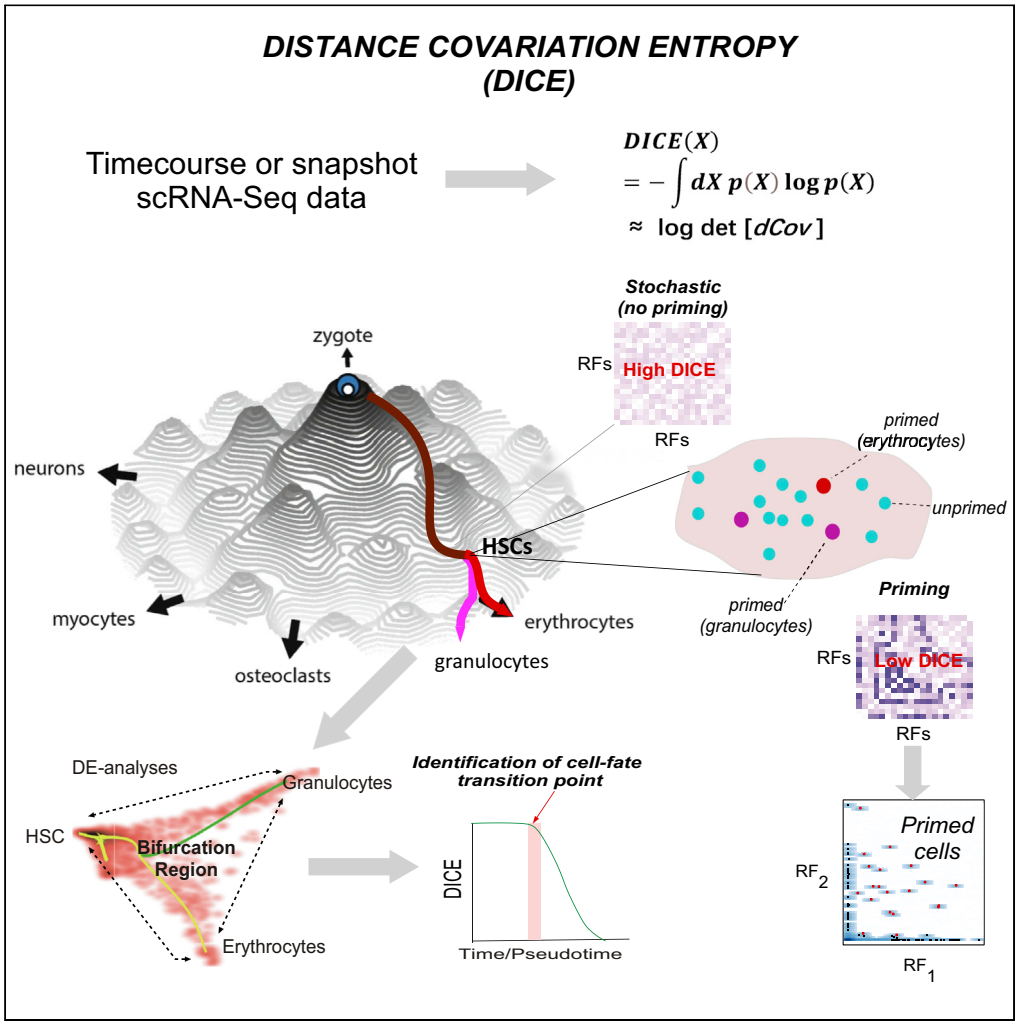
iScience

CellPress
OPEN ACCESS

Article

Distance covariance entropy reveals primed states and bifurcation dynamics in single-cell RNA-Seq data



Qi Luo, Alok K. Maity, Andrew E. Teschendorff

andrew@sinh.ac.cn

Highlights

Covariation entropy of regulatory factors can reveal multipotent primed states

Detecting multipotent primed states from scRNA-Seq data only is possible

Distance covariation entropy improves the identification of tipping points

Primed states co-express an epigenetic regulator and a cell-lineage specific factor

Luo et al., iScience 25, 105709
December 22, 2022
https://doi.org/10.1016/j.isci.2022.105709

iScience

CellPress
OPEN ACCESS

Article

Distance covariance entropy reveals primed states and bifurcation dynamics in single-cell RNA-Seq data

Qi Luo,[1,2] Alok K. Maity,[1,2] and Andrew E. Teschendorff[1,3,*]

SUMMARY

Cell-fate transitions are fundamental to development and differentiation. Studying them with single-cell omic data is important to advance our understanding of the cell-fate commitment process, yet this remains challenging. Here we present a computational method called DICE, which analyzes the entropy of expression covariation patterns and which is applicable to static and dynamically changing cell populations. Using only single-cell RNA-Seq data, DICE is able to predict multipotent primed states and their regulatory factors, which we subsequently validate with single-cell epigenomic data. DICE reveals that primed states are often defined by epigenetic regulators or pioneer factors alongside lineage-specific transcription factors. In developmental time course single-cell RNA-Seq datasets, DICE can pinpoint the timing of bifurcations more precisely than lineage-trajectory inference algorithms or competing variance-based methods. In summary, by studying the dynamic changes of expression covariation entropy, DICE can help elucidate primed states and bifurcation dynamics without the need for single-cell epigenomic data.

INTRODUCTION

Cell-fate transitions are fundamental to development and homeostasis, and are orchestrated by regulatory factors that include transcription factors (TFs) and chromatin-state modifiers.[1,2] A relatively large number of these cell-fate transitions are driven by sets of antagonistic TFs that exhibit highly specific differentiation activity in one particular lineage, while being switched off in the alternative lineages,[1–3] as epitomized by the GATA1-PU1 antagonism that controls the erythroid-myeloid fate transition in early hematopoiesis.[4] In general, abrupt changes in the expression ratio of TFs has been proposed as a means of identifying the TFs controlling alternative cell-fates,[3] including the identification of primed states in multi-or-pluripotent cell populations,[5,6] defined as multi/pluri-potent cellular states that are already committed to differentiate into specific downstream lineages. Studying priming is critical not only to improve our understanding of the cell-fate commitment process, but also for potential regenerative medicine purposes.[5]

