trajectories based on expression similarity are informative in many aspects, they cannot accurately identify the progenitors of a cell population due to the complexity of interpreting the high-dimensional expression data²⁴.

Recent technical advances in mapping the cell phylogenetic trees of complex organisms, including fruit fly²⁵, zebrafish²⁶, mouse²⁷ and humans⁹, have enabled us to seek for a phylogeny-based solution to the above challenge. This is because, in principle, the ontogeny of a multicellular organism can be fully described by a phylogenetic tree in which the root represents the zygote, the bifurcations represents mitotic cell divisions, and the terminal nodes (leaves) represent terminal cells present in the organism at the time of sampling²⁸.

In this Article, in the first part, we show that a classical theory in population genetics, namely the coalescent theory²⁹, can be applied to a cell phylogenetic tree to quantify the progenitor number of any cell population defined in an organism. Here coalescence refers to two cell lineages that are merged into one ancestral lineage in a phylogenetic tree. Notably, we and others have previously applied coalescent theory to studying cell populations for a variety of purposes^{25,30–35}. Targeting coalescent analysis (TarCA), the method we have proposed here, requires no progenitor-specific information but only a cell phylogenetic tree reconstructed from a small fraction of the cells of a complex organism. Both computer simulations and real data support the validity and robustness of the method in quantifying a progenitor population. In the second part of this article, we show that TarCA enables a systematic search for lineage-specific gene expression upregulations during embryogenesis, revealing incipient cell fate commitments in mouse embryos. This study has shown how classical evolutionary theory, in conjunction with new experimental techniques may resolve a long-standing challenge in developmental cell biology, adding to the emerging field of cell phylogenetics^{36,37}.

Results

Theory for TarCA

In conventional cell biology studies, the strategy for characterizing progenitors of a cell population is prospective¹⁵. As in Fig. 1a, activation of green fluorescent protein (GFP) in a specific progenitor permanently labels the descendant cells, which reveals the progenitor-descendant lineage relationships. In Fig. 1b, we propose a new strategy for quantifying the progenitors of a cell population based on retrospective statistical inferences, which requires no prior information about the progenitors. Suppose that during the development of an organism there are n progenitors each producing m descendant cells that together form a cell type. According to the theory of probability, the probability that two random cells of the cell type share a progenitor is $1/n$, the inverse of the number of progenitors. When cell phylogeny of the organism is available, this probability (Pr) can be calculated as

$$\text{Pr} = \sum C(m_i, 2) / C(\sum m_i, 2), \quad (1)$$

where m_i refers to the number of terminal nodes of the i th monophyletic clade of the focal cell type, $C(m_i, 2)$ the number of possible cell pairs for the m_i cells in the i th monophyletic clade, and $C(\sum m_i, 2)$ the total number of cell pairs within all cells of the focal cell type (Methods). Here a monophyletic clade of a cell type refers to a phylogenetic clade with two properties: first, all terminal nodes of the clade are cells of the focal cell type; second, its parental clade contains cells of other cell types (Fig. 1b). Notably, the inference by equation (1) works well even if m varies substantially, evidenced by the simulations that consider different density distributions of m (Fig. 1c and Supplementary Note 1). Accordingly, the inverse of Pr measures the effective number of progenitors (N_p) that are active in founding the descendant cell population.

Because it is less likely to sample the descendants of the progenitors with a disproportionately smaller m , those progenitors tend not to be represented by N_p (Supplementary Note 1); in addition, for progenitors that are quiescent with no descendant cells, they will be simply ignored by N_p , which is fair. Since the estimation of N_p is based on coalescences in targeting monophyletic clades, we termed the method targeting coalescent analysis or TarCA. Notably, here the coalescence of two cells in a monophyletic clade ignores the fine coalescent position in the clade.

If the complete cell phylogeny of an organism is available, the progenitor number of a cell type would be readily obtained by counting the monophyletic clades of the cell type. As such, no statistical inferences are needed. However, it is currently unrealistic to map the complete cell phylogeny of a complex organism with millions to billions of cells. The cell phylogenies so far available in fruit fly, zebrafish, mouse and humans each comprise only a tiny fraction of the cells of the corresponding organisms. For instance, the two cell phylogenies we recently obtained in fruit fly each contain a few thousand sampled cells, which are approximately 0.1/100 to 1/100 of the total number of cells the organism has²⁵. Hence, the input data of TarCA are often a sampled cell phylogeny in which the terminal nodes represent cells sampled from a whole organism comprising diverse cell types. We anticipate that TarCA will make reasonably good estimations even when the input cell phylogeny comprises only a small number of sampled cells, because the number of cell pairs considered in equation (1) is in the order of the square of the number of sampled cells (Fig. 1d).

Testing TarCA under simple simulated scenarios

We conducted simulations to mimic a developmental process, beginning with a zygote and progressing through cell divisions resulting in 1,000 progenitor cells. These progenitor cells collectively established ten distinct cell types (O0–O9), with each progenitor producing approximately 1,000 descendant cells on average (Fig. 2a and Methods). As such, the simulated *in silico* organism contains approximately one million terminal cells of ten cell types, each founded by a certain number (5–240) of progenitors (Supplementary Fig. 1). Then, the terminal cells were sampled at a given sampling rate and the phylogeny of the sampled cells was analyzed by TarCA. Specifically, for cells of a given cell type in the sampled cell phylogeny we used equation (1) to obtain Pr, the inverse of which is the N_p of the cell type.

As shown in Fig. 2b, the obtained N_p of each cell type can well approximate the actual progenitor number at a sampling rate of 0.1% (that is, 1,000 cells) (Pearson's $r = 0.996$, $n = 10$, $P = 9.1 \times 10^{-10}$, two-sided t -test). The estimation accuracy, defined by 1 minus the ratio of the absolute difference between N_p and actual progenitor number to the actual progenitor number (that is, $1 - (\text{abs}(N_p - \text{actual})/\text{actual})$; Methods), ranges from 63% to 90% across the ten cell types. The similarity of N_p to actual progenitor number increased quickly with the increase of sampling rate (Fig. 2c and Supplementary Fig. 2). In addition, TarCA outperformed the sheer number of monophyletic clades, which has been considered as a measure of progenitor number²⁷, in approximating the actual progenitor number particularly when the sampling rate is small (Extended Data Fig. 1). For example, under the sampling rate of 0.1% the average estimation accuracy by the sheer number of monophyletic clades is ~60%, a number much smaller than that of TarCA (82%; Fig. 2d). Second, in the above analyses the number of sampled cells of each cell type is proportional to the total number of cells of the cell types. In addition to the proportional sampling strategy, we may consider a fixed sampling strategy by which the same number of cells are sampled for each cell type (Supplementary Fig. 3). We thus sampled 100 cells for each of the ten cell types (hence 1,000 cells in total) and assessed the performance of TarCA. Again, TarCA outperformed the sheer number of monophyletic clades in approximating the actual progenitor number (Fig. 2d). This analysis suggests a flexibility in cell sampling, such as enriching rare cell types through flow cytometric cell sorting, to achieve a deep characterization. Third, the above analyses

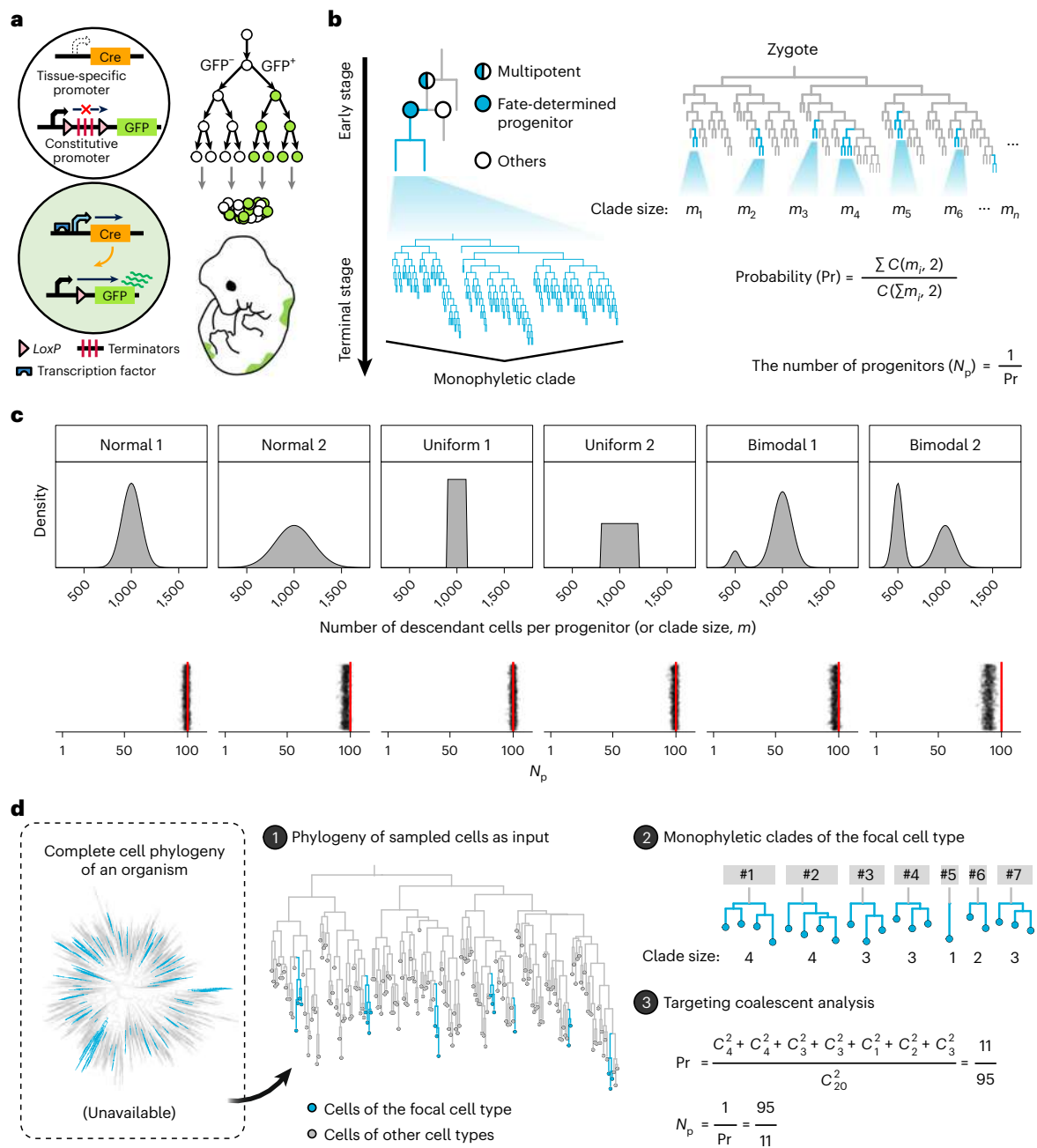


Fig. 1 | Theory for TarCA. **a**, A typical prospective strategy for tracking progenitor-descendant relationships using the Cre-loxP system. A progenitor-specific transcription factor drives the expression of Cre recombinase, which activates the expression of a marker gene (for example, GFP), resulting in permanent fluorescence labeling in the progenitor population as well as the descendant cells (in green). **b**, Schematic of the coalescent method. During development, a fate-committed progenitor gives rise to a specific type of descendant cells that form a monophyletic clade (in blue). Within the cell phylogeny of an organism comprising different cell types, the probability (Pr) that two random cells of the focal cell type are found in the same monophyletic clade is the inverse of the number of progenitors (or clades) of the focal cell type, by assuming the equal sizes of the monophyletic clades. **c**, The method of using

Pr to estimate the progenitor number is robust to variable sizes of monophyletic clades. Each density plot shows a size distribution of the monophyletic clades (see Methods for the corresponding functions). For 100 clades (that is, 100 progenitors) with the sizes randomly picked from the above size distribution, the calculated N_p , the inverse of Pr, is often close to 100, the actual progenitor number. In each case the results of 1,000 replicates are shown with jitter for visual purpose, with a red vertical line marking 100. **d**, Workflow of TarCA for calculating N_p . The input data of TarCA are often a cell phylogeny in which the terminal nodes are cells sampled from a whole organism comprising diverse cell types. Each time a focal cell type is considered, and the monophyletic clades of the focal cell type are identified. The clade sizes are used to calculate Pr and then N_p of the focal cell type.

considered a development process with 1,000 progenitors that found about one million descendant cells of ten cell types. We considered two additional development processes with 2,000 and 3,000 progenitors, respectively, while keeping the total number of descendant cells unchanged (Methods). We found TarCA less sensitive to the

changes than the sheer number of monophyletic clades although both showed a poorer performance under the scenarios of more progenitors (Fig. 2e). The poorer performance is probably due to the reduced proportion of progenitor lineages sampled at a given sampling rate of the descendant cells.

