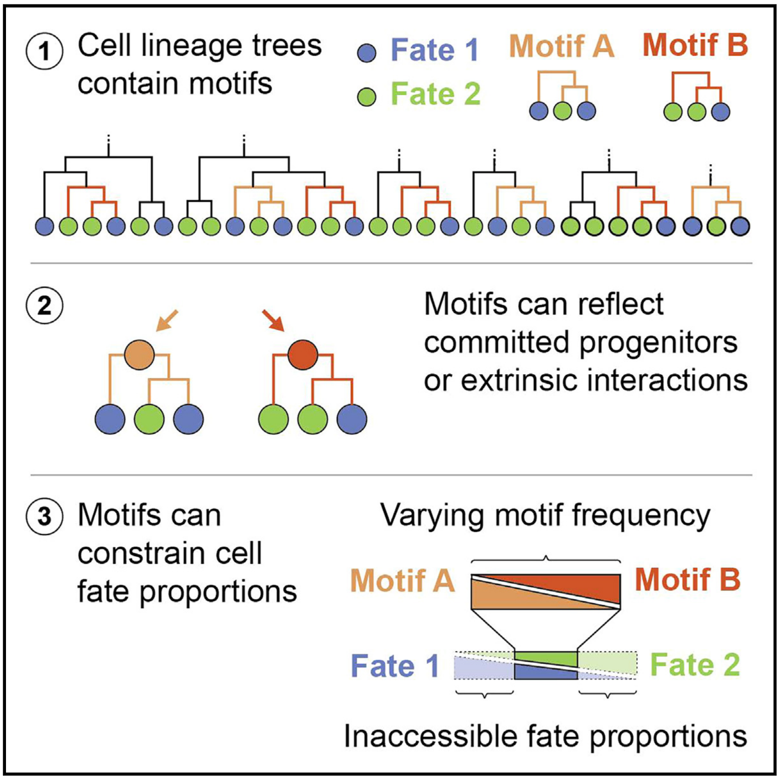
Developmental Cell

Technology

Lineage motifs as developmental modules for control of cell type proportions

Graphical abstract


1 Cell lineage trees contain motifs
Fate 1
Fate 2
Motif A
Motif B
2
Motifs can reflect committed progenitors or extrinsic interactions
3 Motifs can constrain cell fate proportions
Varying motif frequency
Motif A
Motif B
Fate 1
Fate 2
Inaccessible fate proportions

Highlights

- Lineage motifs are overrepresented patterns of cell fates on lineage trees
- Lineage motifs could reflect committed progenitors or extrinsic interactions
- Lineage motifs are identified in existing retina and early embryo lineage datasets
- Lineage motifs could facilitate adaptive variation in cell type proportions

Authors

Martin Tran, Amjad Askary, Michael B. Elowitz

Correspondence

amjada@g.ucla.edu (A.A.),
melowitz@caltech.edu (M.B.E.)

In brief

Biological tissues require fine-tuned cell type proportions for optimal function, but how this process is regulated remains poorly understood. Tran et al. suggest that lineage motifs reflect modular developmental programs that could constrain variation in cell type proportions.

Tran et al., 2024, Developmental Cell 59, 812–826
March 25, 2024 © 2024 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.devcel.2024.01.017

CellPress
OPEN ACCESS

Developmental Cell

Technology

Lineage motifs as developmental modules for control of cell type proportions

Martin Tran,[1] Amjad Askary,[2,*] and Michael B. Elowitz[1,3,4,*]
[1]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA
[2]Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, Los Angeles, CA 90095, USA
[3]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
[4]Lead contact
*Correspondence: amjada@g.ucla.edu (A.A.), melowitz@caltech.edu (M.B.E.)
https://doi.org/10.1016/j.devcel.2024.01.017

SUMMARY

In multicellular organisms, cell types must be produced and maintained in appropriate proportions. One way this is achieved is through committed progenitor cells or extrinsic interactions that produce specific patterns of descendant cell types on lineage trees. However, cell fate commitment is probabilistic in most contexts, making it difficult to infer these dynamics and understand how they establish overall cell type proportions. Here, we introduce Lineage Motif Analysis (LMA), a method that recursively identifies statistically overrepresented patterns of cell fates on lineage trees as potential signatures of committed progenitor states or extrinsic interactions. Applying LMA to published datasets reveals spatial and temporal organization of cell fate commitment in zebrafish and rat retina and early mouse embryonic development. Comparative analysis of vertebrate species suggests that lineage motifs facilitate adaptive evolutionary variation of retinal cell type proportions. LMA thus provides insight into complex developmental processes by decomposing them into simpler underlying modules.

INTRODUCTION

Most tissues comprise multiple specialized cell types that appear in appropriate proportions to support proper tissue-level functions. In many cases, cell type proportions vary spatially within the tissue. For example, the center of the primate retina is cone-dense, allowing for high visual acuity, while the periphery is rod-dense, enabling greater sensitivity in low light conditions.[1] Cell type proportions also vary between species. For instance, the ratio of rod and cone photoreceptors varies depending on the visual needs associated with the lifestyle of each species.[2] Tissue development thus faces the fundamental challenges of (1) generating cell types in correct proportions, and (2) facilitating spatial and evolutionary changes in those proportions.[3,4]

One prevalent mechanism for specifying cell type proportions occurs through regulating cell fate differentiation. As progenitor cells undergo successive rounds of cell division, they progressively become restricted in their fate potential, eventually committing to terminal cell fates. This process can be described in terms of a collection of cell states and the rates at which cells in each state transition to other states, i.e., a cell state transition map[5] (Figures 1A and 1B). In some cases, like the nematode *C. elegans*, cell state transitions are deterministic, producing a stereotyped lineage tree in all individuals.[6] However, in most other organisms, one cannot infer a quantitative cell state transition map from any single lineage tree due to variability. For example, in the mammalian retina, individual progenitor cells can give rise to a wide distribution of cell numbers and types with no apparent fixed ratios between different types. This observation prompted investigators to initially suggest a stochastic view of cell fate determination.[7,8] However, other studies of terminally dividing progenitors with particular expression patterns provided evidence for consistent cell-intrinsic biases in cell fate decisions.[9–15] These biases also appear in earlier, non-terminal divisions.[11,16–22] Cell state transition dynamics can also integrate extrinsic signals, developmental time, and stochastic "noise" with internal progenitor states.[23,24] Thus, even in well-studied systems such as the retina, it remains a major challenge to quantitatively elucidate cell state transition maps.

Different cell state transition maps can generate distinct distributions of cell fates on lineage trees. One simple transition map comprises a multipotent progenitor that can directly and probabilistically differentiate into multiple terminal fates (Figure 1A). A system employing such a direct, memoryless transition map would not exhibit fate correlations between related cells. Alternatively, a more complex transition map could involve the probabilistic generation of various types of committed progenitors, each predetermined to give rise to an invariant set of descendant cell types (Figure 1B). In this case, each type of progenitor would produce a characteristic distribution of descendant cell fates, introducing fate correlations on lineage trees. These fate correlations represent lineage

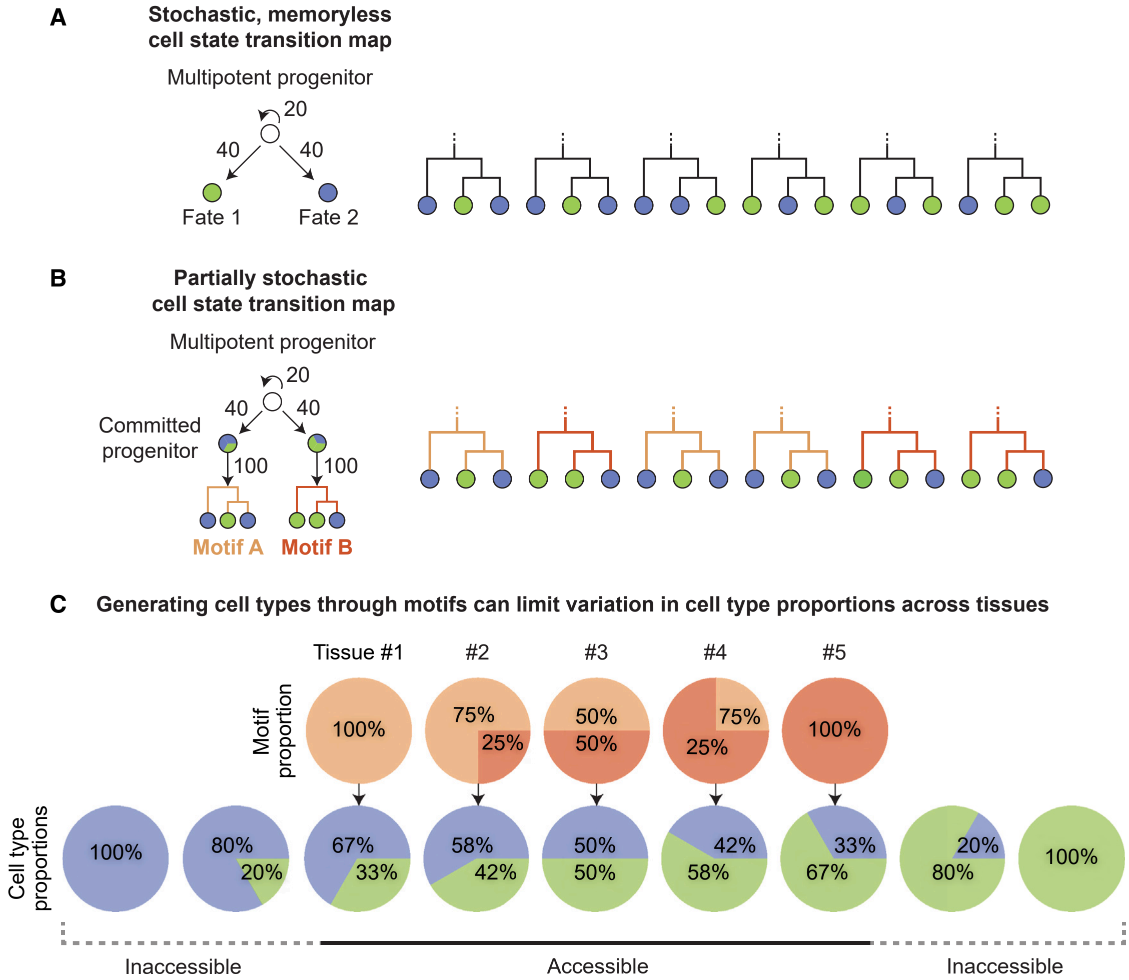


Figure 1. Cell type proportions can be controlled using partially stochastic cell state transition maps that specify defined groups of cell types as motifs

(A) A completely stochastic cell state transition map where a multipotent progenitor can self-renew or give rise to different fates in a memoryless manner. Lineage trees (only triplets shown) generated under this transition map would not exhibit fate correlations between related cells.

(B) A partially stochastic cell state transition map where a multipotent progenitor can self-renew or give rise to different types of committed progenitors. The committed progenitors differentiate, and each gives rise to a defined set of cell types (motif A or B). Lineage trees generated under this transition map would exhibit fate correlations between related cells, representative of the committed progenitors present within the transition map.

(C) In tissues that specify cell types solely by modulating the frequency of motif A and B, variation in cell type proportions is capped such that a cell type can be at most twice as abundant as the other type.

motifs that reflect otherwise hidden progenitor states. Furthermore, based on what motifs are used to specify cell types in developing tissues, this could in turn limit variation in overall cell type proportions (Figure 1C).