Theoretically, cell-fate transitions and switch-like behavior has been modeled in terms of gene-regulatory networks (GRNs) and an associated system of ordinary differential equations (ODEs) that describe the dynamic changes in TF concentrations.[4,7–9] In most cases however, the full underlying GRNs are unknown, and the associated ODEs typically include many unknown parameters whose precise values may strongly affect the resulting dynamics, rendering this ODE-modeling approach impractical. In the face of these difficulties, single-cell RNA-Seq (scRNA-Seq) data[10] and other single-cell omic data types[11] offer the unprecedented opportunity to study the underlying bifurcation dynamics from a more data-driven perspective, as exemplified by the development of numerous lineage-trajectory inference algorithms[12,13] and statistical methods to detect TFs controlling the cell-fate commitment process.[5,14–16] In principle, with scRNA-Seq data it may be possible to identify the full repertoire of relevant regulatory factors,[14] to more accurately pinpoint the bifurcation or tipping points underlying cell-fate transitions,[15–21] or to better characterize system-wide properties such as priming in pluri-or-multipotent cell populations.[5,6] However, challenges remain. Accurately identifying cell-fate transition points from scRNA-Seq data is difficult because sampling of cells from the transition region is often very sparse, the implication being that lineage-trajectory inference algorithms can't accurately pinpoint the bifurcation point. It also remains unclear how best to identify the TFs controlling cell-fate transitions, partly because of the scRNA-Seq assay's low sensitivity to detect

[1]CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China

[2]These authors have contributed equally

[3]Lead contact

*Correspondence: andrew@sinh.ac.cn

https://doi.org/10.1016/j.isci.2022.105709

dynamic changes in TF expression.[22–24] Finally, there are conflicting opinions regarding a key phenomenon such as priming in multipotent cell populations, with some studies reporting primed states that are detectable from scRNA-Seq data alone,[25] with others advocating the need to consider epigenomic (e.g. scATAC-Seq) data.[26,27]

Here we address these challenges by proposing a computational framework based on the concept of distance covariance entropy (DICE). DICE identifies candidate regulatory factors (RFs) controlling primed states and cell-fate transitions by analyzing the entropy of their expression covariation patterns, either in snapshot populations or as a function of time/pseudotime. In effect, DICE quantifies the randomness of the expression covariation between RFs, and is thus ideally suited to explore single-cell phenomena such as priming. Indeed, using DICE we demonstrate that priming in multipotent cell populations is an ubiquitous phenomenon that is detectable from scRNA-Seq data only, without the need for matched scATAC-Seq profiles. In doing so, we discover that primed states are often defined by epigenetic regulators or pioneer factors, consistent with these factors altering the chromatin accessibility landscape that precedes and enables priming and fate commitment. Importantly, the identification and validation of primed states hinges entirely on the use of metrics that quantify the covariation in expression, as indeed simpler variance-only based metrics fail to discover these states.

RESULTS

Rationale of the DICE algorithm

The identification of primed states, bifurcation points, and TFs that control cell-fate decisions from scRNA-Seq data is complicated by the sparseness of sampling (i.e. relatively low coverage of cells) around cell-fate transition points, and the sparseness of the count-data per cell, which can hamper the reliable identification of TFs controlling such fate transitions.[14,22,23] Recent work has aimed to elucidate bifurcation dynamics by studying the variation in expression of TFs, based on the principle that variation of relevant TFs increases in the vicinity of cell-fate transition points.[14,28] Here, we propose an extension of this concept based on how the covariation of TF regulatory activity patterns change with time/pseudotime. Briefly, given a scRNA-Seq dataset, with cells collected either in real time or binned according to pseudotime, DICE first identifies candidate regulatory factors (RFs) that display interesting covariation patterns in time/pseudotime (Figure 1A, STAR Methods). We consider two classes of RFs: TFs and epigenetic factors (EFs), i.e., chromatin regulators/modifiers, because these are known to also control the epigenetic process of differentiation. Although prior selection of RFs is possible if appropriate biological knowledge is available, in general, candidate RFs are identified by the requirement that they display increased differentiation activity with time/pseudotime, but only along one differentiation path corresponding to the specification of one particular cell-fate (Figure 1A, STAR Methods). To measure differentiation activity of RFs we rely on their measured mRNA expression levels. However, for TFs, we preferably use a corresponding TF-regulon designed to measure differentiation activity, assuming such regulons are available.[22] To clarify, in the regulon approach we infer regulatory (differentiation) activity of the TF from the expression levels of its regulon target genes, a strategy that improves the sensitivity and precision to detect true TF differentiation activity changes[22] (STAR Methods). Of note, candidate RFs are further required to display clear antagonistic switch-like behavior between opposing lineages and across the whole time course, a strategy that is well justified on principles from dynamical systems theory (STAR Methods).[6,17] In addition, this strategy allows more robust identification of key RFs, as their inference is drawn from using all timepoints and not just from the lower number of cells collected at or near the bifurcation event.

Having identified candidate RFs, we next compute a covariation metric of these RFs across all cells from a given timepoint or pseudotime bin and for all available timepoints/bins (Figure 1B). To account for potential non-linear or non-monotonic dependencies between RFs, we use the notion of distance covariation/correlation,[29,30]which can elegantly encapsulate such complex dependencies (Figure S1, STAR Methods) in a way that allows subsequent global quantification of these dependencies using the covariation entropy,[6,31] leading to the notion of distance covariance entropy (DICE) (Figure 1B). One potential application of DICE is to explore its variation as a function of timepoint/pseudotime, to better pinpoint the timing of cell-fate transitions, which may not be well characterized from ordinary diffusion maps (Figure 1C). Another potential application of DICE is to the detection of primed states in say multipotent cell populations, as well as the RFs that control such priming (Figure 1D). Of note, although the acronym DICE stands for DIstance Covariance Entropy, we henceforth and interchangeably use the term DICE to also refer to the whole algorithmic framework described above.