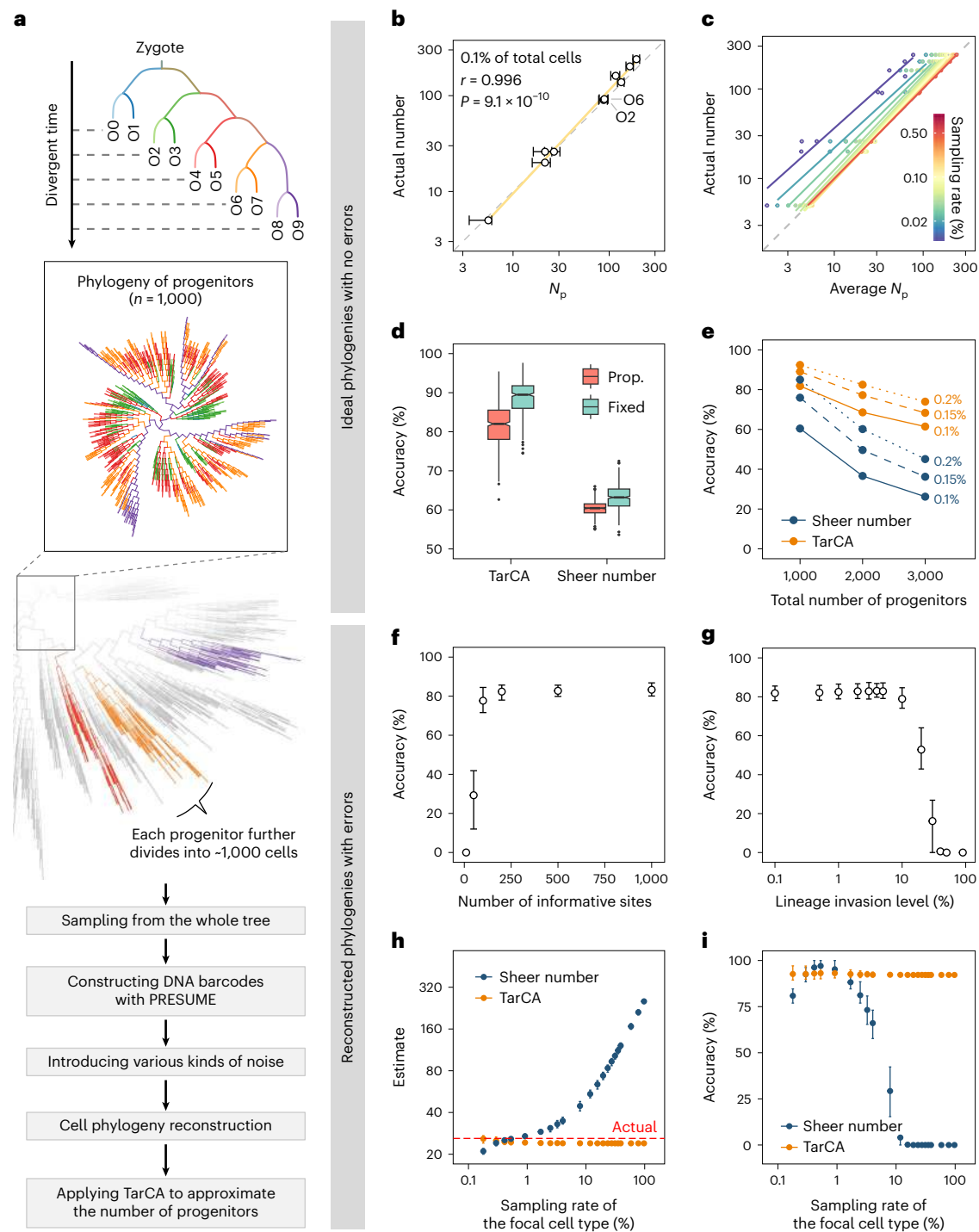


Fig. 2 | Validating the robustness of TarCA with in silico phylogenies.

a, The schematic shows the simulation process, including the definition of the differentiation trajectory, generation of progenitor phylogeny and subsequent expansion of each progenitor cell. Each color represents a specific cell state in a terminal cell population. **b, c**, The impact of sampling rate on TarCA (sampling rate = 0.1% shown in panel **b**). Each dot denotes the average N_p in 1,000 repeats against the actual number of progenitors for a given cell type. Error bars show the lower and upper quartiles, with a trendline colored corresponding to the sampling rate. Pearson's correlation coefficient (r) and P value (P) obtained from two-sided correlation test are attached. **d**, Boxplots compare the performance of TarCA and sheer numbers of monophyletic clades in estimating progenitor size under 0.1% total cells sampling with two strategies: proportional (Prop.) in red and fixed (Fixed) in blue. The center line shows the median accuracy, and the bounds denote the upper and lower quartiles in 1,000 repeats. Whiskers denote the region between median $\pm 1.5 \times$ interquartile range, and outlying

points are shown individually. **e**, The impact of larger progenitor pool size on the performance of TarCA (in orange) and sheer number (in blue), with different proportions of sampled cells, including 0.1% (solid), 0.15% (dashed) and 0.2% (dotted). **f**, The impact of loss of informative sites. Each dot shows the average estimation accuracy for a given number of informative sites, ranging from 10 to 1,000 sites. The error bars show quartiles of 500 replicates. **g**, The performance of TarCA under different levels of lineage invasion. Each dot shows the average estimation accuracy and the error bar shows the lower and the upper quartile of 500 replicates. **h**, Estimates of TarCA and sheer number are presented as the number of cells of the focal cell type (O1) increases with an error rate of 1%. Dots denotes the average estimates, and the error bars show quartiles of 500 replicates. The red dashed line shows the actual number of progenitors. **i**, The estimation accuracy corresponding to **h**. Accuracy = $1 - \frac{\text{abs}(\text{actual} - \text{estimate})}{\text{actual}}$, or zero if $\text{abs}(\text{actual} - \text{estimate})/\text{actual} > 1$.

Testing TarCA under complex simulated scenarios

Because the cell phylogeny of an organism is obtained mostly by phylogenetic reconstruction using lineage-informative mutations, errors in the phylogenetic reconstruction would undermine the performance of TarCA. We thus examined a variety of factors that may introduce errors during phylogenetic reconstruction. These factors include the phylogenetic reconstruction algorithm, the number of lineage-informative mutations, and lineage invasions. Using the algorithm PRESUME³⁸ we simulated a set of DNA barcodes (with 1,000 informative sites and on average -0.01 mutation per site per cell division) by referring to the topology of the cell phylogenies examined in Fig. 2b (Methods). We then reconstructed the cell phylogenies using the assigned DNA barcodes. We considered five widely used phylogenetic reconstruction algorithms and found TarCA insensitive to the algorithms despite fine topological differences among the obtained cell phylogenies (analysis of variance (ANOVA) $P = 0.78$; Supplementary Fig. 4). In addition, since the quality of reconstructed cell phylogenies relies heavily on the number of lineage-informative mutations, we considered only a fraction of the informative sites to assess the potential effects. We found TarCA able to produce robust estimations (Fig. 2f) even in the cell phylogenies with an average ultrafast bootstrap value of -83% (with 100 informative sites considered; Supplementary Fig. 5). We noted that the lack of lineage-informative mutations generates polytomies in cell phylogeny, which would underestimate the number of coalescent events and thus overestimate N_p (Extended Data Fig. 2). Furthermore, experimental contamination or misidentification of cell types can lead to the splitting of a monophyletic clade, a phenomenon we termed as lineage invasion (Extended Data Fig. 3a). To simulate this process we randomly switched the cell type identity for a proportion (ranging from 0.1% to 90%) of the cells in the cell phylogenies examined in Fig. 2b. We found TarCA tolerates well the mislabeling of up to 10% of the cells (Fig. 2g and Extended Data Fig. 3b).

Lastly, we considered scenarios of increasing the sampling rate of a specific cell type. While it is extremely difficult to achieve a high sampling rate for a whole organism with millions to billions of cells, it is plausible to do so for a specific cell type. Again, in the simulations we started with a cell phylogeny as in Fig. 2b (comprising 0.1% of the total cells of the in silico organism) and gradually added the cells of a focal cell type (here, O1) to the phylogeny with a given rate of errors that place the added cells at erroneous positions (Extended Data Fig. 4 and Methods). While TarCA keeps producing stable estimations, we found the sheer number of monophyletic clades of the focal cell type increases constantly with the increase of the sampling rate of the cell type (Fig. 2h and Extended Data Fig. 4). As a result, the estimation accuracy by sheer number first increases to 100% and then decreases to nearly zero as a function of the sampling rate (Fig. 2i and Extended Data Fig. 4). Although the observation appears counterintuitive, it is reasonable because the increased number of added cells would produce more misplaced cells and thus more erroneous phylogenetic branches, directly inflating the sheer number of monophyletic clades of the cell type. Importantly, by considering pairwise coalescent events in monophyletic clades TarCA turned out to be insensitive to such errors. Taken together, these results demonstrate the robustness of TarCA in dealing with the errors of reconstructed cell phylogenies, and caution against the use of sheer number of monophyletic clades to estimate progenitor number.

To further assess the ability of TarCA to handle complex developmental histories, we considered an additional in silico developmental dataset provided by Fang et al.³⁷ In that study, Fang et al. designed 331 quantitative fate maps (QFMs) based on which 3,310 simulated cell phylogenies are generated. We applied TarCA to these cell phylogenies and found TarCA works consistently well in such a large set of cell phylogenies for estimating progenitor number (Supplementary Note 2).