Recently, new methods have begun to allow for lineage tree reconstruction at scale. Long-term *in toto* live imaging allows direct tracking of dividing progenitor cells.[25] Additionally, a new generation of engineered lineage reconstruction systems has emerged.[26–32] These advances provoke the question of how fully resolved lineage trees with endpoint cell fates can be used to infer cell state transition maps.

To address this challenge, we introduce Lineage Motif Analysis (LMA), a computational approach for inferring statistically overrepresented patterns of cell fates on lineage trees. LMA is based on the principle of motif detection, which has been used to identify the building blocks of complex regulatory networks,[33] DNA sequences,[34,35] and other biological features,[36,37] but has not to our knowledge been applied to understand cell fate differentiation. As a "bottom-up," data-driven approach, LMA does not require specific assumptions about underlying molecular mechanisms and can be applied to diverse systems for which sufficient cell lineage information is available. Biologically, motifs could be generated by progenitors intrinsically programmed to autonomously give rise to specific patterns of descendant cell fates. They could also reflect more complex cell state transition maps involving extrinsic cues and cell-cell signaling that generate correlated cell fate patterns on lineage trees.

Here, we first define LMA and demonstrate how accurately it performs using simulated datasets. We then identify lineage motifs in published zebrafish and rat retina development datasets, as well as a dataset of early mouse embryonic development. These results reveal spatial and temporal differences in cell fate determination across different progenitors. Further, the appearance of shared retinal motifs across different species suggests that motifs may be evolutionarily conserved features of development. Computationally, we explore how various dataset characteristics affect motif identification. We demonstrate how the use of lineage motifs facilitates adaptive variation in retinal cell type composition and show that this theory is consistent with known variation in vertebrate retinal cell type proportions. Together, these results support LMA as a broadly useful tool to understand cell fate differentiation.

DESIGN

A previous study analyzed sister cell fate correlations by comparing the frequency of two-cell clones with that predicted by random association of cell types given their observed proportions.[24] Another study analyzed triplet fate correlations by comparing the frequency of triplet patterns with that observed in simulated lineage trees using a stochastic model where each starting progenitor can self-renew or differentiate into all possible cell types within the dataset under a set of probabilities.[38] These studies provide evidence for fate correlations between related cells. However, a framework that can be recursively applied to any lineage tree dataset to systematically identify lineage motifs of varying size remains lacking.

We first simulated a dataset of lineage trees with two terminal cell fates (Figure 2A; STAR Methods). We then applied LMA to analyze the tree dataset, starting by enumerating all possible doublet and triplet cell fate patterns (with varying fate composition and order of fate differentiation) and counting the number of times each occurred within the observed trees (Figure 2B). Then, we compared these counts with those expected in a "null" model without fate correlations. This can be done by randomly shuffling the cell fates at the leaves of the lineage trees to generate resampled trees, followed by counting the number of times each pattern occurs across the resampled trees. We then repeat the resampling process many times. Since the arrangement of cell types in the resampled trees are randomized, the average of counts obtained within the null model represents the expected count if there is no relationship between lineage and cell fate. To identify larger motifs that span more than one cell division, the resampling was done in a manner that preserves the frequency of sub-patterns within each pattern (STAR Methods).

For each pattern, we computed a *Z* score to quantify the degree of over-representation, as well as a false discovery rate (FDR)-adjusted p value[39,40] to measure significance (STAR Methods). In the identified lineage motifs, higher over-representation can be interpreted as stronger intrinsic commitment of a given progenitor toward generating a particular fate pattern. Alternatively, it could represent the strength of extrinsic interaction that generates a particular fate pattern. Finally, anti-motifs, defined as patterns that are underrepresented in the observed trees, were identified using the same approach.

LMA is distinct from a related approach termed Kin Correlation Analysis (KCA). KCA infers cell state transition dynamics from lineage trees and endpoint cell state datasets but is mainly applicable to systems governed by Markovian dynamics, in which sister cell transitions are independent of one another.[41,42]

To demonstrate that LMA can recover lineage motifs that reflect progenitor states in cell state transition maps, we simulated lineage tree datasets using either a competence progression model (Figure 3A) or a binary fate model (Figure S1A). We used differentiation probabilities that generate roughly equal cell type proportions in the overall dataset (Figures 3B and S1B). Applying LMA to both datasets, we found that the resulting motifs reflected the structure of the generative model and captured multiple levels of progenitor commitment over time. For example, in trees generated using a competence progression model (Figures S2A–S2C), where cell fates A through F are generated progressively over time, only symmetric doublet patterns, such as (F,F), were statistically overrepresented within all possible doublet patterns (Figure 3C).

We next analyzed triplet patterns, in which a single progenitor divides to produce a terminal cell, X, and a second progenitor cell that divides once more to produce a doublet of terminal cells, Y and Z, producing a triplet denoted as (X,(Y,Z)). Only triplet patterns including two sequential levels of progenitor commitment, such as (E,(F,F)), were significantly overrepresented (Figure 3D).

LMA can be scaled up to analyze larger asymmetric patterns. Given a reasonable number of trees (500 total), the motifs successfully captured up to 5 levels of the competence progression model. Similar to the triplet results, the significant higher-order motifs exclusively involved sequentially generated cell fates. As motif size grows larger, the size of the dataset required for detection also increases (Figure 3E). Together, these results confirm that LMA can be used to recursively identify lineage motifs in large patterns.

We also analyzed trees generated using a binary fate model in which progenitors make binary choices which restrict their fate potential over time (Figures S2D–S2G). The doublet and quartet motifs reflect the structure of the generative model as expected (Figures S1C and S1D). However, no octet patterns were significantly over- or underrepresented (Figures S1E and S1F). Taken together, these results indicate that LMA is capable of recursively identifying lineage motifs of multiple sizes in different models of development and is especially powerful when applied to the competence progression dynamics, likely due to the lower number of possible patterns per level of progenitor commitment.

Having demonstrated that LMA can recover motifs in lineage trees generated using an intrinsic program, we next sought to demonstrate that LMA could do so in trees generated using an extrinsic program. More specifically, we considered a simplified model of the classic developmental mechanism of lateral inhibition, in which cells of one fate inhibit similar fates in their neighbors[43–45] (Figures S3A–S3C). Our model assumes a two-dimensional grid of progenitors, which self-renew or differentiate into two cell fates, A or B, each of which inhibits differentiation of its neighbors into its own fate. The inhibitory

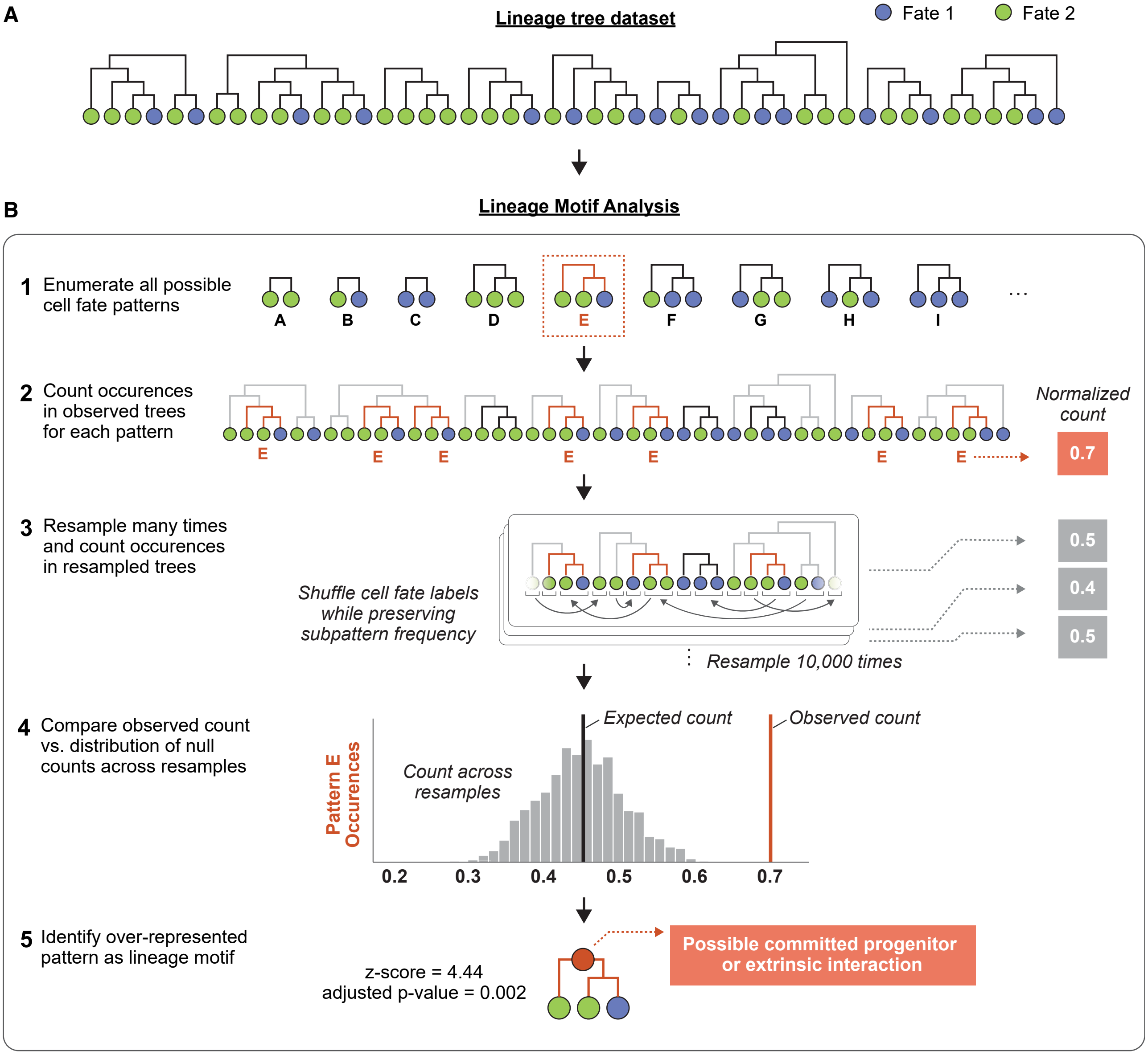

A
Lineage tree dataset
Fate 1
Fate 2
B
Lineage Motif Analysis
1 Enumerate all possible cell fate patterns
A
B
C
D
E
F
G
H
I
...
2 Count occurences in observed trees for each pattern
E
E
E
E
E
E
E
Normalized count
0.7
3 Resample many times and count occurences in resampled trees
Shuffle cell fate labels while preserving subpattern frequency
Resample 10,000 times
0.5
0.4
0.5
4 Compare observed count vs. distribution of null counts across resamples
Pattern E Occurences
Count across resamples
Expected count
Observed count
0.2
0.3
0.4
0.5
0.6
0.7
5 Identify over-represented pattern as lineage motif
z-score = 4.44
adjusted p-value = 0.002
Possible committed progenitor or extrinsic interaction

Figure 2. Lineage Motif Analysis identifies fate correlations in lineage trees by statistical resampling

(A) Lineage trees with two cell types were simulated (STAR Methods).