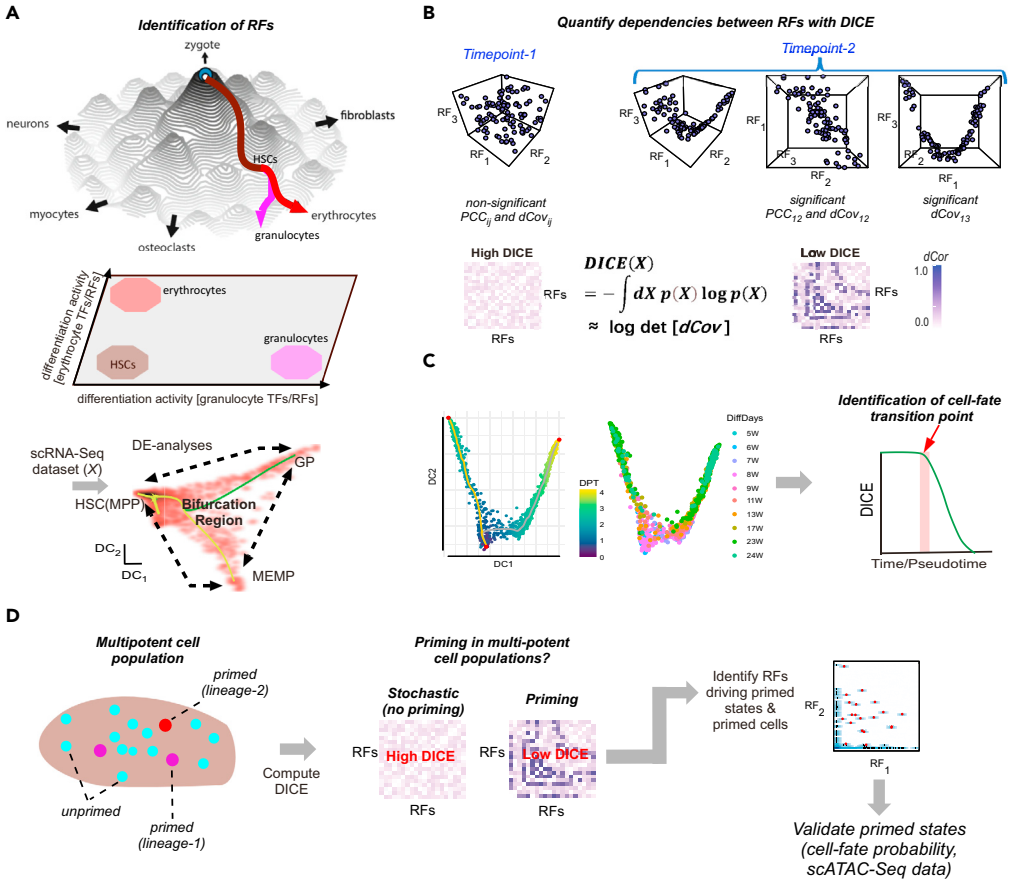


Figure 1. The DICE metric, algorithm and applications

(A) Given a differentiation process into fully differentiated cell-types (depicted here as a path in Waddington's landscape), and an associated scRNA-Seq dataset *X*, the first aim is to identify pools of regulatory factors (RFs) that may contribute to this process. This involves identification of terminal branch points in the context of diffusion maps, as well as differential expression (DE) analyses in order to identify RFs that become more active in the mature cell-types compared to (1) their multipotent progenitors and (2) the mature cell-types from competing lineages, as indicated.

(B) The next step is to quantify the dependencies of these RFs as a function of timepoint or pseudotime bin. Scatterplots of RF activity/expression for 3 RFs intended to display the cases of (1) no association (non-significant Pearson Correlation Coefficients-PCCs or distance covariance-dCov), corresponding to high distance covariance entropy (DICE), and (2) the case of a significant association (as measured by dCov). Illustrated are examples where the relation between RFs is linear and non-monotonic non-linear, the latter requiring the dCov metric to find such associations.

(C) One application of DICE is to compute the DICE metric for each timepoint/pseudotime bin and study its variation as a function of time/pseudotime, to better pinpoint the cell-fate transition point. The latter may or may not be easily identifiable from an ordinary diffusion map, as indicated.

(D) Another application of DICE is to identify primed states in say multipotent cell populations using only scRNA-Seq data. After identification of candidate RFs (as described in a)), DICE can be computed to determine if it is significantly lower compared to that of a fully stochastic randomized data matrix. If DICE is significantly lower, this indicates the existence of associations between RFs and correspondingly of "primed cells" that overexpress these particular RFs. In this work, we validate the RFs and cells involved defining these primed states using cell-fate probability calculations as well as matched scATAC-Seq data.

DICE captures general cell-fate transitions and outperforms the ordinary covariance entropy

First, we aimed to demonstrate that DICE can capture complex cell-fate transitions and that it is preferable over the ordinary covariance entropy measure[31] which cannot account for non-linear patterns. We initially studied this in the context of simulated data, by considering a GRN describing the canonical cross-antagonism of 2 TF (e.g. GATA1/PU1) repressing each other whilst also driving their own positive feedback loop (Figure 2A). This GRN has 9 parameters, but we considered the symmetric scenario, where there are in effect only five (a, b, k, θ, n), describing the strength of autoactivation (a), mutual repression (b) and