Testing TarCA in nematodes and fruit fly

To test the performance of TarCA in real cell phylogenies, we first examined the nematode *Caenorhabditis elegans* (Methods). The complete cell phylogeny of the nematode, which has been available since the 1980s (ref. 39), contains 671 terminal cells of eight functional types, including blast cells (Bla), epidermal cells (Epi), germ cells (Ger), gland cells (Gla), intestinal cells (Int), muscle cells (Mus), neuron cells (Neu) and structural cells (Str). For each cell type, the actual number of progenitors is readily available by counting the corresponding monophyletic clades in the complete cell phylogeny. Interestingly, the N_p obtained by TarCA well approximated the actual number of progenitors (Pearson's $r = 0.76$, $n = 8$, $P = 0.028$, two-sided t -test), with a nearly perfect match for intestine, germ, epithelial and blast (Extended Data Fig. 5). This is a bit surprising because the number of cells analyzed in each cell type is quite small (ranging from 2 to 226 with a median of 42.5); in addition, the number of descendant cells per progenitor is often not much larger than 1, which is, however, a prerequisite for good performance of N_p in approximating progenitor number (Methods). For some cell types there are substantial deviations between estimated N_p and actual number. For the overestimated cell types the deviations result from a large proportion of single-cell clades (for example, 73.9% for Str and 69.2% for Gla) wherein progenitor does not further produce descendants. For the underestimated cell types like Neu, most descendant cells are produced by a few dominant lineages. This highlights N_p as a measure of the effective number of progenitors that actively participate in founding the descendant population. Similar results were found in another nematode *Pellioditis marina*⁴⁰ whose complete cell phylogeny is also available with six defined cell types Neu, Epi, pharynx (Pha), Mus, Ger and Int, and one unknown (Unknown) cell type (Pearson's $r = 0.85$, $n = 7$, $P = 0.016$, two-sided t -test; Extended Data Fig. 5). Hence, the basic logic of TarCA applies to real cell phylogenies. Nevertheless, the true power of TarCA should be tested in complex organisms whose cell phylogenies can only be reconstructed with errors and from a small proportion of the total cells.

Using a base-editing cell lineage tracing system SMALT, we recently obtained two high-quality cell phylogenetic trees from two third instar larvae of the fruit fly *Drosophila melanogaster*²⁵. Each cell phylogeny comprises about 5,000 cells sampled from over a dozen tissues or organs, including imaginal discs such as wing (Wg), haltere (HI), legs (L1, L2 and L3) and eye-antennal discs (Ey), and a variety of larva-specific tissues/organs such as fat body (Fb), salivary glands (Sg) and hindgut (Hg). It is estimated that approximately 0.1–1% of the total cells of each tissue/organ were sampled^{41,42}. Figure 3a shows the details of one cell phylogeny (Fly-1), in which cells of different tissues/organs are highly mixed with each other, suggesting lineage separations among the tissues/organs occur at rather late embryonic stages. The number of terminal nodes of each tissue/organ-specific monophyletic clades is shown in Supplementary Fig. 6. We applied TarCA to the two cell phylogenies to estimate N_p (with 95% credible intervals) for each of the tissues/organs (Methods). The estimated N_p values are overall consistent between the two individuals (Pearson's $r = 0.69$, $n = 16$, $P = 0.003$, two-sided t -test; Fig. 3b), although midgut (Mg) differs substantially between the two individuals, probably due to the distinct sampling rates as described in the original study²⁵. The imaginal discs of fruit fly have been extensively studied⁴³. Previous microscopic examination of fly embryos reported that the number of progenitor cells is about 20–70 for the leg, wing, eye, haltere and genital discs⁴⁴. Remarkably, the N_p derived from TarCA is 36–105 for each of the discs, very close to the reported numbers (Supplementary Table 1). For the relatively larger brain tissue, the average N_p is 291, a number also close to the reported number (228) of early neuron Bla⁴⁵. It should be noted that a delay in the observation may result in an overestimation of the actual progenitor number, particularly for tissues essential to the early development of *D. melanogaster* larvae. For example, the hindgut is formed at embryonic stages 6–8 but observed at the stage 13 (refs. 46,47). Overall, the

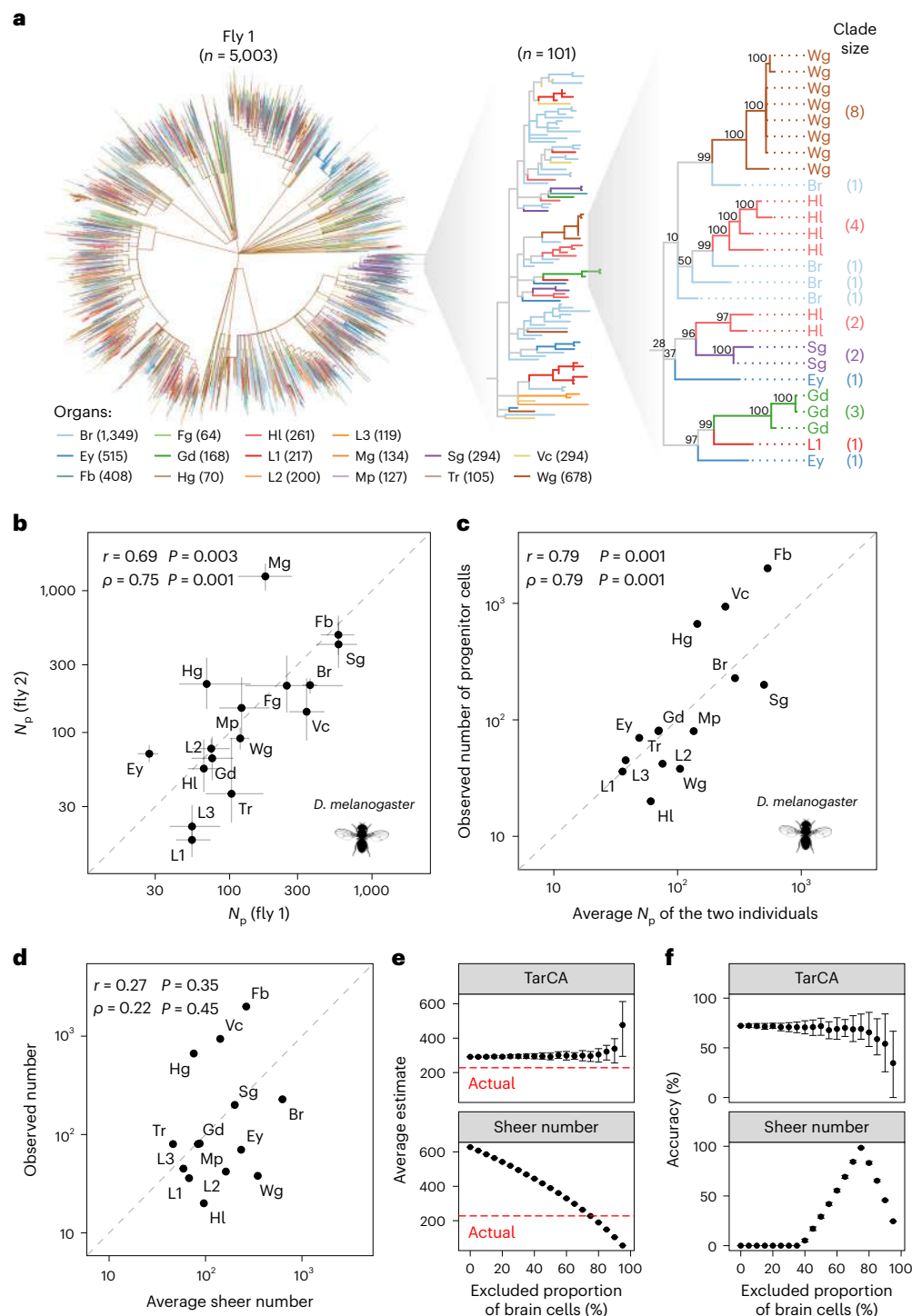


Fig. 3 | Testing TarCA in *D. melanogaster*. **a**, The cell phylogenetic tree of a fly individual comprising 5,003 cells sampled from 16 tissues/organs. The zoomed-in subtrees show details of monophyletic clades of the different tissues/organs in different colors, with the numbers on internal branches showing ultrafast bootstrap (UFboot) values. The clade sizes of a tissue/organ are directly used to estimate N_p of the tissue/organ by referring to equation (1). **b**, The estimated N_p for the tissues/organs is consistent between the two fly individuals. The gray lines show the 95% credible intervals. The attached Pearson's correlation coefficient (r), Spearman's correlation coefficient (ρ) and the corresponding P value (P) were determined by two-sided test of correlation. **c**, The N_p (average of two individuals) matches well the observed actual progenitor number. The attached Pearson's correlation coefficient (r), Spearman's correlation coefficient (ρ) and the corresponding P value (P) were determined by two-sided test of

correlation. **d**, The sheer numbers of monophyletic clades have no significant correlation with the observed actual progenitor numbers. The average of two fly phylogenies is shown for sheer numbers of monophyletic clades. The attached Pearson's correlation coefficient (r), Spearman's correlation coefficient (ρ) and the corresponding P value (P) were determined by two-sided test of correlation. **e**, The estimate of TarCA and sheer number, as a function of the downsampling proportion of the brain cells in the fly phylogenies. The average between the two individuals in 1,000 replicates is shown. Each dot denotes the average estimate, and the error bar shows the lower and the upper quantile of 1,000 replicates. The red dashed line shows the actual progenitor number of brain. **f**, The estimation accuracy corresponding to **e**. Accuracy = $1 - (\text{abs}(\text{actual} - \text{estimate})/\text{actual})$, or zero if $\text{abs}(\text{actual} - \text{estimate})/\text{actual} > 1$.

N_p computed by TarCA is quite consistent with the observed number of progenitors for the 14 tissues/organs with related experimental data (Pearson's $r = 0.79$, $n = 14$, $P = 0.001$, two-sided t -test; Fig. 3c and Supplementary Table 1). This consistency cannot be explained by the number of cells included in the cell phylogenies (ANOVA $P = 0.99$; Supplementary Fig. 7). By contrast, the sheer numbers of monophyletic clades are not correlated with the observed progenitor numbers (Pearson's $r = 0.27$, $n = 14$, $P = 0.35$, two-sided t -test; Fig. 3d). Considering the 0.1–1% sampling rate for each of the fly tissues/organs, TarCA turns out to be a powerful method.

Since brain was heavily sampled and accounted for nearly a fifth of all sampled cells in each of the fly cell phylogenies²⁵, we conducted a downsampling analysis by excluding a proportion (from 5% to 95%) of the brain cells from the phylogenies. We found TarCA produces very stable estimations even if 95% of the cells are excluded while the sheer number of monophyletic clades of brain reduces constantly and quickly with the growing proportion of cells excluded (Fig. 3e). As a result, the estimation accuracy by sheer number fluctuates dramatically as a function of the downsampling proportion (Fig. 3f). The results echo our preceding simulation results shown in Fig. 2h,i, highlighting the limitation of using the sheer number of monophyletic clades to estimate progenitor number in reconstructed cell phylogenies with errors.