(B) The LMA workflow consists of the following steps. First, all possible cell fate patterns are enumerated. Second, the occurrence of each pattern within the observed lineage trees is counted (triplet pattern "E" is shown here as an example). Third, the cell fate labels at the leaves of the trees are randomly shuffled to obtain a resampled set of trees with no fate correlations. This process is then repeated across many resamples. To identify the higher-order motifs that span multiple cell divisions, the shuffling process is done in a manner that preserves sub-pattern frequency (STAR Methods). The occurrence of each cell fate pattern is then counted for each resample. Fourth, the count in the observed lineage trees is compared with the distribution of counts across resamples, whose average is approximately equal to the expected count if there were no fate correlations. Finally, overrepresented patterns are classified as lineage motifs, which represent possible committed progenitors or extrinsic interactions.

effects of multiple neighbors are assumed to combine additively (Figure S3B). As expected, applying LMA to a null dataset generated without lateral inhibition revealed no significantly overrepresented doublet patterns (Figure S3D). By contrast, symmetric sister doublets (A,A) and (B,B) were underrepresented in the lateral inhibition model, whereas the asymmetric sister doublet (A,B) was overrepresented. These results show that extrinsic developmental programs can generate signatures of fate correlation on lineage trees, which can be reliably detected using LMA.

To enable the identification of lineage motifs across diverse developmental contexts, we created a Python package, termed "linmo." The package is available on a GitHub repository (https://github.com/labowitz/linmo), which includes supporting documentation and tutorials for processing the following lineage tree datasets analyzed here.

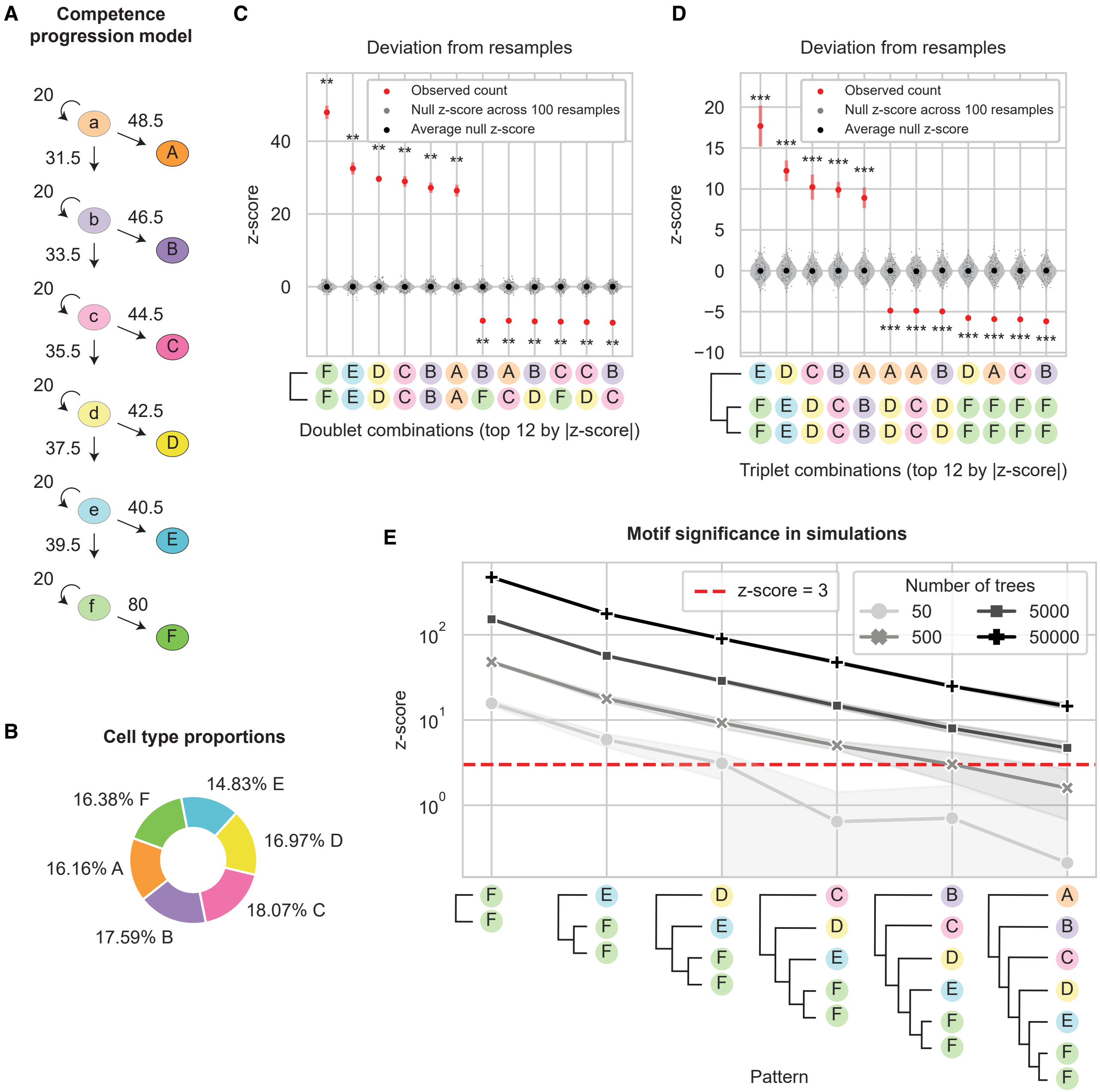


Figure 3. Lineage motifs reflect sequential progenitor states in a competence progression model

(A) Lineage trees were simulated using a competence progression model.

(B) Cell type proportions in 500 simulated lineage trees.

(C) Deviation score for doublet patterns. Null Z scores were calculated by comparing a random resample dataset with the rest of the resample datasets. 10 datasets containing 500 simulated trees each were used, with the standard deviation across the datasets plotted as error bars (** and *** represent adjusted p value < 0.005 and < 0.0005, respectively).

(D) Deviation score for triplet patterns.

(E) Deviation score for select patterns that reflect sequential differentiation of cell types using datasets of varying size. Shading indicates 95% confidence interval across 10 datasets for each point.

See also Figures S1–S3.

RESULTS

LMA reveals spatial organization of zebrafish retina development

Retina development provides a well-studied example of cell fate diversification. It involves generation of a conserved set of terminal cell fates across diverse vertebrate species. At the same time, it also exhibits substantial spatial and inter-species variation in cell type proportions,[1] making it an ideal target tissue for LMA. Therefore, we examined a zebrafish retina development dataset spanning 32 to 72 h post fertilization (hpf),[46] during which progenitors terminally differentiate to form major neuronal and glial cell types, including ganglion (G), amacrine (A), bipolar (B), photoreceptor (P), horizontal (H), and Müller glia (M) (Figure 4A). He et al. used time-lapse confocal microscopy in reporter zebrafish lines to track every cell division event for 60 retinal progenitors spanning the nasal-temporal axis. Their data supported previous work showing that a wave of differentiation starts in the nasal region and gradually progresses to the temporal region.[47,48] Clonal cell type composition was generally observed to be variable, with weak fate correlations between related cells. A key exception, however, was the frequent appearance of symmetric terminal pairs of photoreceptor, bipolar, and horizontal cells.

We sought to identify lineage motifs and characterize how their frequency varies across spatial regions in the zebrafish retina. Therefore, we partitioned lineage trees based on the progenitor spatial location and applied LMA, beginning with doublet patterns, representing the terminal cell division. We found that the (H,H), (B,B), and (P,P) doublet patterns had significantly higher observed counts in the lineage trees, compared with the distribution of counts across resamples and expected count, in a similar manner across the three spatial regions (Figures 4B–4D). Therefore, these doublet patterns are statistically overrepresented in the dataset and represent lineage motifs (Figure 4E). The exception was a lack of (H,H) and (B,B) doublets in the nasal region, likely because those cell types were only present at very low levels in this region (Figures 4D and 4E). These results were consistent with key findings from He et al., while extending the analysis to assess regional variation.

LMA also found motifs not previously identified in the He et al. study and revealed how their frequency varies across space. For example, even though amacrine and bipolar cells appear at similar frequencies across all three retinal regions, the (A,B) doublet was specifically overrepresented in the nasal region (Figures 4D and 4E). Also, doublets comprising one P cell and all other cell types were generally underrepresented across all regions, constituting anti-motifs. We also searched for higher-order motifs that involve multiple cell divisions but found that no patterns were significantly over- or underrepresented, possibly due to the limited size of the dataset (Figure S4). Overall, the observed motif profile suggests that amacrine and bipolar cells frequently share a common progenitor at the terminal cell division, specifically in the nasal region of the zebrafish retina, whereas photoreceptor and non-photoreceptor cells do not share a common progenitor at the terminal cell division in all regions.

Shared retinal lineage motifs across species suggest conservation of cell fate determination

Are retinal lineage motifs conserved between different species? To address this question, we analyzed a dataset of post-natal rat retinal progenitor cells grown *in vitro* at clonal density, consisting of 129 lineage trees with at least 3 cells.[38] During post-natal development, rat retinal progenitor cells gave rise to mostly rod cells (R), some bipolar and amacrine cells (respectively, B and A), and few Müller glia (M) (Figure 5A). In this work, the authors showed that a stochastic model based on independent fate decisions could explain the observed frequencies of most triplet patterns. However, some triplets may be generated by fate-committed progenitors that give rise to sets of correlated cell fates.

Applying LMA to this rat retina dataset confirmed some of these conclusions, such as over-representation of (B,(A,B)) triplets (Figures 5C and 5E). However, it also revealed additional features of rat retinal development. For example, using LMA, we found that (A,B), (B,M), and (A,A) doublets were overrepresented, whereas (B,R) doublets were underrepresented (Figures 5B and 5D). Correcting for sub-pattern frequencies in the triplet analysis revealed that the apparent over-representation of the (R,(A,A)) triplet in the previous study[38] could be entirely explained by the (A,A) doublet motif frequency. This highlights the importance of the recursive nature of LMA.

Because this dataset excluded two-cell lineages, this could potentially introduce biases in three-cell motifs. Therefore, we analyzed cell type proportions in triplets and compared this with those across all other cells (Table S1). We found that there are no obvious differences in cell type proportions between triplet and non-triplet populations, suggesting that the lack of two-cell lineages in the dataset does not substantially bias the triplet motifs detected here.

We next compared the motif profile between zebrafish and rat retina. Because the time period analyzed in these datasets is different and involves the generation of different cell types, we limited this analysis specifically to cell types that are shared between the analyzed datasets (i.e., amacrine, bipolar, and Müller glia). Notably, the (A,B) and (A,A) motifs are observed in both species, suggesting that the committed progenitors that these motifs possibly represent are at least partially evolutionarily conserved. In contrast, the (B,B) motif appears specifically in the zebrafish retina, whereas the (B,M) motif appears specifically in the rat retina. Overall, these data suggest that cell fate allocation in retina across species can occur in a biased and evolutionarily conserved manner, in which amacrine and bipolar cells share a common progenitor at the terminal cell division. At the same time, other aspects of cell fate differentiation may be more species-specific. For example, bipolar and Müller glia tend to share a common progenitor in rat, but not zebrafish, retina at the terminal cell division. More generally, these results provide a case example for how LMA can be used to assess the evolution of cell fate differentiation.