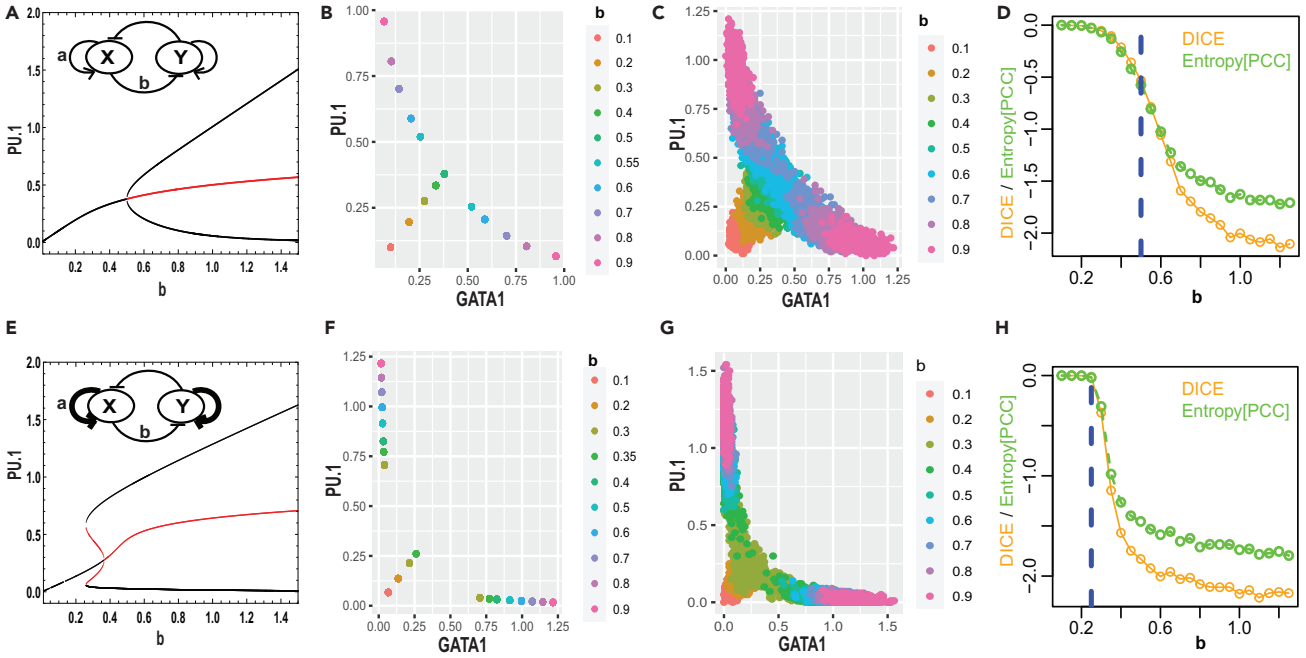


Figure 2. DICE analysis in the GATA1-PU.1 system

(A) Supercritical bifurcation diagram displaying the activation level of PU.1 against the bifurcation parameter. The black and red lines in the bifurcation figures indicate stable and unstable states, respectively. Inset figure shows the mutually repressing with auto-activation GRN representing the GATA1-PU.1 system. Arrows represent auto-activation (rate constant, A) and short bars represent mutual repression (rate constant, (B).

(B) Scatterplot displaying the activation level of the TFs (PU.1 and GATA1) for different bifurcation parameter values. Each dot represents an individual cell.

(C) As (B), but now displaying the values for many cells per bifurcation value.

(D) Plot of DICE (orange) versus bifurcation parameter, with the entropy of the ordinary covariance matrix (PCC) shown in green.

(E-H) As (A-D), but for the scenario of a subcritical pitchfork bifurcation.

degradation (k), with (θ, n) representing additional parameters controlling the autoactivation and mutual repression functions (STAR Methods). For particular parameter choices, e.g. (a, k, θ, n)=(0.008, 1, 0.5, 4), the system is known to undergo a supercritical pitchfork bifurcation as b (the bifurcation parameter) varies from 0.05 to 1.25 (Figure 2A). In this case, the expression state of any one of the two TFs is monostable up until a value 0.5, with larger b-values leading to a bistable regime corresponding to the two possible fates (Figure 2A). By simulating the expression dynamics of the 2 TF for 1000 cells and different choices of b, we verified this bifurcation pattern (Figures 2B and 2C). Next, we computed DICE across the 2 TF and over the 1000 cells for each choice of bifurcation parameter b, which revealed a sharp decrease at the known bifurcation value (Figure 2D), reflecting the anti-correlative expression pattern of the 2 antagonistic TFs. Importantly, DICE displayed a more pronounced decrease compared to the ordinary covariance entropy (Figure 2D, STAR Methods), thus demonstrating that even in the simplest of GRNs, TFs may display weak but significant non-linear dependencies (Figures 2B and 2C) that are better captured by the distance covariation metric. Of note, similar findings were observed if the GRN model undergoes a subcritical bifurcation (Figures 2E–2H). To confirm that DICE can capture bifurcation dynamics in other more complex GRNs, we generated TF expression data for (1) a 4 TF GRN describing the transition from multipotent lymphoid progenitor cells to unipotent early T cells[32] and (2) a 52-gene GRN describing the transformation of somatic cells to induced pluripotent stem cells[33,34] (STAR Methods). We note that for these two systems, the patterns of dependency between TFs are more complex. Despite this, DICE exhibited abrupt patterns of change at the known transition points which were also more pronounced than those derived using the ordinary covariance entropy (Figures S2 and S3). Thus, all these results validate and justify the use of DICE to capture complex cell-fate transitions in different GRNs. This is an important requirement to justify application to general scRNA-Seq datasets where the underlying GRNs may be unknown or incomplete.

DICE identifies primed states in fetal liver hematopoiesis

Next, we applied our DICE algorithm to real scRNA-Seq data, specifically to a scRNA-Seq dataset representing fetal liver hematopoiesis from Ranzoni et al.,[27] to explore if DICE can detect any evidence of priming in

hematopoietic multipotent progenitor (MPP) cells. We here adopt the same definition of priming as in Ranzoni et al., i.e. as the presence of non-random positive associations (possibly non-linear/non-monotonic) between common lineage-specifying regulatory factors in a pool of MPP-cells. This non-random positive association would be driven by a proportion of cells primed to the given cell-fate. Whether such priming is present in multi-and-pluripotent cell populations generally is a question of paramount interest given previous reports that such priming may not be detectable from scRNA-Seq data alone, requiring in addition chromatin-state data (e.g., scATAC-Seq).[26,27] To ensure cross-comparability with Ranzoni et al., we used the normalized data and cell-type annotation provided by the authors. Out of 4504 cells that passed QC, we focused on 1569 non-cycling cells annotated as MPP (n = 1200), megakaryocyte-erythroid-mast progenitors (MEMPs, n = 192) and granulocyte progenitors (GPs, n = 177). More mature cell-types (e.g., B-cells, granulocytes, erythroid cells, dendritic cells) were excluded from our analyses, as this would add substantial data variation that could obscure the inference of the MEMP-GP bifurcation. Using Diffusion Maps[35] on the full normalized data matrix defined over the 1569 non-cycling cells, confirmed the existence of a clear bifurcation from MPPs into either GPs or MEMPs (Figure 3A). To identify candidate regulatory factors controlling the GP fate decision, we performed differential overexpression analysis comparing GPs to MPPs and separately also to MEMPs. An analogous overexpression analysis was performed to identify candidate RFs regulating the MEMP fate decision. These DE-analyses were done using a combined list of 1994 RFs from the Molecular Signature and DOROTHEA databases,[23,36] which not only includes TFs but also EFs. In all, we identified 46 GP- and 79 MEMP-specific RFs (Figure 3B). The list of 46 GP RFs included well-known granulocyte-specific TFs such as IRF8 as well as epigenetic factors such as UHRF1. The list of 79 MEMP RFs included well-known MEMP-specific TFs such as KLF1, GATA1 and TAL1, but also epigenetic factors like HDAC1. Of interest, we observed that GP and MEMP-specific RFs fell into two broad categories, that we termed type-1 and type-2 (Figure 3B), depending on their frequency of expression in MPP cells. For instance, KLF1 and GATA1 displayed low frequency of expression among MPP cells, and displayed clear antagonistic patterns between MEMPs and GPs. On the other hand, a MEMP-specific factor like FOXP1, which also exhibited significantly higher expression in MEMP-cells compared to either MPP or GP-cells, still displayed frequent expression in the MPP and GP-cells. Because the frequency of primed cells among MPPs is expected to be low, we reasoned that the RFs responsible for priming would fall into the former "type-1" category. In support of this, type-1 RFs displayed stronger overexpression in the mature differentiated cells from the corresponding lineage when compared to the mature differentiated ones from the opposite lineage, while also displaying stronger enrichment for biological terms related to lineage differentiation (Figure S4, STAR Methods). Computing DICE for the 26 type-1 GP-RFs over the MPP cells and separately also for the 45 type-1 MEMP-RFs, demonstrated clear evidence of priming, i.e. their expression covariation patterns were distinctively non-random (Figure 3C). We stress that this is a non-trivial finding because the selection of the RFs never depends on such covariation patterns across MPP cells.