Testing TarCA in mouse embryos

Using a Cas9-based lineage tracing system, a previous study reconstructed the cell phylogenies of a few mouse embryos²⁷. The examined embryos are roughly at the stage of gastrulation when fate commitment is ongoing for most terminal cell types. Since the reconstructed cell phylogenies are of moderate quality, we excluded the low-quality phylogenies containing less than 1,000 unique cell barcodes or harboring huge polytomies covering more than half of cells, leaving two phylogenies (embryo 3 and embryo 6) for further analysis (Supplementary Note 3). The available scRNA-seq data of the two embryos enabled the characterization of cell types (Methods). We applied TarCA to the two cell phylogenies and obtained N_p for 15 cell types (Methods). The estimated N_p values are quite consistent between the two embryos (Pearson's $r = 0.73$, $n = 15$, $P = 0.002$, two-sided t -test; Fig. 4a). However, we noted that the estimations should have overall been inflated due to the prevalent polytomies in the cell phylogenies. Hence, the absolute N_p values must be treated with cautions while their relative values should be meaningful. The results have some notable features. For example, among all examined cell types, two extraembryonic structures, extraembryonic ectoderm (ExE) and extraembryonic endoderm, have the smallest N_p (~1,000). In particular, the N_p of ExE is smaller than that of extraembryonic endoderm. This cannot be explained by sampling bias across the cell types (ANOVA $P = 0.50$) but is consistent with the fact that the two cell types are both early fate determined and separated from the rest of an embryo in a sequential order before the stage of epiblast (Epb; Fig. 4a)⁴⁸. In sharp contrast, the N_p of another extraembryonic structure, the extraembryonic mesoderm, is ~10-fold larger (~10,000) and comparable to that of many mesoderm-derived cell types. This is in line with the fact that the extraembryonic mesoderm originates from primitive streak (PrS)⁴⁹, an embryonic region also producing mesoderm and endoderm and expected to have a complex lineage composition (Fig. 4a).

A previous study assembled time-series scRNA-seq data of mouse embryos and mapped the pseudo-time developmental trajectories for all kinds of cell types in early developmental stages⁵⁰, including the 15 cell types examined in Fig. 4a (Methods). It is expected that a cell type with complex pseudo-time developmental trajectories tends to have a large N_p . We quantified the complexity of pseudo-time trajectories using an entropy-like statistic for each of the 15 cell types, which measures the heterogeneity of ancestral cell populations contributing to the focal cell type (Fig. 4b, Supplementary Fig. 8 and Methods). For instance, gut endoderm cells are originated from the anterior PrS,

definitive endoderm and embryonic visceral endoderm at embryonic day (E)7.5, and the latter two continuously contribute to gut tube formation in the subsequent stages, indicating a complex developmental history of gut endoderm. Interestingly, the obtained pseudo-time trajectory complexity of the cell types is correlated with their N_p (Pearson's $r = 0.77$, $n = 15$, $P = 0.001$, two-sided t -test; Fig. 4c), thus providing further support to TarCA.

Since TarCA can apply to any cell populations defined in a cell phylogeny, we examined ten artificially defined cell populations, some of which represent mere anatomic classifications corresponding to key early embryonic events (Extended Data Fig. 6). They include inner cell mass, ExE, Epb, hypoblast, PrS, ectoderm, endoderm, mesoderm, surface ectoderm and neuroectoderm^{48–50}. We estimated N_p for each of the ten defined cell populations and obtained highly similar results in the two embryos (Pearson's $r = 0.95$, $n = 10$, $P = 3.53 \times 10^{-5}$, two-sided t -test; Extended Data Fig. 6). As expected, the N_p increases generally as a function of the time when the cell populations appear in an embryo: the average N_p is 116, 329, 228 and 613, respectively, for inner cell mass, ExE, Epb and hypoblast, the four early cell populations appearing before the formation of PrS; the N_p is up to tens of thousands for the other cell populations.

TarCA reveals incipient fate commitments by identifying LUGs

Due to its the simple design and robust performance in quantifying the lineage diversity of a cell population, TarCA can be applied to various research questions as long as the cell phylogeny is available. Below we will show how TarCA can inform cell fate commitments by tracking the lineage-dependent gene expression upregulations during embryogenesis.

In development, cell fate commitment refers to the transition of expression state from being lineage independent to lineage dependent such that all descendant cells of a fate-committed progenitor have the same characteristic expression state; the transition is supposed to be a stepwise process with some early driver or marker genes activated and maintained in specific lineages at an incipient stage (Fig. 5a). The terminal stage of a fate commitment is often traceable because the whole transcriptome can be readily mapped to known cell types; however, tracking the incipient stage with only a few upregulated driver or marker genes in a small number of cells is difficult. We reasoned that TarCA can be extended to identifying lineage-specific upregulated genes (LUGs) from cells with both phylogeny and scRNA information. Specifically, for a focal gene that is upregulated in a set of cells, the lineage diversity (that is, progenitor number N_p) of these cells in cell phylogeny can be quantified by TarCA; the obtained N_p will then be compared to the expected lineage diversity of the same number of randomly selected cells in the cell phylogeny; the focal gene would be defined as an LUG if the obtained N_p is significantly smaller than the expected (Methods). Because cell fate commitment is characterized by lineage-coupled gene expression changes, identification of LUGs would help uncover the incipient stage of a cell fate commitment when a transcriptome-wide fate commitment signature is not yet established.

The gut endoderm cells of the two mouse embryos examined in Fig. 4 (embryo 3 and embryo 6) are at the stage of approximately E8.0–E8.5 (Supplementary Fig. 9a). It is known that the gut endoderm has a spatial pattern along the anterior–posterior axis at ~E7.5 and will form seven organs, including thyroid, thymus, lung, liver, pancreas, small intestine and colon, at ~E9.0 (refs. 51,52). Accordingly, it is expected that the gut endoderm cells of the two embryos are undergoing incipient fate commitments. We first examined embryo 3, whose gut endoderm cells form five clusters according to their scRNA-seq transcriptomes (Fig. 5b and Methods). There are no clear transcriptome-wide expression signatures toward a specific organ among the clusters (Supplementary Fig. 9b). The only exception is cluster 3, which showed high expression in *Rhox5*, *Trap1a* and *Xlr3a*,

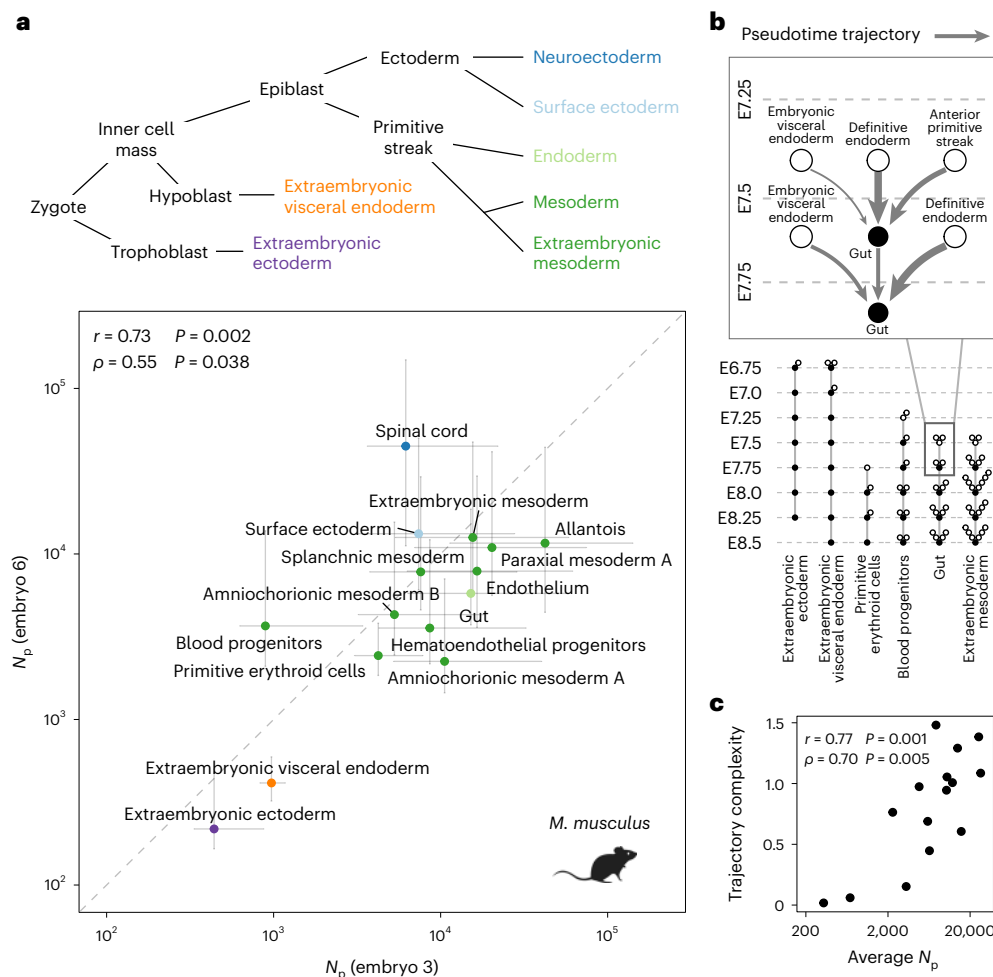


Fig. 4 | Testing TarCA in mouse embryos. a, The N_p is shown for the 15 cell types existing in the mouse embryos, with the gray lines showing the 95% credible intervals. The major early embryonic events are also provided for a better illustration of the relationships among the cell types/populations. **b**, Measuring the complexity of pseudo-time trajectories that contribute to a cell type. During development, a focal cell type (solid circle) is often contributed by other cell populations (hollow circle) according to the pseudo-time trajectories derived