Computational simulations reveal how various dataset characteristics affect motif identification

To what degree the limited size of available lineage tree datasets, coupled with sampling variation, affects the accuracy of motif identification is unclear. Therefore, we simulated lineage tree

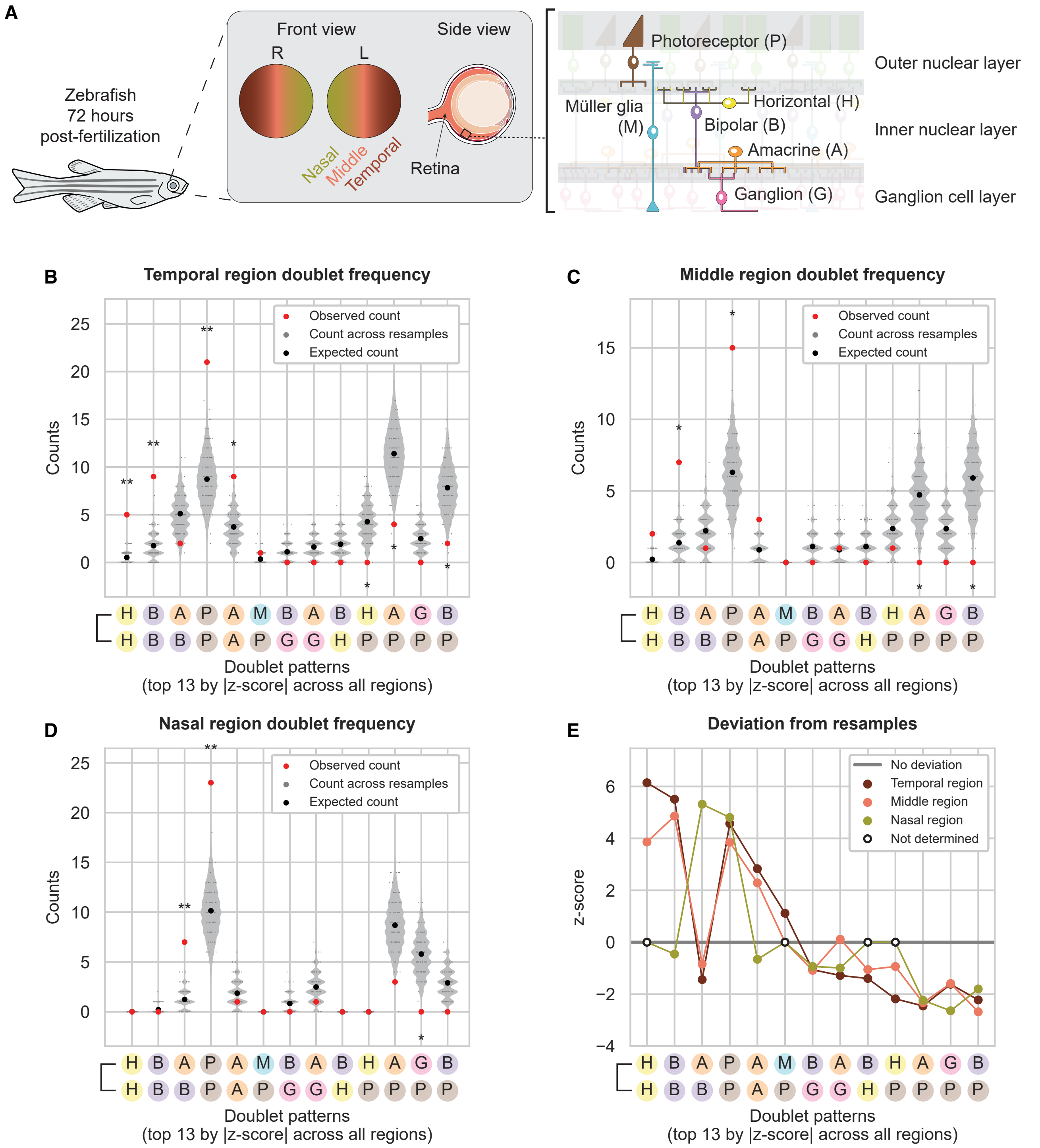


Figure 4. Doublet lineage motifs in zebrafish retina development show spatial organization of fate commitment

(A) Schematic of cell type organization in the zebrafish retina.

(B) Counts for doublet patterns in the observed zebrafish retina trees from He et al.[46] in the temporal region and across 10,000 resamples (* and ** represent adjusted p value < 0.05 and < 0.005, respectively). The expected count was calculated analytically (STAR Methods).

(C) Counts for doublet patterns in the middle region of zebrafish retina and across 10,000 resamples.

(D) Counts for doublet patterns in the nasal region of zebrafish retina and across 10,000 resamples.

(E) Deviation score for doublet patterns in the temporal, middle, and nasal region. Doublet patterns with an observed and expected count of 0 were omitted from the analysis.

See also Figure S4.

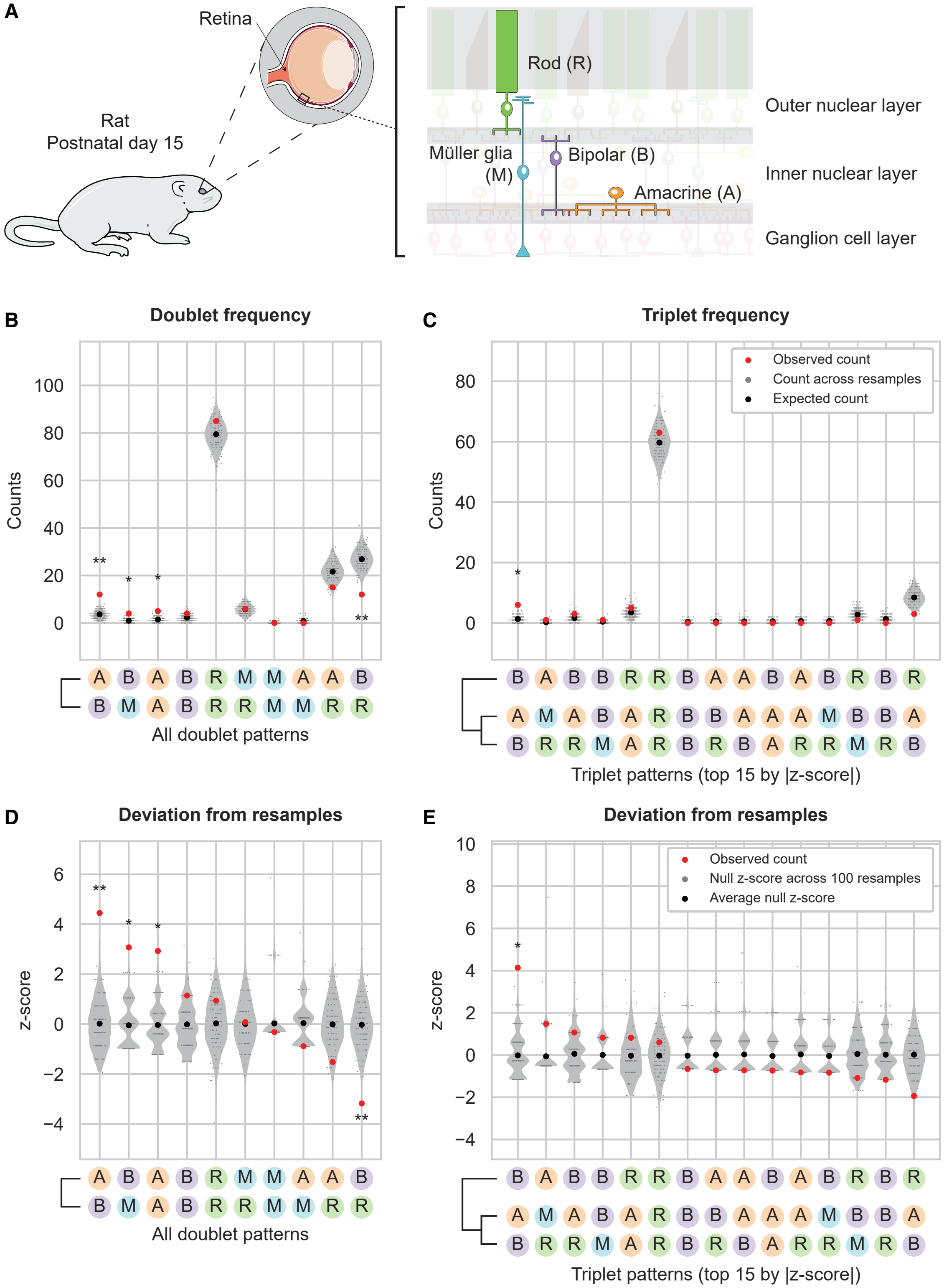


Figure 5. Doublet and triplet lineage motifs reveal fate commitment patterns in rat retina development

(A) Schematic of cell type organization in the rat retina.

(B) Counts for doublet patterns in the observed lineage trees from Gomes et al.[38] and across 10,000 resamples (* and ** represent adjusted p value < 0.05 and < 0.005 respectively). The expected count was calculated analytically (STAR Methods).

(C) Counts for triplet patterns in the observed lineage trees and across 10,000 resamples.

(legend continued on next page)

datasets using a stochastic model, using set division and fate probabilities based on the Gomes et al. dataset (Figure S5A), while varying dataset size. We also added varying levels of sister fate correlation, quantified as conditional probabilities, into the model. For example, the conditional probability of a rod cell also having a rod sister, P(R sister | R), would be the same as the overall frequency of rod cells, P(R), if no sister fate correlations were present. However, this conditional probability could be increased or decreased to reflect sister fate correlation or anti-correlation, respectively. To further explore how differences in absolute cell type frequency impact motif identification, we allowed sister fate correlations for either rod or Müller glia fates, which are present at 75% or 3%, respectively.

Analysis of the simulated lineage tree datasets revealed several conclusions. First, no motifs were detected when lineage trees were generated without sister fate correlations, across all tested dataset sizes (Figures S5B and S5C). This indicates that our thresholds for significance are stringent and ensures the absence of false positive motifs. Second, fate correlations are more strongly detected in larger datasets (Figures S5D and S5E). Conversely, they often go undetected within small datasets even if sister cell fates are completely correlated or anti-correlated. This suggests that analysis of small datasets is likely to be affected by high false-negative rates. Third, fate correlations are more strongly detected when the conditional probability deviates further from overall cell type frequency (Figures S5D and S5E). Fourth, fate correlations are more strongly detected for more abundant cell types. Overall, this analysis suggests that the fate correlations detected in both zebrafish and rat retina were very strong, since those datasets were relatively small (60 and 129 trees, respectively). However, there may also be weak fate correlations that were not detected as motifs due to limited dataset size.

LMA reveals temporal differences in cell commitment during early mouse development

Early embryonic development features conserved cell types across mammals and spatially restricted cell fate specification, making it an ideal system to apply LMA. We therefore used LMA to analyze a dataset of early mouse embryo development.[49] In a previous study, Morris et al. used time-lapse confocal microscopy to trace individual progenitor cells starting at the 8- to 16-cell division within 20 mouse blastocysts until their final fate is known at the late blastocyst stage (~E4.5). Beginning at the 16-cell stage, progenitors can either be located inside, contributing to the inner cell mass, or be located outside along the periphery (Figure 6A). (For shorthand, we will refer to the progenitors at the 16-cell stage simply as “inside” and “outside” progenitors.) During the next two rounds of cell divisions, outside progenitors can become internalized, adding to the inner cell mass, or continue to remain outside, eventually differentiating into trophectoderm (T). Cells within the inner cell mass either undergo apoptosis (A) or further differentiate into either epiblast (E) or primitive endoderm (P) fates. The authors found that inside progenitors are biased to give rise to epiblast, whereas outside progenitors that internalize later are biased to give rise to primitive endoderm. However, it remained unclear whether individual progenitors give rise to sets of correlated cell fates in the final few divisions prior to fate assignment.