Next, to identify the specific RFs driving the primed states into GP and MEMP lineages, we devised a dual perturbation entropy and Spearman rank correlation strategy, designed to confidently identify RF pairs exhibiting the most significant positive covariation as assessed over the MPP population (STAR Methods). In the case of the GP and MEMP lineages, this identified IRF8-UHRF1 and GATA1-KLF1 as the most highly ranked pairs, respectively (Figures 3D and 3E). Thus, we posited that these RF-pairs control early priming and cell-fate commitment into the GP and MEMP lineages, respectively.

To validate this prediction in the context of the same scRNA-Seq data, we next identified the GP-primed cells within the MPP population as those cells co-expressing IRF8 and UHRF1, and similarly for the MEMP-lineage using GATA1 and KLF1 (STAR Methods). In total, we identified 88 and 24 MPP cells primed for the GP and MEMP lineages, respectively, and their positions in the Diffusion Map were consistent with that of their respective differentiation paths (Figure 3F). To confirm this, and thus to validate the RF-pairs defining the primed states, we used the Palantir algorithm[37] to compute fate probabilities for each of the MPP cells (STAR Methods, Figure 3G). Validating our assignments, we observed that MPP cells primed for the GP lineage exhibited significantly higher probabilities of differentiating into GP-lineage cells compared to unprimed cells or cells primed for the MEMP-fate (Figure 3H). Conversely, cells primed for the MEMP lineage exhibited lower probabilities of differentiating into the GP-fate as compared to unprimed and GP-primed cells (Figure 3H).

Validation of primed states using scATAC-Seq data

To further validate the above findings, we posited that the TFs defining the primed states would exhibit higher TF-activity scores in the primed cells as measured using an independent assay such as scATAC-Seq. Although

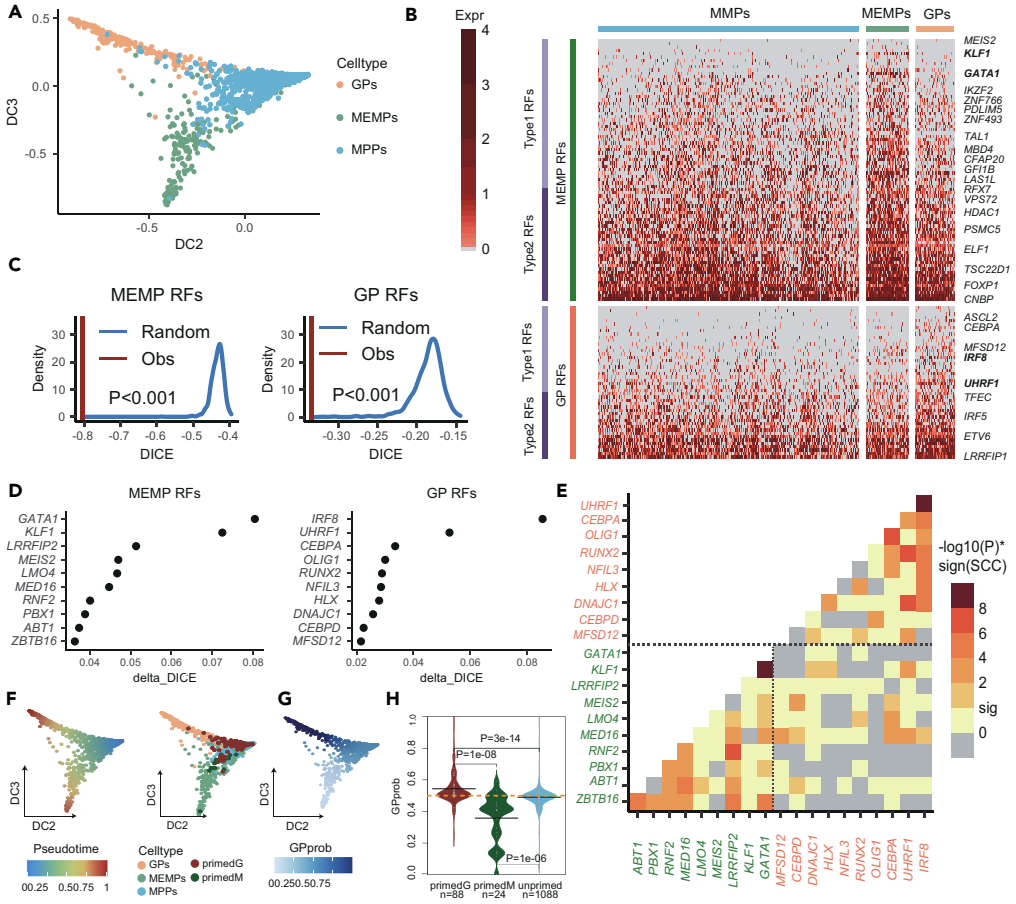


Figure 3. DICE identifies priming and primed states in fetal liver hematopoiesis

(A) Diffusion map derived from the scRNA-seq dataset encompassing multipotent progenitors (MPPs), megakaryocyte-erythroid-mast progenitors (MEMPs) and granulocyte progenitors (GPs) during fetal liver hematopoiesis, demonstrating the bifurcation from MPPs into GPs and MEMPs.