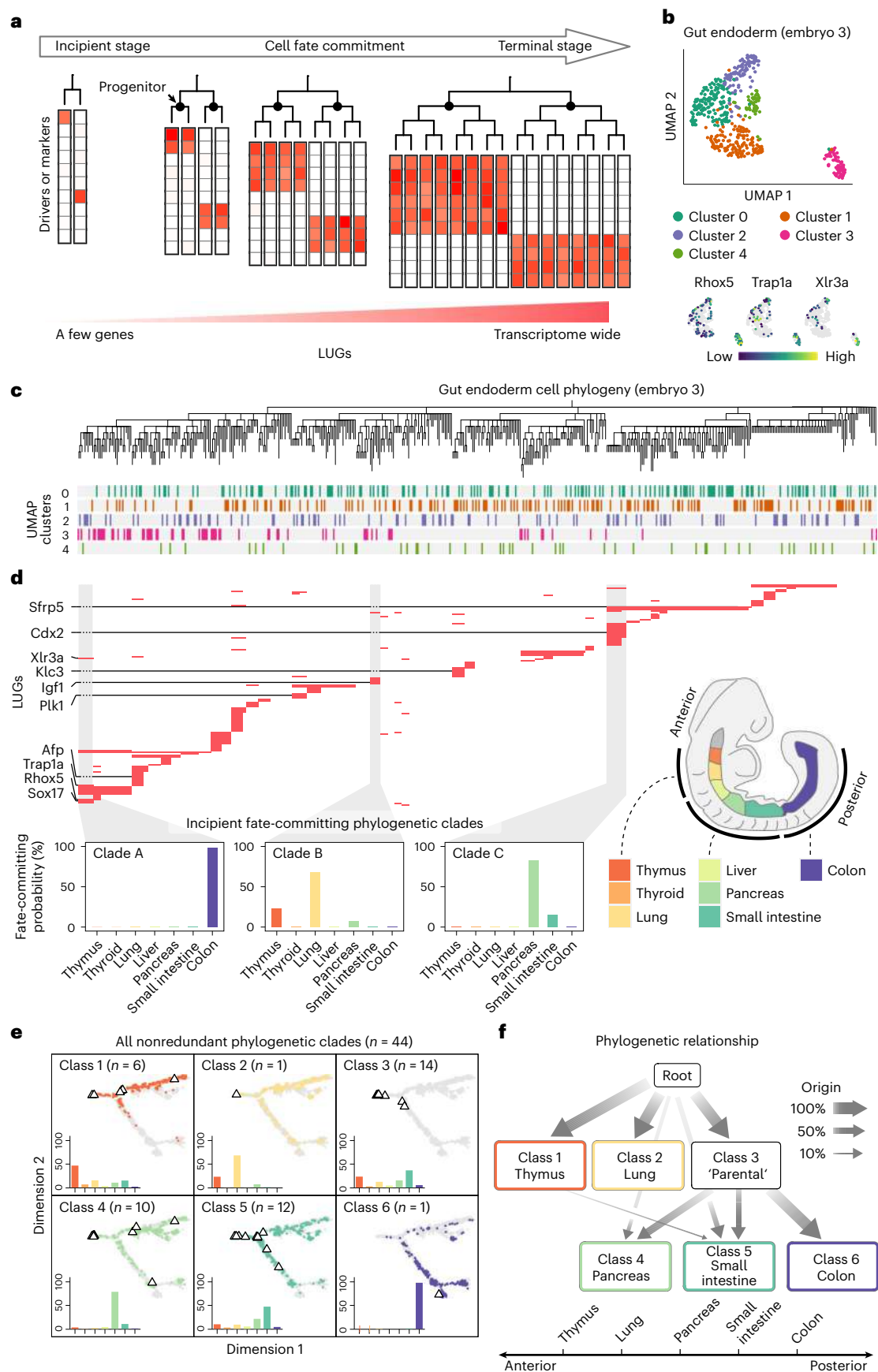
from cellular expression state (from E6.75 to E8.5). The more contributing cell populations (hollow circles) during the process suggests a higher complexity of the pseudo-time trajectories. See also Supplementary Fig. 8 for the other cell types examined in **a**. **c**, The pseudo-time trajectory complexity is strongly correlated with the N_p . Each dot represents a cell type. The attached Pearson's correlation coefficient (r), Spearman's correlation coefficient (ρ) and the corresponding P value (P) were determined by two-sided test of correlation.

a well-known gene set for maintaining visceral endoderm origin lineage of the gut tube in hindgut development, suggesting the hindgut signature originated from visceral endoderm rather than definitive embryonic endoderm (Fig. 5b); meanwhile, we noted that the convergence of some visceral endoderm cells toward definitive endoderm cell state may have shaped the phylogenetic distribution of cluster 3 cells in Fig. 5c^{48,53–55} (Fig. 5c). Additionally, we conducted a randomization analysis and found that the four out of five clusters, except for cluster 3, are dispersedly distributed on the cell phylogenetic tree ($P > 0.05$; Fig. 5c and Supplementary Fig. 10). These results together indicate that the gut endoderm cells in embryo 3 are generally far from the terminal stage of fate commitments.

To obtain LUGs in the gut endoderm cells to track incipient fate commitments, we performed a systematic gene-by-gene search. Specifically, each time a single gene is tested, and a small proportion (here 10%) of the cells that show the highest expression level for the focal gene are treated as a population whose N_p in cell phylogeny is then quantified by TarCA. As explained before, the gene will be defined as an LUG if the obtained N_p is significantly ($P < 0.01$) smaller than the expected (Methods). After testing all individual genes of the transcriptome, we obtained 155 LUGs in embryo 3 (Fig. 5d, Extended Data Fig. 7, Supplementary Fig. 11 and Methods). There are a few notable

findings. First, three classic markers of visceral endoderm cells, including *Trap1a*, *Rhox5* and *Xlr3a*, are identified, consistent with the partial extraembryonic origin of gut tube⁵⁶. Second, a few well-known regulators, such as *Igf1*, *Sfrp5* and *Cdx2*, that contribute to the organogenesis of lung⁵⁷, pancreas⁵⁸ and intestine⁵⁹, respectively, are also identified. Third, the LUGs are not dispersedly distributed in the cell phylogeny; instead, they tend to co-exist in some phylogenetic clades.

To annotate the phylogenetic clades with LUGs we referred to the single-cell transcriptome collected from mouse gut endoderm at E8.75, when the majority of cells exhibit clear organ-specific transcriptomic signatures⁵³. By evaluating the expression state of the LUGs at E8.75, we trained a support vector machine (SVM) classifier to predict for each phylogenetic clade the fate-committing probabilities toward seven organs (Fig. 5d, Supplementary Fig. 12 and Methods). As such, each clade has seven fate-committing probabilities that are summed up to be 1. Some clades show a dominant propensity toward an organ, with the fate-committing probability larger than 50%. For instance, as shown in Fig. 5d, clade A is composed of 10 cells with 13 LUGs that represent a strong colon-specific signature, with the corresponding fate-committing probability $>90\%$; clade B is composed of seven cells with six LUGs that represent a strong lung-specific signature; clade C is composed of 13 cells with 14 LUGs that represent



a strong pancreas-specific signature. Hence, despite the absence of transcriptome-wide organ-specific signatures in the gut endoderm cells of embryo 3, a few LUGs can inform the fate commitments of an

incipient stage. This being said, we noted that LUGs may not necessarily be the drivers of differentiations although differentiation drivers are more likely to be LUGs.

Fig. 5 | TarCA informs incipient cell fate commitments by identifying LUGs.

a, Cell fate commitment is supposed to be a continuous process, characterized by a few LUGs at the incipient stage and scaling up to a transcriptome-wide scale at the terminal stage. **b**, UMAP visualization of the gut endoderm cells in embryo 3. Cells are colored according to Leiden clustering (top), with the same projections additionally colored according to the expression levels of three extraembryonic marker genes, *Trap1a*, *Rhox5* and *Xlr3a* (bottom). **c**, The phylogenetic tree of the gut endoderm cells of embryo 3, with cell annotations referring to the UMAP projection in **b**. **d**, Left: the distribution of LUGs in the gut endoderm cell phylogeny of embryo 3. Each red tile represents an LUG upregulated in corresponding phylogenetic clade, with 155 LUGs identified and 10 representative LUGs shown. Right: bar plots showing the fate-committing probabilities of three exemplar clades. The anatomic structure of the seven organs derived from gut tube is shown. **e**, The 44 nonredundant phylogenetic

clades are categorized into six classes, each with a distinct profile of fate-committing probability. Each panel represents a clade class, with dots showing the gut endoderm cells of E8.75 stage and triangles showing the phylogenetic clades of embryo 3 based on the expression profiles of the 155 LUGs in two-dimensional projection. Some cells of E8.75 stage are colored because their committed organ matches the major fate-committing organ of the designated clade class of embryo 3. Inset bar plots show the average fate-committing probability of clades within a given class. The *n* in parentheses shows the number of clades analyzed. **f**, The overall phylogenetic relationship of the six classes of clades. Arrows show the parent–child relationship between clades of two classes in the phylogeny, with arrow width representing the frequency of the designated origin (that is, parent) for a given class of clades. The phylogenetic distribution of the six classes of clades echoes the anatomic distribution of their fate-committing organs.

We next checked all clades in the cell phylogeny of embryo 3. After merging the clades with the same LUGs, we obtained 44 nonredundant clades. As shown in Fig. 5e, the fate-committing probability profile of the 44 clades forms six classes (Methods): class 2, class 4 and class 6 each have a dominant propensity toward lung, pancreas and colon, respectively, with the corresponding fate-committing probability >50%. Class 1 and class 5 have a clear tendency toward thymus and small intestine, respectively, with the fate-committing probability ~40%. There is no clear organ-specific tendency observed for class 3. There are informative patterns regarding the parent–child relationship of the clades in the cell phylogeny (Fig. 5f): nearly all class 1, class 2 and class 3 clades are the direct child clades of the root; meanwhile, class 3 clades serve as the major parental clades of class 4, class 5 and class 6 clades. Interestingly, the phylogenetic distribution of the six clade classes echoes the anatomic distribution of their fate-committing organs as shown in Fig. 5d. Specifically, class 1 and class 2 clades have a tendency toward anterior organs (thymus and lung) while class 4, class 5 and class 6 clades have a tendency toward posterior organs (pancreas, small intestine and colon). This observation is further supported by directed differentiation experiments conducted *in vitro*⁶⁰. Hence, the fate-committing trajectories of the organs are shaped by their anatomical distribution, which is captured by our analysis.

Using the same procedure, we also examined the cell phylogeny of embryo 6. This embryo is at an earlier developmental stage than embryo 3 according to our analysis (Supplementary Fig. 9a) as well as the original study²⁷. As expected, we found a generally much weaker fate-committing tendency toward a specific organ for the phylogenetic clades of embryo 6 (Extended Data Fig. 8 and Supplementary Figs. 13–15).

Applying TarCA to analyze static cell barcoding data

In addition to evolving cell barcodes that are required for cell phylogeny reconstruction, static cell barcodes have been widely used for cell lineage tracing by defining cell clones¹⁶. In principle, TarCA can also be applied to analyze such cell clones defined by static cell barcodes, because clonal cells with the same static barcode resemble a monophyletic clade in cell phylogeny. As such, we can apply TarCA to search for LUGs from single-cell data with static cell barcode information.

To demonstrate this issue we analyzed a dataset that records with static cell barcodes the *in vitro* differentiation process of mouse hematopoietic stem cells (Methods)⁶¹. After the barcoding treatment at day 0, all descendant cells derived from a progenitor would harbor the same barcode as the progenitor. The barcoded cells were then sampled at days 2, 4 and 6, respectively, with both barcode sequences and scRNA profiles characterized. We applied TarCA to search for the LUGs during hematopoiesis using the same procedure as in Fig. 5 except that cell clones are treated as monophyletic clades (Supplementary Note 4). In total, 1,825 LUGs were identified (Extended Data Fig. 9), covering the core differentiation driver genes *Cpa3*, *Lgals1*, *Cyp11a1* and *Srgn* defined by Weinreb et al.⁶¹.

Notably, an alternative computational tool, CoSpar, has been used to analyze this dataset to infer fate-associated genes during hematopoiesis, with 377 genes identified⁶². There are 260 overlaps between the LUGs defined by TarCA and the fate-associated genes defined by CoSpar (8.2-fold enrichment; $P < 2.2 \times 10^{-16}$, chi-squared test). As shown in Supplementary Fig. 16, the 260 overlaps are enriched in Gene Ontology (GO) terms associated with myeloid cell differentiation toward megakaryocyte (Mk) and granulocyte (Gr). Interestingly, the TarCA-specific LUGs are also enriched in GO terms associated with the differentiation of erythrocyte (Er) and mast cells (Ma), while the CoSpar-specific fate-associated genes show no such GO properties. These results demonstrated the adaptability of TarCA in dealing with static barcoding information.