We therefore partitioned lineage trees into those generated from inside or outside progenitors and applied LMA to both sets. We found that most doublet patterns (90%) within both types of progenitors are either motifs or anti-motifs (Figure S6). For example, symmetric sister pairs such as (P,P), (E,E), and the apoptotic doublet (A,A) were overrepresented among descendants of both inside and outside progenitors. (T,T) was also overrepresented among outside progenitors (Figure S6A). These results suggest that by E4.5, most cells have already committed to one of the three lineages before the previous cell division and therefore produce symmetric doublets. We also observed doublet motifs comprised multiple cell types, such as (A,P), (A,E), and (E,P), which were overrepresented in trees from outside progenitors while underrepresented in trees from inside progenitors. Trophectoderm, unlike the other cell fates, was part of all anti-motifs among outside progenitors. Overall, the weaker motif signatures for inside progenitors suggest less commitment compared with the strong, and usually symmetric, doublet motifs among the descendants of the outside progenitor cells (Figure S6C).

Analyzing higher-order motifs among outside progenitors, we strikingly observed both triplet motifs with multiple cell fates, such as (A,(P,P)) and (T,(A,P)), and the homogeneous triplet motif (P,(P,P)), suggesting the existence of committed progenitors at least two generations earlier (Figure 6B). In contrast, all triplet patterns for inside progenitors did not significantly deviate from null expectations, suggesting that they remain uncommitted toward defined fates (Figure 6C). Taken together, these results suggest that some outside progenitors may commit to give rise to defined groups of cell types at least two cell divisions before blastocyst formation, while inside progenitors remain plastic and uncommitted toward certain fates (Figure 6D).

Lineage motifs facilitate adaptive variation in cell type frequencies

The fitness landscape over cell type frequency space could in principle have peaks of high fitness, plateaus of relatively constant fitness, or valleys of low fitness. Inspired by the concept of Pareto optimality in evolutionary trade-offs,[3,4,50] we asked whether it is possible to structure the cell state transition map in such a way that allows cell type frequencies to disproportionately populate the high fitness, or more adaptive, regimes. We reasoned that lineage motifs could address this problem. Mathematically, lineage motifs represent a linear transformation from a set of motif frequencies to a set of cell type frequencies. If most cell fate decisions resulted in generation of cell types as motifs, then a developing tissue could indirectly control the frequencies of individual cell types by specifying the frequency of each motif (Figure 1C).

(D) Deviation score for doublet patterns in the observed lineage trees. Null *Z* scores were calculated by comparing a random resample dataset with the rest of the resample datasets.
(E) Deviation score for triplet patterns in the observed lineage trees.
See also Figure S5 and Table S1.

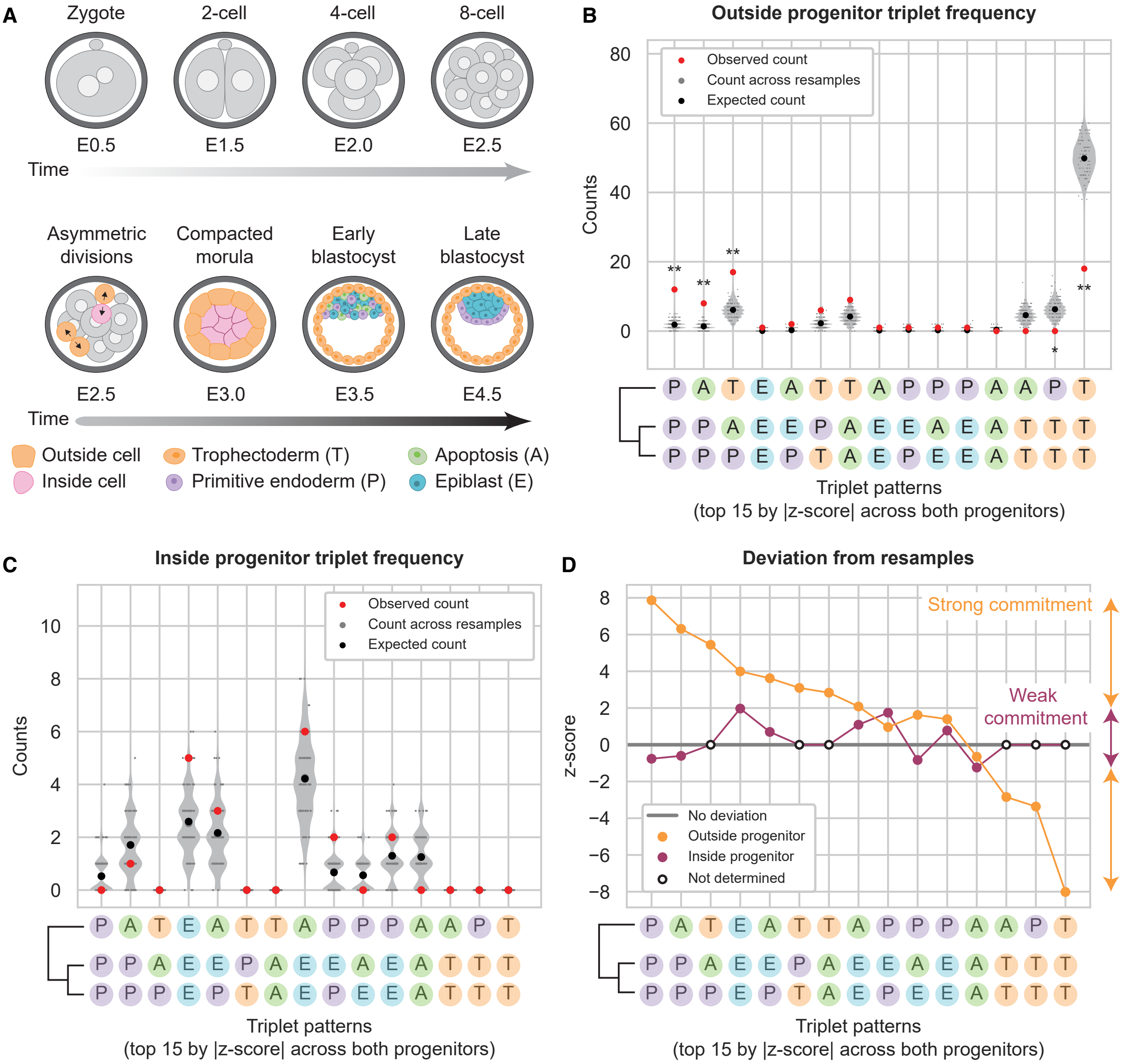


Figure 6. Triplet lineage motifs in mouse blastocyst development suggest that “outside” progenitors initiate fate commitment earlier than “inside” progenitors

(A) Schematic of early mouse embryo development and the cell types involved.

(B) Counts for triplet patterns in the observed mouse blastocyst trees from Morris et al.[49] in the outside progenitors and across 10,000 resamples (* and ** represent adjusted p value < 0.05 and < 0.005, respectively). The expected count was calculated analytically (STAR Methods).

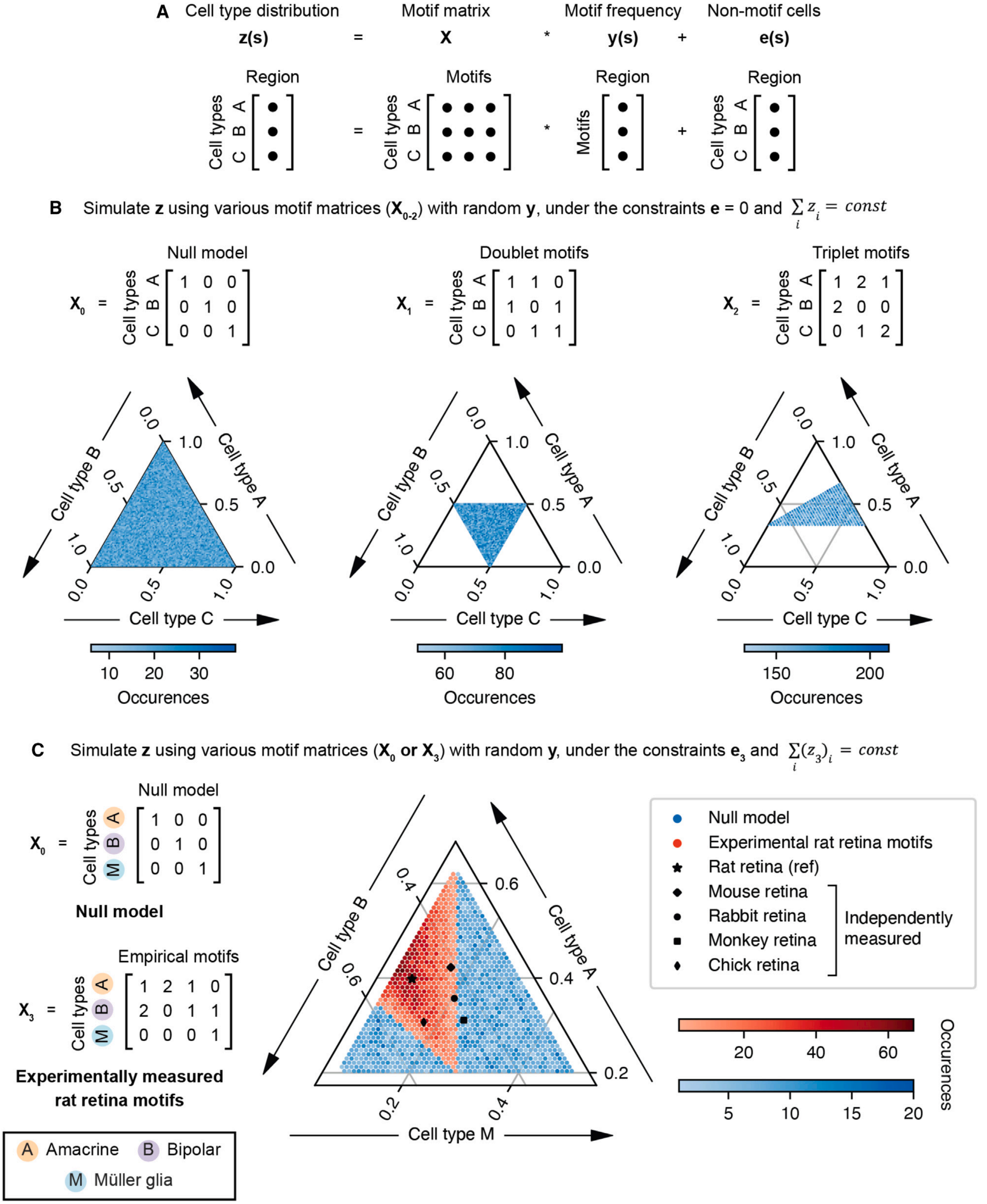
(C) Counts for triplet patterns in the observed mouse blastocyst trees in the inside progenitors and across 10,000 resamples.

(D) Deviation score for triplet patterns in the outside and inside progenitors. Triplet patterns with an observed and expected count of 0 were omitted from the analysis.

See also Figure S6.