(B) Heatmap displaying the mRNA expression levels of selected lineage specific regulatory factors (RF) across the different cell types. Every RF displayed in this heatmap exhibits significantly higher expression in one of the two lineages (GPs or MEMPs) relative to both the MPPs as well as the cells from the other lineage. Type-1 RFs refer to RFs with relatively low expression frequency in MPPs, whereas type-2 RFs refer to those with relatively high expression frequency in MPPs.

(C) Vertical red line denotes the observed DICE value for the type-1 lineage specific RFs as estimated over the MPPs. The blue curve represents the null distribution obtained by randomizing the data matrix (1000 Monte-Carlo randomizations). Empirically derived P-values are given.

(D) Panels depict the change in DICE values obtained on removing each type-1 RF in turn, separately for MEMP and GP TFs. We only depict the top 10 type-1 RFs exhibiting the highest increased DICE values.

(E) Spearman correlation coefficient (SCC) heatmap between the top-10 RFs from (D). The heatmap is colored by-log10-adjusted P-value of the SCC weighted by the sign of the SCC. MEMP specific RFs are displayed in green and GP specific RFs in coral.

(F) As (A), but with cells colored by pseudotime (left) and cell-type including primed cells (right). PrimedM refers to MPPs co-expressing GATA1 and KLF1, and primedG to MPPs co-expressing IRF8 and UHRF1.

(G) As (A), but with cells colored by their cell-fate probabilities, specifically the probability of differentiating into GPs, as calculated with Palantir.

(H) Violin plots comparing the probabilities of differentiating into GP-lineage for different MPP cell groups: PrimedG are MPPs co-expressing IRF8 and UHRF1 and primed for GP-lineage, primedM are MPPs co-expressing GATA1 and KLF1 and primed to MEMP-lineage, unprimed are the rest of MPPs. p values are from a one-sided Wilcoxon rank-sum test. The horizontal dashed orange line represents the GP-fate probability = 0.5. The three horizontal black lines represent the mean values for the different cell groups. Gray points in the violins represent real data points.

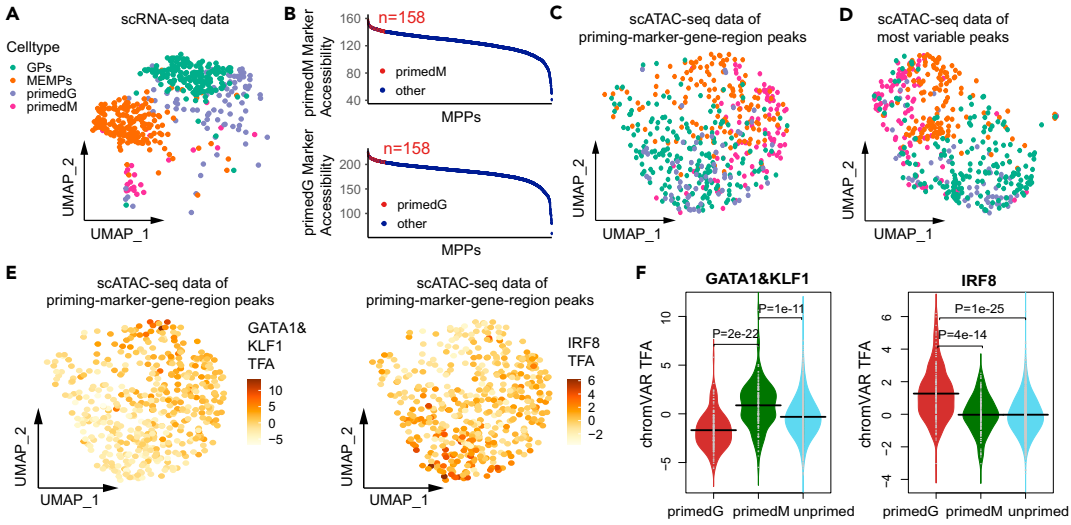


Figure4. Validation of primed states using scATAC-Seq data

(A) UMAP visualization of MEMPs, GPs and the primed cells, as obtained from the scRNA-seq using the derived marker genes.

(B) Scatterplots of scATAC-Seq data accessibility scores for each MPP cell, with cells ranked in decreasing order of overall accessibility as computed over the corresponding marker genes. The top 10% of cells were defined as primed in the corresponding lineage, as shown. Cells ranked among the top 10% in both lineages were not assigned as primed.

(C) UMAP visualization of MEMPs, GPs and the primed MPP cells (as predicted from (B)), as obtained from the scATAC-seq using normalized peak-level data summarized over the marker genes.

(D) As (C), but as obtained from the scATAC-Seq data using the most variable peak-level features.

(E) As (C), but with cells now colored by the TF activity (TFA) obtained by running chromVAR on the scATAC-seq data. Left panel: colored by the average TFA of GATA1 and KLF1. Right panel: colored by the TFA of IRF8.

(F) Violin plots comparing the average chromVAR TFA of GATA1 and KLF1 (left panel) and IRF8 (right panel) between the predicted MEMP and GP primed cells (primedM, primedGP) and unprimed cells. P-values derived from one-tailed Wilcoxon rank sum tests.