Discussion

In this study, we developed TarCA, a phylogeny-based statistical method for estimating the progenitor number of a cell population. The method circumvents the major limitation of conventional experimental methods that rely on precharacterized markers to track progenitors¹⁵. A particular strength of TarCA is its versatility for analyzing organs, tissues, cell types or any cell populations defined by researchers without the need for prior information on the progenitors. Another strength is it works robustly in a cell phylogeny comprising a tiny fraction of the total cells in a complex organism. In addition, TarCA can produce stable estimations in a reconstructed cell phylogeny that contains errors. In contrast, the tentative approach of counting the sheer number of monophyletic clades is highly sensitive to the errors in phylogeny reconstruction. With the growing availability of reconstructed cell phylogenies, we anticipate TarCA would enable the quantification of lineage diversity for all tissues of a complex organism including humans. This would be fundamental to a quantitative model for understanding the between-individual developmental plasticity and the age-dependent degeneration of the tissues^{63,64}.

Lineage diversity has been a central concept in evolutionary studies on organismal populations, including speciation, migration, adaptation, extinction and so on. Hence, it is not unexpected that TarCA can be adapted to address various kinds of questions about cell populations. In this study, we show that TarCA can be used to track incipient cell fate commitments by identifying a few LUGs in a few cells. This circumvents the requirement for transcriptome-wide signatures to track the intermediate states of differentiation with dense sampling. This said, we note that the current analyses are proof-of-concept and can be much improved in the future. Specifically, the relatively low resolution of the cell phylogeny of the mouse embryos prevent a more sensitive discovery of LUGs using a smaller cell population. In the search-for-LUG analysis, we considered the 10% of cells with the highest expression of a focal gene as a testing population, which corresponds to about 50 cells (Methods). In principle, considering a smaller proportion, say 5% or 1%, would reveal even earlier LUGs. However, in our case, this would substantially reduce the power of TarCA because there would be too

few coalescent events detected with the sparsity of effective coalescent events in the cell phylogeny. Provided a high-resolution cell phylogeny with as many as possible cells, we may discover clades covering a full spectrum of fate-committing stages, providing a high-resolution roadmap of cell fate commitments.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41592-024-02189-7>.

References

- Frankham, R. Relationship of genetic variation to population size in wildlife. *Conserv. Biol.* **10**, 1500–1508 (1996).
- Charlesworth, B. Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205 (2009).
- Goodell, M. A., Nguyen, H. & Shroyer, N. Somatic stem cell heterogeneity: diversity in the blood, skin and intestinal stem cell compartments. *Nat. Rev. Mol. Cell Biol.* **16**, 299–309 (2015).
- Bhartiya, D. et al. Evolving definition of adult stem/progenitor cells. *Stem Cell Rev. Rep.* **15**, 456–458 (2019).
- Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, 1983).
- Beerman, I. & Rossi, D. J. Epigenetic control of stem cell potential during homeostasis, aging, and disease. *Cell Stem Cell* **16**, 613–625 (2015).
- Forsberg, L. A., Gisselsson, D. & Dumanski, J. P. Mosaicism in health and disease—clones picking up speed. *Nat. Rev. Genet.* **18**, 128–142 (2017).
- Risques, R. A. & Kennedy, S. R. Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS Genet.* **14**, e1007108 (2018).
- Mitchell, E. et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022).
- Li, R. Y. et al. A body map of somatic mutagenesis in morphologically normal human tissues. *Nature* **597**, 398–403 (2021).
- Coorens, T. H. H. et al. Extensive phylogenies of human development inferred from somatic mutations. *Nature* **597**, 387–392 (2021).
- Moore, L. et al. The mutational landscape of human somatic and germline cells. *Nature* **597**, 381–386 (2021).
- Park, S. et al. Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature* **597**, 393–397 (2021).
- Clevers, H. & Watt, F. M. Defining adult stem cells by function, not by phenotype. *Annu. Rev. Biochem.* **87**, 1015–1027 (2018).
- Das, R. N. & Yaniv, K. Discovering new progenitor cell populations through lineage tracing and in vivo imaging. *Cold Spring Harb. Perspect. Biol.* **12**, a035618 (2020).
- VanHorn, S. & Morris, S. A. Next-generation lineage tracing and fate mapping to interrogate development. *Dev. Cell* **56**, 7–21 (2021).
- Kretzschmar, K. & Watt, F. M. Lineage tracing. *Cell* **148**, 33–45 (2012).
- Shu, H. S. et al. Tracing the skeletal progenitor transition during postnatal bone formation. *Cell Stem Cell* **28**, 2122–2136 (2021).
- Cattaneo, P. et al. Parallel lineage-tracing studies establish fibroblasts as the prevailing in vivo adipocyte progenitor. *Cell Rep.* **30**, 571–582 (2020).
- Wei, Y. et al. Liver homeostasis is maintained by midlobular zone 2 hepatocytes. *Science* **371**, eabb1625 (2021).
- Liu, K., Jin, H. & Zhou, B. Genetic lineage tracing with multiple DNA recombinases: a user's guide for conducting more precise cell fate mapping studies. *J. Biol. Chem.* **295**, 6413–6424 (2020).
- He, L. et al. Enhancing the precision of genetic lineage tracing using dual recombinases. *Nat. Med.* **23**, 1488–1498 (2017).
- Farrell, J. A. et al. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131 (2018).
- Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* **21**, 410–427 (2020).
- Liu, K. et al. Mapping single-cell-resolution cell phylogeny reveals cell population dynamics during organ development. *Nat. Methods* **18**, 1506–1514 (2021).
- McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
- Chan, M. M. et al. Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019).
- Frumkin, D., Wasserstrom, A., Kaplan, S., Feige, U. & Shapiro, E. Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput. Biol.* **1**, 382–394 (2005).
- Fu, Y. X. & Li, W. H. Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theor. Popul. Biol.* **56**, 1–10 (1999).
- Yao, Z., Liu, K., Deng, S. & He, X. An instantaneous coalescent method insensitive to population structure. *J. Genet. Genomics* **48**, 219–224 (2021).
- Hu, Z., Fu, Y. X., Greenberg, A. J., Wu, C. I. & Zhai, W. Age-dependent transition from cell-level to population-level control in murine intestinal homeostasis revealed by coalescence analysis. *PLoS Genet.* **9**, e1003326 (2013).
- Werner, B. et al. Measuring single cell divisions in human tissues from multi-region sequencing data. *Nat. Commun.* **11**, 1035 (2020).
- Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat. Genet.* **48**, 238–244 (2016).
- Smadbeck, P. & Stumpf, M. P. H. Coalescent models for developmental biology and the spatio-temporal dynamics of growing tissues. *J. R. Soc. Interface* **13**, 20160112 (2016).
- Slatkin, M. Simulating genealogies of selected alleles in a population of variable size. *Genet. Res.* **78**, 49–57 (2001).
- Stadler, T., Pybus, O. G. & Stumpf, M. P. H. Phylodynamics for cell biologists. *Science* **371**, eaah6266 (2021).
- Fang, W. et al. Quantitative fate mapping: a general framework for analyzing progenitor state dynamics via retrospective lineage barcoding. *Cell* **185**, 4604–4620 (2022).
- Konno, N. et al. Deep distributed computing to reconstruct extremely large lineage trees. *Nat. Biotechnol.* **40**, 566–575 (2022).
- Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100**, 64–119 (1983).
- Houthoofd, W. et al. Embryonic cell lineage of the marine nematode *Pellioditis marina*. *Dev. Biol.* **258**, 57–69 (2003).
- Lehner, C. F., Jacobs, H. W., Sauer, K. & Meyer, C. A. Regulation of the embryonic cell proliferation by *Drosophila* cyclin D and cyclin E complexes. *Novartis Found. Symposia* **237**, 43–54 (2001).
- Aldaz, S. & Escudero, L. M. Imaginal discs. *Curr. Biol.* **20**, 429–431 (2010).
- Beira, J. V. & Paro, R. The legacy of *Drosophila* imaginal discs. *Chromosoma* **125**, 573–592 (2016).
- Madhavan, M. M. & Schneiderman, H. A. Histological analysis of the dynamics of growth of imaginal discs and histoblast nests during the larval development of *Drosophila melanogaster*. *Wilhelm. Roux's Arch. Dev. Biol.* **183**, 269–305 (1977).

45. Crews, S. T. *Drosophila* embryonic CNS development: neurogenesis, gliogenesis, cell fate, and differentiation. *Genetics* **213**, 1111–1144 (2019).
 46. Lengyel, J. A. & Iwaki, D. D. It takes guts: the *Drosophila* hindgut as a model system for organogenesis. *Dev. Biol.* **243**, 1–19 (2002).
 47. Nakamura, M. et al. Reduced cell number in the hindgut epithelium disrupts hindgut left–right asymmetry in a mutant of pebble, encoding a RhoGEF, in *Drosophila* embryos. *Mech. Dev.* **130**, 169–180 (2013).
 48. Rivera-Perez, J. A. & Hadjantonakis, A. K. The dynamics of morphogenesis in the early mouse embryo. *Cold Spring Harb. Perspect. Biol.* **7**, a015867 (2015).
 49. Carlson, B. M. Gastrulation and germ layer formation. in *Reference Module in Biomedical Sciences* (2015).
 50. Qiu, C. et al. Systematic reconstruction of cellular trajectories across mouse embryogenesis. *Nat. Genet.* **54**, 328–341 (2022).
 51. Nowotschin, S. & Hadjantonakis, A. K. Guts and gastrulation: emergence and convergence of endoderm in the mouse embryo. *Curr. Top. Dev. Biol.* **136**, 429–454 (2020).
 52. Smith, R. J. et al. Single-cell chromatin profiling of the primitive gut tube reveals regulatory dynamics underlying lineage fate decisions. *Nat. Commun.* **13**, 2965 (2022).
 53. Nowotschin, S. et al. The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature* **569**, 361–367 (2019).
 54. Pijuan-Sala, B. et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490–495 (2019).
 55. Li, L. C. et al. Single-cell patterning and axis characterization in the murine and human definitive endoderm. *Cell Res.* **31**, 326–344 (2021).
 56. Kwon, G. S., Viotti, M. & Hadjantonakis, A. K. The endoderm of the mouse embryo arises by dynamic widespread intercalation of embryonic and extraembryonic lineages. *Dev. Cell* **15**, 509–520 (2008).
 57. Wang, Z. et al. Insulin-like growth factor-1 signaling in lung development and inflammatory lung diseases. *BioMed Res. Int.* **2018**, 6057589 (2018).
 58. Cebola, I. et al. TEAD and YAP regulate the enhancer network of human embryonic pancreatic progenitors. *Nat. Cell Biol.* **17**, 615–626 (2015).
 59. Gao, N., White, P. & Kaestner, K. H. Establishment of intestinal identity and epithelial-mesenchymal signaling by Cdx2. *Dev. Cell* **16**, 588–599 (2009).
 60. Ikononou, L. & Kotton, D. N. Derivation of endodermal progenitors from pluripotent stem cells. *J. Cell. Physiol.* **230**, 246–258 (2015).
 61. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, eaaw3381 (2020).
 62. Wang, S. W., Herriges, M. J., Hurley, K., Kotton, D. N. & Klein, A. M. CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information. *Nat. Biotechnol.* **40**, 1066–1074 (2022).
 63. Chen, C., Liao, Y. & Peng, G. Connecting past and present: single-cell lineage tracing. *Protein Cell* **13**, 790–807 (2022).
 64. Lopez-Otin, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
- Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
- Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.
- © The Author(s), under exclusive licence to Springer Nature America, Inc. 2024

Methods

Cell phylogeny as the input data of TarCA

All cells of a multicellular organism often develop from a zygote and their developmental genealogical history can be described by a phylogenetic tree with the root representing the zygote, the bifurcations representing mitotic cell divisions, and the terminal nodes (leaves) representing terminal cells present in the (developed) organism. Hence, a cell phylogeny often refers to the developmental genealogy of a set of terminal cells sampled from a multicellular organism (developmental intermediate/progenitor cells are not available in the organism at the time of sampling). The sampled cells are often from a whole individual comprising diverse organs, tissues and cell types. The cells can be annotated according to their anatomic distribution (for example, brain or blood) or their expression information (for example, scRNA-seq profile). The genealogical relationship of the cells is often traced by a DNA barcoding system engineered in the organism that introduces mutations specifically on a barcode DNA sequence during embryogenesis, such as the SMALT system we recently invented²⁵. As such, the phylogeny of the sampled cells can be obtained by referring to the barcoding sequences of the cells, with a routine phylogeny reconstruction process same as building a species phylogeny using orthologous gene sequences³⁶.