More precisely, we can describe the conversion from motif frequencies to cell type distributions as a linear transformation: $z(s) = X \times y(s) + e(s)$. Here, $z(s)$ is a vector whose components represent the counts of each cell type in position/species s, X is a non-negative integer matrix describing how many cells of each type (rows) are produced by each motif (columns), $y(s)$ denotes the motif frequencies in position/species s, and $e(s)$ represents the number of additional cells of each type in position/species s that cannot be explained through the motifs (Figure 7A).

To understand how motifs constrain cell type frequencies, we first constructed a set of hypothetical motif matrices X_0, X_1, X_2, each reflecting a different motif structure. X_0 is a diagonal matrix representing the null model in which each column trivially corresponds to a single cell type. In contrast, X_1 and X_2 contain



(legend on next page)

exclusively doublet or triplet motifs, respectively, where multiple cell types are generated together (Figure 7B). For each motif matrix, we simulated datasets by randomly choosing frequencies for each of the motifs present within each matrix (STAR Methods). For simplicity, we initially assumed $e = 0$ and constrained cell type frequencies to sum to a constant, $\sum_i z_i = const$, to reflect the limited total capacity of the tissue. We then analyzed the range of cell type distributions produced by each matrix. In this framework, each randomly chosen y(s) would generate a particular z(s), which corresponds to one individual with a particular set of cell type frequencies. Under the null model, cell type frequencies spanned the full space, as expected. By contrast, the other two models restricted cell type frequencies to limited subspaces.

Although motifs can constrain cell type distributions in general, it remained unclear whether the specific motifs observed in the rat retina dataset would be consistent with the distributions of retinal cell types independently observed in different species. Addressing this question requires (1) defining the motif-accessible space of cell type frequencies permitted by the observed rat retina motifs, and (2) determining whether independently measured retinal cell type proportions from other vertebrate species lie within that space.

To define the motif-accessible frequency space, we first need to determine the lower and upper bounds for cell type frequencies. We set the lower bounds at e_3, the cell type counts for all cells in the rat retina dataset born outside of a motif (STAR Methods; Figure S7A). We set the upper bounds by constraining the total number of cells to be the same as in the rat retina dataset, $\sum_i (z_3)_i = const$. Using these constraints, we simulated datasets containing randomly chosen frequencies for each of the observed rat retina motifs in X_3, or as a control, the null model, X_0. The motif model accessed only a subset of the space of type proportions spanned by the null model (Figure 7C). Within this subspace, the motif model showed higher density of fate distributions corresponding to moderate levels of both amacrine and bipolar cells and low levels of Müller glia. Bipolar cells and Müller glia exhibited a reduced maximum proportion relative to the null model, consistent with the observation that both of these cell types are generated with other cell types in the rat retina motifs.

We compared the datasets generated using the motif or null model with independent measurements of retinal cell type proportions across multiple vertebrate species.[2,51] Strikingly, this analysis revealed that all of the independently measured fate distributions of vertebrate retina lie within, or very close to, the subspace accessed by the motif model. To understand how the structure of the motif matrix impacts the resulting cell type distributions, we repeated this analysis omitting the (A,A) motif from the motif matrix X_3 (Figure S7B). This resulted in a smaller subspace achieved by the motif model, specifically lowering the maximum proportion of A cells from 62.8% to 46.0% (Figure S7C). This perturbed model failed to capture the empirical cell type distributions, indicating that the (A,A) motif is required to explain variation in cell type proportions across vertebrate retina. Taken together, these results are consistent with the notion that motifs identified in lineage trees of rat retina could facilitate evolutionary variation in retinal cell type proportions across vertebrates. They further show that the range of cell type proportion space achieved using the motif model can be expanded or constrained by respectively increasing or decreasing the number of different motifs in the model.

DISCUSSION

Producing cell types in optimal ratios is essential for tissue function. In many contexts, these proportions are established during development, when intrinsically committed progenitors or extrinsic interactions generate sets of terminal cell types. Increasing recent attention to the role of lineage in development[52] and the emergence of new methods for reconstructing lineage trees[53,54] provoke the question of how one can infer committed progenitors or extrinsic interactions based on the arrangement of descendant cell fates on lineage trees.

In this work, we introduce a general computational approach, LMA, based on statistical resampling of lineage trees. Using simulations, we demonstrated that LMA can be recursively applied to uncover fate correlations in large patterns that span multiple cell divisions. By applying this framework to three biological datasets, we identified fate correlations, some of which validate known fate patterns. In the retina, motifs can recur across space, or appear specifically in different tissue regions. The presence of shared motifs across zebrafish and rat retina suggests evolutionary conservation of retinal cell fate determination. In early mouse development, inside progenitors at the 16-cell stage appear plastic and uncommitted toward certain fates compared with outside progenitors, when analyzing their last two cell divisions before blastocyst formation. Finally, we showed that the rat retina motifs, if utilized in different frequencies, could explain variation in cell type frequencies across several vertebrate species. Lineage motifs thus provide a useful and biologically meaningful lens through which we can analyze cell fate differentiation.

Lineage motifs could be regarded simply as the consequence of a differentiation process that requires the cells to pass through intermediate states of partial fate commitment. However, this explanation still leaves open the question of why certain commitment states have been selected for during evolution. One potential answer is that lineage motifs play functional roles in controlling cell type distributions. Developing organisms could use various regulatory strategies including intrinsic transcription factors or extrinsic signals as “knobs” to modulate motif frequencies. Since lineage motifs represent groups of cell types

Figure 7. Motifs can facilitate optimal variation in cell type frequencies between species

(A) The $z(s) = X * y(s) + e(s)$ matrix equation describes the linear transformation from motif frequencies to cell type distributions.

(B) Cell type distributions were simulated by randomly varying the frequencies of motifs using three example motif matrices (X_0, X_1, X_2).

(C) Cell type distributions were simulated using a null model (X_0) or the empirical motif matrix based on the rat retina motifs (X_3) in Figure 5. The independently measured cell type proportions of mouse, rabbit, monkey, and chick retinas from Masland[2] and Yamagata et al.[51] were overlaid on the ternary plot.

See also Figure S7.

produced in fixed stoichiometric ratios, this mechanism could restrict variation of cell type proportions during development or evolution to physiologically adaptive regimes. In the future, we anticipate greater availability of high-quality lineage datasets, which should allow more complete tabulation of motifs across different tissue contexts. These data should thus enable more stringent tests of the model proposed here.

A second potential role for lineage motifs could be to create spatially localized neighborhoods of interacting cell types to implement specific functions. In the context of the retina, particular types of interneurons must be synaptically connected. For example, in crossover inhibition, OFF bipolar cells receive input from ON amacrine cells, which are depolarized by ON bipolar cells at light onset.[55] A neural circuit of these cell types in close spatial proximity could be ensured by regulating the generation of these cell types through a lineage motif, such as the (B,(A,B)) motif observed in the rat retina dataset (Figure 5E). Consistent with this hypothesis, recent work has shown that specific synapses develop preferentially among sister excitatory neurons in the mouse neocortex.[56]

Lineage motifs can be compared with other methods for analyzing cell fate differentiation, such as pseudotime, where single cells are densely profiled throughout time to obtain a population-level branching continuum of cell states.[53] A previous study using pseudotime inference suggested that molecularly defined subpopulations of retinal progenitors give rise to different sets of cell types.[21] In particular, neurogenic early stage progenitors give rise to ganglion, amacrine, and horizontal cells, Otx2+ late-stage progenitors give rise to bipolar and rod cells, and other late-stage progenitors give rise to Müller glia. However, in our analysis of both the zebrafish and rat retina, we observe progenitors that are biased to form a sister pair of one amacrine and one bipolar cell. In the rat retina, we also observe progenitors that are biased to form a sister pair of one bipolar cell and one Müller glia. Therefore, individual progenitors during development can generate lineage patterns that deviate from the population-level trajectories inferred using pseudotime.[57] In future work, pseudotime trajectories could be refined by using dynamic information learned through lineage motifs.

Looking forward, LMA should be especially useful for contexts that have systematic spatial or cross-species variation in cell type composition. Deeper tree reconstructions could enable the analysis of developmental hyper-motifs, representing higher level correlations between constituent motifs.[58] Analyzing how signaling or transcription factor dynamics are correlated with the generation of motifs will reveal how this process is regulated during development. Overall, by decomposing complex lineage trees into their functional building blocks, lineage motifs should help provide insight into longstanding questions in development and evolution.

Limitations of the study

Because progenitor states are not directly observed in the datasets analyzed here, we cannot make claims about the exact dynamics that regulate how progenitors differentiate over time to give rise to lineage motifs. Incomplete identification of cell types due to the use of limited numbers of markers in the experimental studies analyzed here could prevent discovery of more complex motifs. Finally, the largest limitation has to do with the limited size of existing datasets. Although we have identified motifs and anti-motifs in three different datasets in this work, it is likely that not all fate correlations will recur with high enough significance to be classified as a motif. Programs with weaker fate correlations and datasets of limited size can hinder motif identification, as explored in Figures S5D and S5E. In the small retina datasets analyzed here, rare lineage combinations could be falsely underrepresented (i.e., absent) or overrepresented in the captured data due to limited sampling size. Future work leveraging lineage recording systems should allow the production of much larger spatially resolved lineage tree datasets that could be analyzed with the approaches introduced here.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Animals
  - ○ Primary cell cultures
- METHOD DETAILS
  - ○ Lineage tree resampling and motif identification
  - ○ Construction of synthetic lineage tree datasets
  - ○ Simulation of cell type proportions with input motif matrices
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.devcel.2024.01.017.

ACKNOWLEDGMENTS

We thank Duncan Chadly, Ben Emert, Dongyang Li, Jacob Parres-Gold, Yodai Takei, and other members of the Elowitz lab for scientific input and feedback. We thank Mykel Barrett, Marianne Bronner, Rusty Lansford, Carlos Lois, Magda Zernicka-Goetz, and Meng Zhu for helpful discussion and feedback. This research was supported by Paul G. Allen Frontiers Group and Prime Awarding Agency under award no. UWSC10142 and the National Institute of Health grant R01MH116508. M.T. was supported with funding from the Tianqiao and Chrissy Chen Institute for Neuroscience at Caltech and by the National Eye Institute of the National Institutes of Health under award number F31EY033220. A.A. was supported with funding from National Eye Institute of NIH under award number R00EY031782. M.B.E. is a Howard Hughes Medical Institute investigator. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This article is subject to HHMI's open access to publications policy. HHMI lab heads have previously granted a nonexclusive CC BY 4.0 license to the public and a sublicensable license to HHMI in their research articles. Pursuant to those licenses, the author-accepted manuscript of this article can be made freely available under a CC BY 4.0 license immediately upon publication.