scATAC-Seq profiles for the same cells analyzed earlier are not available, they are available for the same cell-types, as profiled by the same study.[27] Thus, using the scATAC-Seq data and cell-type annotations provided by Ranzoni et al.,[27] we aimed to identify putative primed cells within the MPP population (STAR Methods). Briefly, this was done by performing differential overexpression analysis to first infer marker genes for the primed-GP and primed-MEMP MPP subpopulations identified earlier with the scRNA-Seq data (Figure 4A), and subsequently computing scATAC-Seq accessibility scores over these marker genes for each of the MPP cells (STAR Methods). This allowed identification of candidate GP and MEMP primed cells as those exhibiting the corresponding highest accessibility scores (Figure 4B). To check that these assignments of candidate primed cells are plausible, we performed UMAP analysis[38] on all cells using normalized peak-level counts over the marker genes (STAR Methods), displaying only the GP, MEMP and corresponding predicted primed cells. This revealed a significant segregation of GP and MEMP cells, with the primed cells of a given lineage displaying preferential co-clustering with the more mature cells of that lineage (Figure 4C). Importantly, consistent segregation of the primed cells of each lineage was also evident when performing UMAP on the most variable peak-level features, without restricting to our marker genes, thus confirming the consistency of the primed states (Figure 4D). To further validate these states, we posited that if the TFs we have identified from the scRNA-Seq data do indeed define GP and MEMP primed states, that the target genes of these specific TFs ought to exhibit higher chromatin accessibility in the predicted primed cells. To this end, we applied chromVar[39] to infer a TF regulatory activity score (TFA) for each of the DNA-binding TFs IRF8, GATA1 and KLF1 in each of the MPP cells. This revealed that TFA levels of GATA1 and KLF1 were higher in the cellular neighborhoods enriched for MEMP-primed cells, while the converse was true for TFA of IRF8 (Figure 4E). Using Wilcoxon rank sum tests we confirmed the statistical significance of these patterns (Figure 4F). Overall, the scATAC-Seq data support the view that GATA1 and KLF1 define MPP states primed to differentiate into the MEMP-lineage, while IRF8 defines MPP states primed for the GP-lineage. Of note, although we used the scATAC-Seq data to validate the primed states, the identification of these states was accomplished using only scRNA-Seq data.

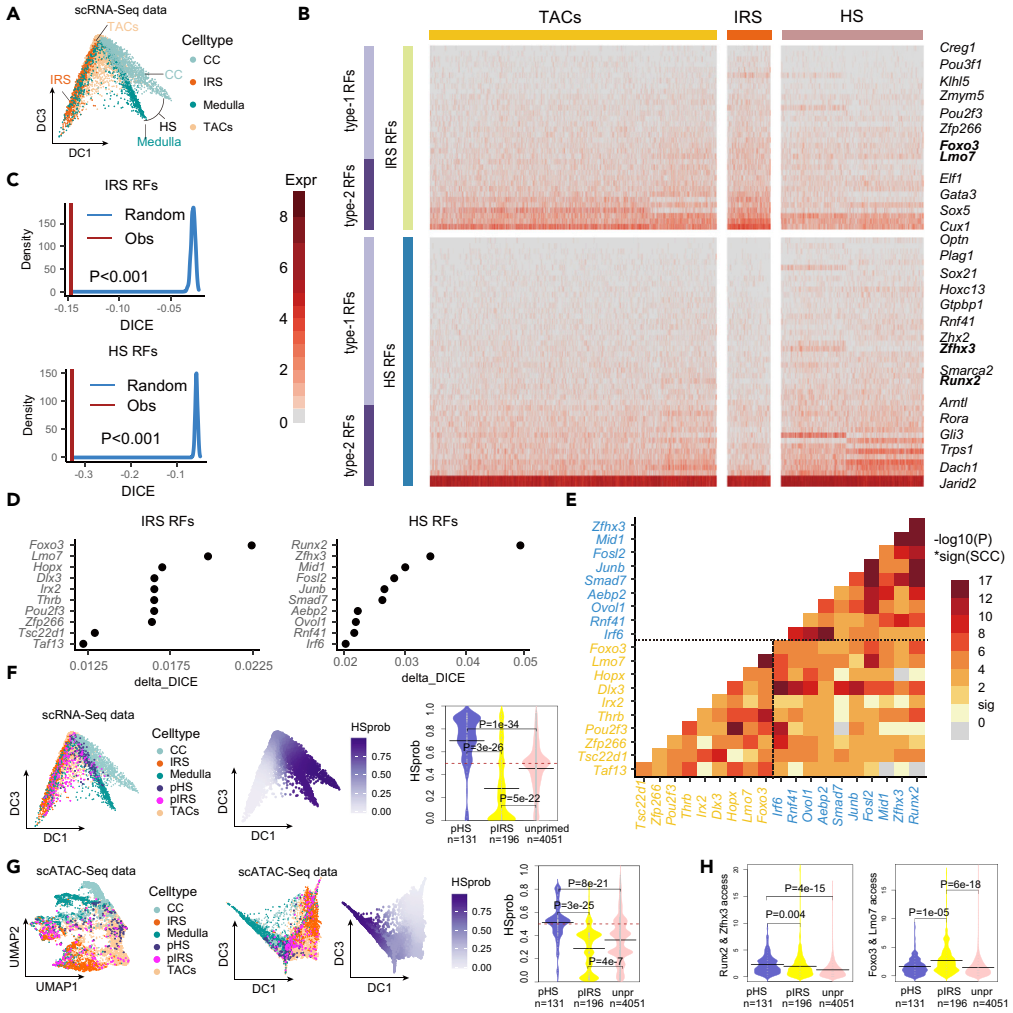


Figure5. Identification and validation of primed states in hair-follicle regeneration

(A) Diffusion map derived from the scRNA-seq dataset encompassing multipotent transit amplifying cells (TACs), differentiated hair-shaft (HS) cells encompassing medulla and cuticle cortex (CC), and differentiated IRS cells during hair-follicle regeneration.

(B) Heatmap displaying the mRNA expression levels of IRS and HS-specific regulatory factors (RF) across the different cell types. Every RF displayed in the heatmap exhibits significantly higher expression in one of the two lineages (IRS or HS) relative to TACs and the opposite lineage. Type-1 RFs refer to RFs with relatively low expression frequency in TACs, whereas type-2 RFs refer to those with relatively higher expression frequency.

(C) Vertical red line denotes the observed DICE value for the type-1 lineage specific RFs as estimated over the TACs. The blue curve represents the null distribution obtained by randomizing the data matrix (1000 Monte-Carlo randomizations). P-values are given.

(D) The change in DICE values obtained on removing each of the type-1 RFs in turn, separately for each lineage. We only depict the top-10 type-1 RFs exhibiting the highest increased DICE values.

(E) Spearman correlation coefficient (SCC) heatmap between the top-10 RFs from each lineage. The heatmap is colored by log10-adjusted P-value of the SCC weighted by the sign of the SCC. HS-specific RFs are displayed in blue and IRS-specific RFs in orange.