Derivation of N_p

Supposing during the development of an organism there are n progenitors each producing m cells of a given cell type. As such, there are n monophyletic clades of the cell type each with m terminal nodes in the complete cell phylogeny of the organism. Here a monophyletic clade refers to a phylogenetic clade with two properties: first, all terminal nodes of the clade are cells of the focal cell type; second, its parental clade contains cells of other cell types. The probability (Pr) that two random cells of the cell type are present in the same monophyletic clade can be calculated as

$$\text{Pr} = \frac{n \times C_m^2}{C_{n \times m}^2}, \quad (2)$$

where n is the number of monophyletic clades (or progenitors), C_m^2 is the number of possible cell pairs for the m cells in a monophyletic clade, and $C_{n \times m}^2$ is the total number of cell pairs for the $n \times m$ cells of the cell type. Then, we have

$$\text{Pr} = \frac{n \times m \times (m-1)}{(n \times m) \times (n \times m - 1)}, \quad (3)$$

$$\text{Pr} = \frac{m-1}{n \times m - 1} \quad (4)$$

In the case of $n \geq 1$ and $m \gg 1$, we have

$$\text{Pr} = \frac{m}{n \times m} = \frac{1}{n} \quad (5)$$

Hence, the inverse of Pr measures n , the number of progenitors.

In reality, often the cell phylogeny of sampled cells is available. Similar to equation (2), the Pr of a focal cell type will be calculated as

$$\text{Pr} = \frac{\sum C_{m_i}^2}{C_{N_s}^2} = \frac{\sum (m_i \times (m_i - 1))}{N_s \times (N_s - 1)}, \quad (6)$$

where m_i is the number of terminal nodes (sampled cells) in the i th monophyletic clade of the focal cell type, and N_s is the total number of sampled cells of the focal cell type. According to the preceding derivation shown in equation (5), the inverse of the obtained Pr is an estimate of the number of progenitors, denoted as N_p :

$$N_p = \frac{1}{\text{Pr}} \quad (7)$$

Assessing the effects of various descendant numbers

Because of the heterogeneity of progenitor cells and environmental fluctuation during development, the numbers of descendants are variable among progenitors. We thus further evaluated the effect of the distributions of descendant numbers, including normal distributions, uniform distributions and bimodal distributions. In the first two panels of Fig. 1c, the number of descendants produced by a progenitor cell follows a normal distribution with the mean of 1,000 and the standard deviation of 100 or 200, respectively. In the third and fourth panels of Fig. 1c, the numbers of descendants follow a uniform distribution that ranges from 900 to 1,100 or from 800 to 1,200. In the fifth panel of Fig. 1c, 90% of the progenitors produce descendants by following a normal distribution with the mean of 1,000 and the standard deviation of 100, and the rest 10% produce descendants by following a normal distribution with the mean of 500 and the standard deviation of 50. In the sixth panel of Fig. 1c, 50% of the progenitors produce descendants by following a normal distribution with the mean of 1,000 and the standard deviation of 100, and the rest 50% produce descendants by following a normal distribution with the mean of 500 and the standard deviation of 50. To carry out the simulation, we created an *in silico* population seeded by 100 progenitors, each with the corresponding number of descendants based on the chosen distribution in Fig. 1c. Then 1% of the descendant cells were used to estimate N_p by referring to equation (1). Each simulation was conducted with 1,000 repeats. Besides, it is worth noting that we considered ever more complex distributions for the number of descendants per progenitor, and the corresponding results were collected in Supplementary Note 1.

Building an *in silico* cell phylogeny

To validate the coalescent method, we conducted simulations that mimicked a typical two-stage developmental process. In the first stage, we started with 1,000 progenitor cells divided from a zygote. These progenitors underwent divergence at different times, committing toward ten distinct cell fates (denoted as O0 to O9). We achieved this by generating a random topology with the R package *ape* and then dividing it into five intervals according to the depth of terminal nodes. For each interval, terminal nodes would be randomly assigned to two different cell fates. In the second stage, each progenitor cell would further produce 1,000 descendants on average, following a normal distribution with a standard deviation of 100. Specifically, we also used *ape* to construct the subtrees of descendants. All descendants were annotated with the cell fate of its ancestral progenitors, resulting an ideal cell phylogeny with a total of 998,093 terminal nodes (descendant cells). Additionally, we conducted similar simulations for phylogenies seeded by 2,000 or 3,000 progenitors. However, the number of descendants in these cases followed different normal distributions, with a mean of 500 or 300, respectively, while maintaining the same standard deviation.

Testing the effects of sampling strategy

First, we conducted a proportional sampling by randomly picking -0.01–0.5% (that is, -100–5,000 cells) of the descendant cells in the corresponding phylogenies to derive N_p , with 1,000 repeats in each case. We confirmed the overall good correlation between estimates and the actual numbers of the ten cell types (O0–O9). Next, we calculated the accuracy for each cell type by the formula:

$$\text{Accuracy} = \max\left(1 - \frac{|\hat{y}_i - y_i|}{y_i}, 0\right), \quad (8)$$

where y_i denotes the actual progenitor number of the i th cell type, and $\hat{y}_i$ the estimated progenitor number (N_p or sheer number of monophyletic clades). Notably, an Accuracy of zero is assigned when

the estimated progenitor number is over two times of the actual progenitor number. The median Accuracy of the ten cell types was considered in Fig. 2. Second, we conducted a fixed sampling by randomly picking 100 cells from each cell type in the corresponding phylogenies to assess the performance of TarCA, and the results are shown in Fig. 2.

Testing the effects of phylogenetic reconstruction algorithms

First, we obtained a set of DNA barcodes using PRESUME³⁸ by referring to a given input phylogeny. By following the manual of PRESUME, the parameter -L 1000 denotes the length of each barcode is 1,000, -m 1 denotes the mean of gamma distribution for relative substitution rates of different sequence positions is 1, and -gtrgamma default is set to be (1) the rate of six types of substitution is 0.03333, (2) the set of four nucleotide frequencies is 0.25, and (3) the global rate parameter is conducted with G4. The input phylogeny is composed of 1,000 cells proportionally sampled from the terminal cell pool of the *in silico* organism shown in Fig. 2a, with 100 replicates. Then, we reconstructed the phylogenies with five widely used softwares: (1) RapidNJ⁶⁵, a method based on neighbor joining algorithm, with parameters -b 1000 for running bootstrap; (2) MPBoot⁶⁶, a method based on maximum parsimony algorithm, with parameters -bb 1000; (3) FastTree⁶⁷, a method based on maximum-likelihood algorithm, with parameters -n 1000, -nt and -gtr; iv. RAXML⁶⁸, a classical software based on maximum likelihood algorithm, with parameters -m ASC_GTRCAT, -asc-corr=lewis, -p 2023; v. IQ-TREE⁶⁹, a time-efficient method based on maximum likelihood algorithm, with parameters -m TESTONLY to choose best model for phylogeny reconstruction.

Testing the effects of reducing the number of informative sites or lineage invasions

Similar to the above analysis, the input phylogenies are each composed of 1,000 cells proportionally sampled from the terminal cell pool of the *in silico* organism. In each case, the phylogeny information was used to simulate DNA barcodes, with a total length of 1,000 base pairs (bp). Then, we randomly excluded a proportion (up to 99%) of the sites on the barcodes and used only the remaining sites for phylogeny reconstruction by IQ-TREE.

We also tested different levels of invasions, ranging from 0.1% to 90%. For each level, the corresponding proportion of the cells in an input phylogeny were selected and their cell type identities were assigned randomly (O0–O9). This process resulted in a cell phylogeny with the original topology but mislabeled cells, simulating the effect of lineage invasions.

Testing the effects of sampling rate in a reconstructed phylogeny with errors

While it is extremely difficult to achieve a high sampling rate for a whole organism with millions to billions of cells, it is plausible to do so for a specific cell type. To simulate this scenario, we started with a guide cell phylogeny comprising 1,000 cells, as one examined in Fig. 2b, and gradually added cells of a focal cell type (O1 in this case) to the phylogeny. Considering the inevitable errors during phylogeny reconstruction, an added cell would attach to an incorrect phylogenetic position at a given probability. Consequently, the number of cells located in incorrect positions would follow a binomial distribution, which can be represented as follows:

$$N_{\text{err}} \sim B(N_{\text{add}}, \mu), \quad (9)$$

where N_{err} denotes the number of incorrectly-placed cells, N_{add} denotes the total number of added cells, and μ represents the error rate. This sampling procedure was repeated 500 times. Then, the performance of TarCA and sheer number of monophyletic clades in estimating the progenitor number of the focal cell type (O1) is compared.

Evaluating the performance of TarCA in a third-party dataset provided by Fang et al.

The simulated phylogenies, based on 331 QFMs, were obtained from the supplementary files provided by Fang et al.³⁷. This dataset comprises a total of 3,310 phylogenies (331 QFMs $\times$ 2 sampling strategies $\times$ 5 repeats). Each QFM contains detailed information about the population history, including the differentiation relationship between cell types, divergence time, differentiation bias and other relevant factors. Following the description of Fang et al., we calculated the population size of progenitors at different time points, and used the initial population size of each cell state as its progenitor number. Subsequently, we applied TarCA to the simulated phylogenies. For each terminal cell state, the progenitor number could be directly estimated using TarCA. For an intermediated cell state, we grouped all their descendants as a population to estimate N_p . In addition, the recovery rate of progenitors is defined by the fraction of recaptured progenitor lineages (that is the sheer number of monophyletic clades; Supplementary Note 2 and Supplementary Fig. 7).

Inferring the number of progenitor cells in nematode and fly cell phylogenies

The *C. elegans* dataset is from Sulston et al.³⁹ and the *P. marina* dataset is from Yuan et al.⁷⁰. The cell type information of both nematodes was downloaded from Yuan et al.⁷⁰. The median number of each cell type is 42.5 in *C. elegans* and 80 in *P. marina* (the lower quantile is 18.25 and 25, and the upper quantile is 100 and 98, respectively). Since the complete cell phylogeny is available, the sheer number of monophyletic clades of each cell type measures the actual progenitor number.