AUTHOR CONTRIBUTIONS

Conceptualization, M.T., A.A., and M.B.E.; methodology, M.T., A.A., and M.B.E.; software, M.T. and A.A.; formal analysis, M.T. and A.A.; writing – original draft, M.T.; writing – review & editing, M.T., A.A., and M.B.E.; visualization, M.T.; supervision, A.A. and M.B.E.; funding acquisition, M.T., A.A., and M.B.E.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 19, 2023
Revised: October 10, 2023
Accepted: January 19, 2024
Published: February 14, 2024

REFERENCES

1. Peichl, L. (2005). Diversity of mammalian photoreceptor properties: adaptations to habitat and lifestyle? Anat. Rec. A Discov. Mol. Cell. Evol. Biol. *287*, 1001–1012.
2. Masland, R.H. (2011). Cell populations of the retina: the Proctor lecture. Invest. Ophthalmol. Vis. Sci. *52*, 4581–4591.
3. Shoval, O., Sheftel, H., Shinar, G., Hart, Y., Ramote, O., Mayo, A., Dekel, E., Kavanagh, K., and Alon, U. (2012). Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. Science *336*, 1157–1160.
4. Kirschner, M.W., and Gerhart, J.C. (2005). The Plausibility of Life: Resolving Darwin's Dilemma (Yale University Press).
5. Stadler, T., Pybus, O.G., and Stumpf, M.P.H. (2021). Phylodynamics for cell biologists. Science *371*, eaah6266.
6. Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev. Biol. *100*, 64–119.
7. Turner, D.L., Snyder, E.Y., and Cepko, C.L. (1990). Lineage-independent determination of cell type in the embryonic mouse retina. Neuron *4*, 833–845.
8. Turner, D.L., and Cepko, C.L. (1987). A common progenitor for neurons and glia persists in rat retina late in development. Nature *328*, 131–136.
9. Hafler, B.P., Surzenko, N., Beier, K.T., Punzo, C., Trimarchi, J.M., Kong, J.H., and Cepko, C.L. (2012). Transcription factor Olig2 defines subpopulations of retinal progenitor cells biased toward specific cell fates. Proc. Natl. Acad. Sci. USA *109*, 7882–7887.
10. Godinho, L., Williams, P.R., Claassen, Y., Provost, E., Leach, S.D., Kamermans, M., and Wong, R.O.L. (2007). Nonapical symmetric divisions underlie horizontal cell layer formation in the developing retina in vivo. Neuron *56*, 597–603.
11. Suzuki, S.C., Bleckert, A., Williams, P.R., Takechi, M., Kawamura, S., and Wong, R.O.L. (2013). Cone photoreceptor types in zebrafish are generated by symmetric terminal divisions of dedicated precursors. Proc. Natl. Acad. Sci. USA *110*, 15109–15114.
12. Rompani, S.B., and Cepko, C.L. (2008). Retinal progenitor cells can produce restricted subsets of horizontal cells. Proc. Natl. Acad. Sci. USA *105*, 192–197.
13. Emerson, M.M., Surzenko, N., Goetz, J.J., Trimarchi, J., and Cepko, C.L. (2013). Otx2 and Onecut1 promote the fates of cone photoreceptors and horizontal cells and repress rod photoreceptors. Dev. Cell *26*, 59–72.
14. Brzezinski, J.A., 4th, Prasov, L., and Glaser, T. (2012). Math5 defines the ganglion cell competence state in a subpopulation of retinal progenitor cells exiting the cell cycle. Dev. Biol. *365*, 395–413.
15. Brzezinski, J.A., 4th, Kim, E.J., Johnson, J.E., and Reh, T.A. (2011). Ascl1 expression defines a subpopulation of lineage-restricted progenitors in the mammalian retina. Development *138*, 3519–3531.
16. De la Huerta, I., Kim, I.J., Voinescu, P.E., and Sanes, J.R. (2012). Direction-selective retinal ganglion cells arise from molecularly specified multipotential progenitors. Proc. Natl. Acad. Sci. USA *109*, 17663–17668.
17. Javed, A., Mattar, P., Lu, S., Kruczek, K., Kloc, M., Gonzalez-Cordero, A., Bremner, R., Ali, R.R., and Cayouette, M. (2020). Pou2f1 and Pou2f2 cooperate to control the timing of cone photoreceptor production in the developing mouse retina. Development *147*, dev188730.
18. Javed, A., Santos-França, P.L., Mattar, P., Cui, A., Kassem, F., and Cayouette, M. (2023). Ikaros family proteins redundantly regulate temporal patterning in the developing mouse retina. Development *150*, dev200436.
19. Liu, S., Liu, X., Li, S., Huang, X., Qian, H., Jin, K., and Xiang, M. (2020). Foxn4 is a temporal identity factor conferring mid/late-early retinal competence and involved in retinal synaptogenesis. Proc. Natl. Acad. Sci. USA *117*, 5016–5027.
20. Zhang, J., Roberts, J.M., Chang, F., Schwakopf, J., and Vetter, M.L. (2023). Jarid2 promotes temporal progression of retinal progenitors via repression of Foxp1. Cell Rep. *42*, 112416.
21. Clark, B.S., Stein-O'Brien, G.L., Shiau, F., Cannon, G.H., Davis-Marcisak, E., Sherman, T., Santiago, C.P., Hoang, T.V., Rajaii, F., James-Esposito, R.E., et al. (2019). Single-Cell RNA-Seq Analysis of Retinal Development Identifies NFI Factors as Regulating Mitotic Exit and Late-Born Cell Specification. Neuron *102*, 1111–1126.e5.
22. Mattar, P., Ericson, J., Blackshaw, S., and Cayouette, M. (2015). A conserved regulatory logic controls temporal identity in mouse neural progenitors. Neuron *85*, 497–504.
23. Eldar, A., and Elowitz, M.B. (2010). Functional roles for noise in genetic circuits. Nature *467*, 167–173.
24. Cepko, C. (2014). Intrinsically different retinal progenitor cells produce specific types of progeny. Nat. Rev. Neurosci. *15*, 615–627.
25. Wolf, S., Wan, Y., and McDole, K. (2021). Current approaches to fate mapping and lineage tracing using image data. Development *148*, dev198994.
26. Raj, B., Wagner, D.E., McKenna, A., Pandey, S., Klein, A.M., Shendure, J., Gagnon, J.A., and Schier, A.F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. Nat. Biotechnol. *36*, 442–450.
27. Chow, K.-H.K., Budde, M.W., Granados, A.A., Cabrera, M., Yoon, S., Cho, S., Huang, T.H., Koulena, N., Frieda, K.L., Cai, L., et al. (2021). Imaging cell lineage with a synthetic digital recording system. Science *372*, eabb3099.
28. Askary, A., Sanchez-Guardado, L., Linton, J.M., Chadly, D.M., Budde, M.W., Cai, L., Lois, C., and Elowitz, M.B. (2020). In situ readout of DNA barcodes and single base edits facilitated by in vitro transcription. Nat. Biotechnol. *38*, 66–75.
29. Chan, M.M., Smith, Z.D., Grosswendt, S., Kretzmer, H., Norman, T.M., Adamson, B., Jost, M., Quinn, J.J., Yang, D., Jones, M.G., et al. (2019). Molecular recording of mammalian embryogenesis. Nature *570*, 77–82.
30. Loveless, T.B., Grotts, J.H., Schechter, M.W., Forouzmand, E., Carlson, C.K., Agahi, B.S., Liang, G., Ficht, M., Liu, B., Xie, X., and Liu, C.C. (2021). Lineage tracing and analog recording in mammalian cells by single-site DNA writing. Nat. Chem. Biol. *17*, 739–747.
31. Yang, D., Jones, M.G., Naranjo, S., Rideout, W.M., 3rd, Min, K.H.J., Ho, R., Wu, W., Replogle, J.M., Page, J.L., Quinn, J.J., et al. (2022). Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. Cell *185*, 1905–1923.e25.
32. Bowling, S., Sritharan, D., Osorio, F.G., Nguyen, M., Cheung, P., Rodriguez-Fraticelli, A., Patel, S., Yuan, W.-C., Fujiwara, Y., Li, B.E., et al. (2020). An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells. Cell *181*, 1693–1694.
33. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. Science *298*, 824–827.
34. D'haeseleer, P. (2006). What are DNA sequence motifs? Nat. Biotechnol. *24*, 423–425.
35. Nevill-Manning, C.G., Wu, T.D., and Brutlag, D.L. (1998). Highly specific protein sequence motifs for genome analysis. Proc. Natl. Acad. Sci. USA *95*, 5865–5871.

36. Granados, A.A., Kanrar, N., and Elowitz, M.B. (2022). Combinatorial expression motifs in signaling pathways. Preprint at bioRxiv. https://doi.org/10.1101/2022.08.21.504714.

37. Geard, N., Bullock, S., Lohaus, R., Azevedo, R.B.R., and Wiles, J. (2011). Developmental motifs reveal complex structure in cell lineages. Complexity *16*, 48–57.

38. Gomes, F.L.A.F., Zhang, G., Carbonell, F., Correa, J.A., Harris, W.A., Simons, B.D., and Cayouette, M. (2011). Reconstruction of rat retinal progenitor cell lineages in vitro reveals a surprising degree of stochasticity in cell fate decisions. Development *138*, 227–235.

39. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. B *57*, 289–300.

40. Benjamini, Y., Krieger, A.M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. Biometrika *93*, 491–507.

41. Hormoz, S., Singer, Z.S., Linton, J.M., Antebi, Y.E., Shraiman, B.I., and Elowitz, M.B. (2016). Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. Cell Syst. *3*, 419–433.e8.

42. Hormoz, S., Desprat, N., and Shraiman, B.I. (2015). Inferring epigenetic dynamics from kin correlations. Proc. Natl. Acad. Sci. USA *112*, E2281–E2289.

43. Wilkinson, H.A., Fitzgerald, K., and Greenwald, I. (1994). Reciprocal changes in expression of the receptor lin-12 and its ligand lag-2 prior to commitment in a C. elegans cell fate decision. Cell *79*, 1187–1198.

44. Sprinzak, D., Lakhanpal, A., Lebon, L., Santat, L.A., Fontes, M.E., Anderson, G.A., Garcia-Ojalvo, J., and Elowitz, M.B. (2010). Cis-interactions between Notch and Delta generate mutually exclusive signalling states. Nature *465*, 86–90.

45. Goodyear, R., Holley, M., and Richardson, G. (1995). Hair and supporting-cell differentiation during the development of the avian inner ear. J. Comp. Neurol. *351*, 81–93.

46. He, J., Zhang, G., Almeida, A.D., Cayouette, M., Simons, B.D., and Harris, W.A. (2012). How variable clones build an invariant retina. Neuron *75*, 786–798.

47. Hu, M., and Easter, S.S. (1999). Retinal neurogenesis: the formation of the initial central patch of postmitotic cells. Dev. Biol. *207*, 309–321.

48. Neumann, C.J., and Nuesslein-Volhard, C. (2000). Patterning of the Zebrafish Retina by a Wave of Sonic Hedgehog Activity. Science *289*, 2137–2139.

49. Morris, S.A., Teo, R.T.Y., Li, H., Robson, P., Glover, D.M., and Zernicka-Goetz, M. (2010). Origin and formation of the first two distinct cell types of the inner cell mass in the mouse embryo. Proc. Natl. Acad. Sci. USA *107*, 6364–6369.

50. Tendler, A., Mayo, A., and Alon, U. (2015). Evolutionary tradeoffs, Pareto optimality and the morphology of ammonite shells. BMC Syst. Biol. *9*, 12.

51. Yamagata, M., Yan, W., and Sanes, J.R. (2021). A cell atlas of the chick retina based on single-cell transcriptomics. Elife *10*, e63907.

52. Domcke, S., and Shendure, J. (2023). A reference cell tree will serve science better than a reference cell atlas. Cell *186*, 1103–1114.

53. Wagner, D.E., and Klein, A.M. (2020). Lineage tracing meets single-cell omics: opportunities and challenges. Nat. Rev. Genet. *21*, 410–427.

54. VanHorn, S., and Morris, S.A. (2021). Next-Generation Lineage Tracing and Fate Mapping to Interrogate Development. Dev. Cell *56*, 7–21.