The two fly cell phylogenies were downloaded from the supplementary file provided online²⁵. The reported numbers of progenitors of the fly tissues/organs were curated from previous studies presented in Supplementary Table 1.

Inferring the credible interval of N_p

To evaluate the statistical reliability of N_p , we employed Bayes' theorem to infer the distribution of posterior probability of TarCA estimation. First, we formulated our prior belief about the probability distribution of N_p as $p(\theta)$, and then the observed Pr , denoted as x , has a likelihood $p(x, |, \theta)$. The posterior probability of N_p , denoted as $p(\theta, |, x)$, is given by:

$$p(\theta, |, x) = \frac{p(x, |, \theta)p(\theta)}{p(x)}, \quad (10)$$

where $\theta = 1, 2, 3, \dots$, and $p(x)$ is a normalizing constant. $p(x)$ is obtained through the summation of all possible values of θ , which is

$$p(x) = \sum p(x, |, \theta_i)p(\theta_i), \quad (11)$$

where θ_i means the i th element within all possible values of θ . To establish the prior distribution with limited θ , we employed three widely used distributions: uniform distribution, normal distribution and exponential distribution, each with or without $\log_{10}$ transformation of θ . Then, we selected the model that exhibited the best performance as $p(\theta)$. In this study, the normal distribution after $\log_{10}$ transformation was used in all analyses.

Due to the small sample size relative to the actual number of cells, the possibility of obtaining a monophyletic clade composed of m cells is approximated by a Poisson distribution. Given a hypothetical cell population formed by θ progenitors, the expected clade size for each monophyletic clade is N_s/θ . Hence, we have

$$\{m_i\} \sim \text{Pois}(N_s/\theta), \quad (12)$$

where N_s denotes the exact sample size of the focal cell population, and m_i denotes the number of descendants of the i -th progenitor.

The availability of m_i enabled the estimation of Pr according to equation (1). After 1,000 repeats we obtained the likelihood $p(x, \theta)$, and then the posterior probability distribution was derived according to equation (10). Finally, the 95% equal-tailed credible interval for each N_p was inferred from the posterior probability distribution.

Inferring progenitor numbers in mouse cell phylogenies

Based on the previous study of Chan et al., we utilized maximum-parsimony-based Cassiopeia to selectively reconstruct the cell phylogeny of embryo 3 and embryo 6 (ref. 27). We selected the two out of six embryos examined in that study on the basis of two criteria (Supplementary Note 3): first, the embryo should contain at least 1,000 unique barcodes; second, over 50% of the cells should be separated from each other by clear bifurcations in the phylogenies. To minimize the effects of polytomies (multiple lineages descending from a single node) in the two phylogenies, we removed internal nodes with over 100 children. Meanwhile, we integrated the scRNA-seq data of the two embryos with TOME database, which systematically describes the differentiation of cell states along mouse embryogenesis⁵⁰, using the standard pipeline of data integration by identifying ‘anchors’ in Seurat/v4. As a result, we annotated the cell types for cells of the two selected embryos. There are 15 cell types with at least one coalescent event in both cell phylogenies. These cell types are included in Fig. 4.

Assessing the expression trajectory complexity in mouse embryos

For a focal cell type at a given stage (shown in Fig. 4b), the cells are contributed often by multiple earlier source cell types. The data provided by the Pijuan-Sala et al. dataset of TOME tells the number of source cells that contribute to the focal cell type for each source cell type. As such, we obtain $p_i = N_i / \sum N_i$, where N_i is the number of source cells contributing to the focal cell type for the i th source cell type, and p_i is the relative contribution of the i th source cell type to the focal cell type. Then, a Shannon’s diversity index of the focal cell type at the given stage can be calculated as

$$\text{Shannon's diversity} = - \sum p_i \log_2 p_i \quad (13)$$

After considering all applied stages, the average Shannon’s diversity of different stages is used to measure the trajectory complexity of the focal cell type.

Investigating the transcriptome-wide expression divergence of gut endoderm cells in mouse embryos

For the gut endoderm cells we annotated in the two embryos we performed a re-clustering analysis using the standard workflow of Seurat/v4. We used the top 30 principal components for nearest-neighbor graph construction and set the parameter resolution to 1 for cluster determination. As a result, we identified five subpopulations in embryo 3 and six subpopulations in embryo 6. To assess if a subpopulation is concentrated in the corresponding cell phylogeny, we applied TarCA to estimating the progenitor number of the subpopulation. We then compared the observed N_p with the expected N_p of the same number of randomly selected cells in the phylogeny (with 1,000 replicates). If the observed N_p is statistically smaller than the expectation, the cells of this subpopulation are supposed to be derived from specific lineages; in other words, they are supposed to be fate committed. Otherwise, these cells are not fate committed yet.

Extracting LUGs

Each time one gene is tested to see if it is an LUG. The two cell phylogenies examined in this study each consisted of approximately 500 gut endoderm cells. We selected from a phylogeny the top 10% (about 50 cells) of cells that show the highest expression of the focal gene. Our main objective was to assess whether the selected cells as a whole

tend to have a low lineage diversity compared to the same number of randomly selected cells in the phylogeny. To achieve this, we employed TarCA to estimate the N_p of the selected cells as a population, and then compared it with the expectations that were derived from 1,000 populations of the same number of randomly selected cells. The focal gene was defined as an LUG if the obtained N_p is significantly ($P < 0.01$) smaller than the expected. To account for dropout events in scRNA-seq data, which would bias the estimation of gene expression level, we here used a loose definition of monophyletic clade. Specifically, a monophyletic clade of a focal cell population refers to a clade that is significantly enriched with cells of the focal population compared to its parental clade. For more details, the codes used in the above analyses can be found at GitHub.

Building a SVM to infer the fate-committing probability of phylogenetic clades with LUGs

There are a total of 155 LUGs defined in embryo 3. For each phylogenetic clade we constructed a binary vector with 155 elements to describe the states of the 155 LUGs. The elements were assigned ‘1’ if the corresponding genes were defined as an LUG in the focal clade, and ‘0’ if not. To reduce noise, phylogenetic clades composed of less than five cells were excluded from further analyses. Then, we obtained a $l \times k_Q$ query Boolean matrix Q , where $l = 155$, and k_Q is the number of clades analyzed. Meanwhile, we examined mouse gut endoderm cells at E8.75 and constructed for each cell a binary vector that stores the expression states of the 155 LUG genes. Specifically, the element was assigned ‘1’ if the corresponding gene is highly expressed (top 10%) in the focal cell compared to the other cells, and ‘0’ if not. As such, we obtained a $l \times k_R$ reference Boolean matrix R , where $l = 155$, and k_R is the number of gut endoderm cells under examination.

Based on the reference matrix R and the corresponding committed organs at E8.75, we built an SVM using 70% of the gut endoderm cells to classify the seven organs (thymus, thyroid, lung, liver, pancreas, small intestine and colon). We evaluated the performance of the SVM classifier using the macro average of multi-class ROC in the 30% testing cells (area under curve = 0.91). Then, the classifier was applied to the query data Q to predict the fate-committing probability toward the seven organs for each phylogenetic clade. Using essentially the same procedure we also built an SVM classifier (area under curve = 0.93) for embryo 6.

The phylogenetic clades of an embryo were grouped into several classes according to their fate-committing probability profiles by using an R package *apcluster* with cosine similarity. In addition, in Fig. 5e we performed a reversed graph embedding using a R package *DDRTree* to project the $l \times (k_Q + k_R)$ Boolean matrix $[Q|R]$ for the clades and the gut endoderm cells at E8.75 into a two-dimensional space. For more details the codes used in the above analyses can be found at GitHub.

Analyzing the static barcoding dataset that traces the in vitro differentiation trajectory of hematopoiesis

We obtained the single-cell transcriptome dataset of hematopoietic progenitor cells from Wang et al. (<https://cospar.readthedocs.io/>). The tracing of hematopoietic progenitor cell differentiation in culture by serial sampling at days 2, 4 and 6 after barcoding was conducted by Wang et al. to find genes associated with the differentiations in the *Gata1*⁺ states. *Gata1*⁺ states contain two types of progenitor. The first type of progenitor gives rise to megakaryocytes and erythrocytes (MEP), and the other type produces basophils, eosinophils and mast cells (BEMP). We first excluded the cells without barcode information and then examined the datasets of the three sampling time points separately.

We applied the LUG searching algorithm, as used in the mouse gut endoderm cells, to each dataset to identify genes associated with cell fate commitments (see Supplementary Note 4 for details). To adapt to the static barcoding data we modified our LUG searching algorithm by

treating cells with the same static barcode as a monophyletic clade. For a to-be-tested focal gene, we selected the top 10% of cells that show the highest expression of the focal gene. We then calculated the N_p of the selected cells and compared the obtained N_p to the expected number derived from the same number of randomly selected cells (with 1,000 repeats). The focal gene will be identified as an LUG if the obtained N_p is significantly ($P < 0.01$) smaller than the expected. We used the ‘clusterProfiler’ R package to perform GO enrichment analysis on the three fate-associated gene sets, respectively. The hierarchy of the enriched GO terms were then extracted by QuickGO, a web-based tool for browsing GO and the associated annotations. For more details the codes used in the above analyses can be found at GitHub.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The Yuan et al. dataset of *C. elegans* and *P. marina* is available at https://github.com/helloicyvodka/DELTA_code. Cell phylogenies of *D. melanogaster* can be found in source data files of Liu et al. (<https://doi.org/10.1038/s41592-021-01325-x>). The data about cell phylogenies of mouse embryogenesis are available in the Gene Expression Omnibus database under accession number GSE117542 and data of annotation from Pijuan-Sala et al. are available via <http://tome.gs.washington.edu>. Simulated phylogenies constructed by Fang et al. are available at <https://doi.org/10.5281/zenodo.7112097>. The data of hematopoiesis can be accessed at <https://cospar.readthedocs.io/>. All source data supporting the findings of the present study are available at https://github.com/shadowdeng1994/TarCA_sourcedata. Source data are provided with this paper.

Code availability

All custom codes for processing the data are available at <https://github.com/shadowdeng1994/TarCA>.

References

65. Simonsen, M., Mailund, T. & Pedersen, C. N. S. in *International Workshop on Algorithms in Bioinformatics* 113–122 (Springer, 2008).
66. Hoang, D. T. et al. MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC Evol. Biol.* **18**, 11 (2018).
67. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).

68. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
69. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
70. Yuan, M. et al. Alignment of cell lineage trees elucidates genetic programs for the development and evolution of cell types. *iScience* **23**, 101273 (2020).

Acknowledgements

We are grateful to W. Zhai, Z. Hu, J. Yang and Y. Zhang for comments. This work was supported by the National Key R&D Program of China (2021YFA1302500 and 2021YFA1302501), the National Natural Science Foundation of China (32293190, 32293191, 31970570 and 32200492).

Author contributions

X.H. and S.D. conceived the study. S.D. and H.G. did the simulation and analyzed the data. D.Z. and M.Z. analyzed the third-party data. X.H. supervised the study. X.H., S.D. and H.G. wrote the manuscript with contribution from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41592-024-02189-7>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41592-024-02189-7>.

Correspondence and requests for materials should be addressed to Xionglei He.

Peer review information *Nature Methods* thanks Manu Setty, Michael Stumpf and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Madhura Mukhopadhyay, in collaboration with the *Nature Methods* team. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.