55. Demb, J.B., and Singer, J.H. (2015). Functional Circuitry of the Retina. Annu. Rev. Vis. Sci. *1*, 263–289.

56. Yu, Y.C., Bultje, R.S., Wang, X., and Shi, S.H. (2009). Specific synapses develop preferentially among sister excitatory neurons in the neocortex. Nature *458*, 501–504.

57. Weinreb, C., Wolock, S., Tusi, B.K., Socolovsky, M., and Klein, A.M. (2018). Fundamental limits on dynamic inference from single-cell snapshots. Proc. Natl. Acad. Sci. USA *115*, E2467–E2476.

58. Adler, M., and Medzhitov, R. (2022). Emergence of dynamic properties in network hypermotifs. Proc. Natl. Acad. Sci. USA *119*, e2204967119.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Rat retina lineage trees	Gomes et al.[38]	https://doi.org/10.1242/dev.059683
Zebrafish retina lineage trees	He et al.[46]	https://doi.org/10.1016/j.neuron.2012.06.033
Mouse blastocyst lineage trees	Morris et al.[49]	https://doi.org/10.1073/pnas.0915063107
Mouse, rabbit, and monkey retina cell type proportions	Masland[2]	https://doi.org/10.1167/iovs.10-7083
Chick retina cell type proportions	Yamagata et al.[51]	https://doi.org/10.7554/eLife.63907
Software and algorithms		
Python	Python Software Foundation	N/A
linmo (Python)	This paper	https://github.com/labowitz/linmo and https://doi.org/10.22002/kn8yx-kmb24
Simulation and analysis code (Python)	This paper	https://labowitz.github.io/linmo/ and https://doi.org/10.22002/kn8yx-kmb24

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Michael B. Elowitz (melowitz@caltech.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- All code used in this study has been deposited at GitHub (https://github.com/labowitz/linmo) as well as the CaltechDATA research repository (https://doi.org/10.22002/kn8yx-kmb24) and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

This study involves the analysis of published datasets on zebrafish retina development,[46] rat retina development,[38] and mouse early embryonic development.[49] Below we provided details on the experimental models drawn directly from each of the published studies.

Animals

He et al.[46] bred the UAS:Kaede and MAZe transgenic zebrafish lines at 26.5°C to obtain embryos, which were grown and imaged from 32 to 72 hours postfertilization at 28.5°C. Phenylthiourea (PTU, 0.0003%) was applied at 8 hours postfertilization to delay pigmentation, and MS-222 (0.04%) was applied prior to live imaging to anaesthetize the embryos. All animal work was approved by the Local Ethical Review Committee at the University of Cambridge and performed under protocols from the UK Home Office license PPL 80/2198. The husbandry and housing of the animals and sex of the embryos was not provided in the published study.

Morris et al.[49] crossed spontaneously ovulating C57BL/6xCBA female mice with CAG::GFP-GPI transgenic male mice and collected eight-cell embryos in M2 medium containing BSA. The embryos were then grown in KSOM medium and imaged until the embryos reached late-blastocyst stage (embryonic day 4.5). The husbandry and housing of the animals, sex of the embryos, culture temperature, institutional permission, and oversight information was not provided in the published study.

Primary cell cultures

Gomes et al.[38] harvested retinal cells from E20 Sprague Dawley rats. The authors filtered the cells twice in a 8 μm nylon mesh and resuspended them in serum-free media with the following components: 1:1 mixture of DMEM-F12 medium with N2 supplement and Neurobasal medium with B27 supplement, penicillin/streptomycin, NT-3 (10 ng/mL), BDNF (10 ng/mL), EGF (50 ng/mL), FGF-2 (10 ng/mL), insulin (20 μg/mL), N-acetyl-L-cystein (6.3 mg/mL), forskolin (25 μM), 8-(4-chlorophenylthio) adenosine 3'5'-cyclic monophosphate (cpt-cAMP, 0.1 mM). The dissociated cells were left at 37°C in a CO_2 incubator for a couple hours to settle, then imaged until all progenitors had terminally differentiated (around 9-14 days total). During imaging, the cells were kept at 37°C with an 8% CO_2, 12% O_2 environment. The sex, authentication of the cells, institutional permission, and oversight information was not provided in the published study.

METHOD DETAILS

Lineage tree resampling and motif identification

NEWICK-formatted lineage trees were first sorted to have doublet and quartet patterns arranged in alphabetical order according to their cell type annotations. All patterns were then aligned in order of earliest to latest born cells. For example, before alignment, triplet patterns could be present in the raw lineage tree data as ((X,X),X) or (X,(X,X)), and were therefore aligned to match the latter format in all cases. A similar procedure was followed for higher-order patterns, like asymmetric quartets, quintets, sextets, and septets.

All cell types and cell type patterns were then enumerated and counted for the number of occurrences within the lineage trees. The datasets were then appropriately resampled according to the type of motifs to be identified. For doublet motif identification, each cell type in the lineage tree dataset was replaced by a random cell type drawing from a list of all cell types within the dataset. This procedure maintains tree topology and overall cell type proportions but eliminates fate correlations between related cells. Our results were not sensitive to replacing with vs. without replacement. To detect larger motifs, it is necessary to control not only for overall cell type frequencies but also for the frequencies of any "sub-patterns" within the pattern of interest. For example, a triplet pattern comprising a sister cell doublet and their common cousin could appear over-represented solely because the sister doublet is itself a motif. To account for this, the lineage trees were resampled in a manner that preserves sub-pattern frequency, by drawing from a pool of similar sub-patterns across all trees. Therefore, for triplet motif identification, each singlet and doublet in the lineage tree dataset was respectively replaced by a random singlet or doublet drawing from a list of all singlets or doublets in the dataset. In this way, the overall frequencies of singlets and doublets remains the same across the resampled dataset while eliminating fate correlations between particular singlets and doublets. For quartet motif identification, each doublet in the lineage tree dataset was replaced by a random doublet drawing from a list of all doublets in the dataset. A similar procedure was followed for increasingly larger patterns.

The occurrences of each pattern were counted for each resampled dataset, then used to calculate an average number of occurrences and standard deviation across all resamples. The average and standard deviation were then used to calculate a z-score as follows:

$$z = \frac{x - \bar{x}}{\sigma}$$

where x is the observed count in the original set of lineage trees, $\bar{x}$ is the average count across all resamples, and σ is the standard deviation across all resamples.

For plotting, the expected count of each pattern was calculated by multiplying the marginal probabilities of observing each of the two sub-patterns by the total number of that type of pattern across the entire dataset. Additionally, if the sub-patterns were not identical, the expected number was multiplied by two. For example, the expected number of the triplet (A,(B,C)) would be calculated as P(A) * P((B,C)) * total number of triplets * 2, and the expected number of the quartet ((A,B),(A,B)) would be calculated as P((A,B)) * P((A,B)) * total number of quartets. The null z-scores were calculated by repeating the same resampling procedure above for randomly chosen resampled datasets.

Construction of synthetic lineage tree datasets

To generate the example lineage trees shown in Figure 2, lineage trees were simulated using a one-step model in which a progenitor can self-renew or differentiate using probabilities of 10 or 90% respectively. For all differentiated cells, one of three possible fates were assigned with equal probability: blue, green, or a committed progenitor. Finally, two cell divisions were simulated for all committed progenitors, such that each gave rise to a triplet of (green, (green, blue)) with 100% probability. Cell divisions and fate differentiation were repeatedly simulated for all progenitors present within the tree until all cells reached terminal fates.

To test the recursive nature and accuracy of LMA in Figures 3 and S1, synthetic lineage tree datasets were simulated using a competence progression model or binary fate model of development. Each tree started as an 'a' or 'i' progenitor for each respective model, and a cell division was simulated producing two descendant cells, whose fates were chosen probabilistically based on the transition probabilities of the parental progenitor type. Cell divisions and fate differentiation were repeatedly simulated for all progenitors present within the tree until all cells reached terminal fates (A-F or A-H for each respective model).

To test LMA on lineage trees generated using an extrinsic model of development in Figure S3, synthetic lineage tree datasets were generated by populating a grid with 61 x 61 progenitor cells. A cell division was simulated for a randomly chosen progenitor, producing two descendant cells. One of the descendent cells remains at the same position in the grid, while the other cell is randomly placed in the immediately adjacent up, down, left, or right position, shifting all other cells within the column or row in an outwards manner. Fates were then chosen for each of the descendent cells probabilistically based on the transition probabilities as defined in Figure S3B. Cell divisions and fate differentiation were repeatedly simulated for all progenitors present within the grid in a similar manner until all cells reached terminal fates. For the extrinsic model, P(A)* was calculated as (40 - 10n_A) * SF, where n_A is the number of neighboring A cells with 4-connectivity, SF is 80 / ((40 - 10n_A) + (40 - 10n_B)), and n_B is the number of neighboring B cells with 4-connectivity.

To test various dataset characteristics on motif identification, synthetic lineage tree datasets were generated by simulating self-renewal and differentiation in a progenitor cell, using probabilities as defined in Figure S5A. To account for rod sister fate correlations, each doublet of terminally differentiated cells was first assigned one fate based on the standard set of fate probabilities. If the assigned fate was indeed a rod cell, then the fate probability of the remaining cell within the terminal doublet was modified to be the conditional probability being tested. Cell divisions and fate differentiation were repeatedly simulated for all progenitors present within the tree until all cells reached terminal fates. This process was repeated in a similar manner to incorporate Müller glia sister fate correlations.

Simulation of cell type proportions with input motif matrices

Cell type proportions were first simulated using input motif matrices (Figure 7B) by choosing random frequencies for each motif and taking sets of cell types that were of total size 100 cells (for X_0 and X_1) or 99 cells (for X_2). For the motif transformation using the motifs measured in the rat retina dataset, e_3 was computed as $z_3 - X_3 y_3$ where z_3 is the total number of A, B, and M cell types in the dataset, X_3 is the empirical motif matrix based on the observed motifs, [((B,(A,B)), (A,A), (A,B), and (B,M)], and y_3 is the number of occurrences of each motif within the dataset (Figure S7A). Similarly, e_4 was computed as $z_4 - X_4 y_4$ where z_4 is the total number of A, B, and M cell types in the dataset, X_4 is the empirical motif matrix based on the observed motifs, [((B,(A,B)), (A,B), and (B,M)], and y_4 is the number of occurrences of each motif within the dataset (Figure S7B). The total number of 113 A, B, and M cells across the entire dataset was used for $\sum_i (z_3)_i$ and $\sum_i (z_4)_i$. R cells were omitted from this analysis because no rat retinal motifs contained R cells.

QUANTIFICATION AND STATISTICAL ANALYSIS

For resampling lineage trees, 10^4 datasets were used to generate counts for each pattern across the resamples that resemble a normal distribution. For the motif transformation, 10^5 datasets were generated to sample the possible space of cell type proportions. The p-value for all patterns in the paper was calculated by (1) determining whether the observed count is higher or lower than the average across the resamples (the null distribution), (2) counting the number of resamples that have counts at least as extreme as the observed counts, and (3) dividing this by the total number of resamples to obtain a one-sided p-value. P-values were adjusted to the total number of patterns analyzed using the Benjamini/Hochberg correction with false discovery rate (α) = 0.05 and a two-stage linear step-up procedure with estimation of the number of true hypotheses.[39,40]

ADDITIONAL RESOURCES

GitHub repository: https://github.com/labowitz/linmo

linmo package documentation: https://labowitz.github.io/linmo/