(F) Left panel: as (A) but with primed TAC cells annotated as primed into IRS or HS lineages. Middle panel: as (A) but will cells labeled by their Palantir probability of differentiating into HS-lineage. Right panel: Violin plots depicting the Palantir-derived probability (averaged over 50 runs) of differentiating into HS for all TACs, stratified according to being primed to HS (pHS), primed to IRS (pIRS) or unprimed. P-values derive from one-tailed Wilcoxon rank sum tests.

(G) Left panel: UMAP visualization for 15 topics generated with cisTopic on matched scATAC-Seq peak data with cells colored by celltype. Middle panels: Diffusion maps generated with Palantir on 15 topics with cells colored by celltype and by the probability of differentiating into HS-lineage. Right panel: Violin plots depicting the Palantir-derived probability (averaged over 50 runs) of differentiating into HS for all TACs (TACs grouped as in panel-(F), with the Palantir cell-fate

Figure5. *Continued*

probability computed over the 15 topics (derived from scATAC-Seq data). P-values are from a one-sided Wilcoxon rank-sum test. The horizontal dashed red line represents the HS-fate probability = 0.5.

(H) Left panel: Violin plots depicting the average accessibility score for HS driving RFs (*Runx2* and *Zfhx3*) across the 3 categories of TACs. P-values from a one-sided Wilcoxon rank-sum test. For each RF, its accessibility is calculated by adding the accessible counts across peaks within a region 3 kb upstream of the gene and the gene body, and the summed accessibility means adding the accessibility scores of the two driving RFs. Right panel: as left panel but for the IRS-priming RFs *Lmo7* and *Foxo3*.

DICE identifies primed states during hair-follicle development

To further demonstrate that primed states are identifiable from scRNA-Seq data only, we turned our attention to a scRNA-Seq study of hair-follicle regeneration, where hair-follicle stem cells initially give rise to short-lived transit-amplifying cells (TACs) that subsequently divide to produce differentiated cell types of the mature hair-follicle, including the inner root sheath (IRS) keratinocyte layer, and the hair shaft (HS), which itself consists of the cuticle cortex (CC) and medulla (M).[26] Thus, we asked if DICE would be able to identify multipotent TACs primed for differentiation into the alternative IRS and HS layers. To explore this, we selected RFs upregulated in one of the IRS and HS lineages relative to TACs and relative to the alternative lineage. Diffusion Map analysis over these RFs using destiny[40] confirmed the well-known bifurcation into competing IRS and HS lineages, with the M & CC bifurcation occurring at a later stage (Figure 5A). As before, we observed that upregulated RFs in each of the IRS and HS lineages encompassed two subtypes, depending on the frequency of expression in the multipotent (here TAC) population (Figure 5B), with the type-1 RFs displaying lower frequencies of expression and being more strongly enriched for biological terms related to differentiation of that lineage (Figure S5). Focusing on type-1 RFs, we verified that their DICE computed over TACs was significantly lower compared to a completely randomized TAC expression data matrix (Figure 5C), suggesting that priming is present within this TAC population. Using our dual perturbation Spearman rank correlation analysis we identified candidate pairs of RFs driving this priming (Figures 5D and 5E): *Foxo3* and *Lmo7* for priming into IRS, and *Runx2* and *Zfhx3* for HS. Next, we identified IRS-primed and HS-primed TACs as those co-expressing the respective pair of IRS or HS-priming RFs, resulting in 131 and 196 HS and IRS-primed cells, respectively. To confirm that these cells are primed to each lineage, we ran Palantir to estimate cell-fate probabilities for each TAC cell (Figure 5F). This revealed a stronger probability to differentiate into the HS-lineage for cells that DICE had predicted to be HS-primed, as required (Figure 5F).

To further validate these primed TAC states, we took advantage of the fact that all cells had been profiled with SHARE-Seq, a technology that generates joint scRNA-Seq and scATAC-Seq profiles in the same cells.[26] To process the scATAC-Seq data, we used cisTopic,[41] a dimensional reduction and Latent Dirichlet Allocation (LDA) framework based on topic modeling, to identify latent sources of variation termed ''regulatory topics''. With cisTopic, cells are clustered on the basis of their contributions to each regulatory topic. Using UMAP visualization over this latent space confirmed segregation of all TAC, IRS and HS cells by cell-type, with the medulla and cuticle-cortex cells displaying more similarity to each other compared to IRS cells, as required (Figure 5G). Next, we estimated cell-fate probabilities into IRS and HS lineages, by running Palantir on the cisTopic latent space (15 topics), revealing that the TAC cells DICE had previously predicted to be HS-primed are more likely to differentiate into the HS-lineage (Figure 5G). Thus, the cells we identified with DICE as being primed into HS or IRS lineages using scRNA-Seq data, are validated using matched scATAC-Seq profiles. Next, to validate the RF-pairs that DICE predicts to be driving these primed states (as derived from the scRNA-Seq data), we used the scATAC-Seq profiles to compute chromatin accessibility scores for the corresponding RFs in each of the TACs. This was done in two ways: one approach focused on the accessibility of the RFs themselves, whilst in the second we used chromVAR[39] to measure accessibility at the predicted RF-targets (STAR Methods). Validating their roles in HS-priming we observed that the HS-primed cells displayed higher accessibility scores for the predicted HS-RFs compared to unprimed TACs or TACs primed for the competing IRS-lineage (Figure 5H). The converse was also true for the IRS-priming RFs (Figure 5H). chromVAR-derived accessibility scores for the HS and IRS TFs with available target-information, were also higher in their corresponding primed cells (Figure S6), further validating our findings. Thus, overall, our data demonstrate that DICE can identify primed states from scRNA-Seq alone.

DICE identifies priming-RFs with higher confidence than other methods

Of note, if following the differential expression analysis (Figure 5B) we had selected candidate priming RFs using a criterion of highest variance across TAC cells, as opposed to using DICE, we would have selected a

distinct set of RFs (*Lef1* and *Dach1* for HS-lineage, *Sox5* and *Cux1* for IRS-lineage), with the primed states defined by these failing to validate in the scATAC-Seq data (Figure S7). This supports the view that consideration of expression covariation can improve the identification of primed states and their regulators over methods based on variance only. In support of this, we also compared DICE to an alternative method dubbed "Palantir-only" where we use Palantir directly to identify primed states using a threshold on the estimated cell-fate probabilities (STAR Methods). Applying this method to both the Ranzoni fetal liver hematopoietic as well as the hair-follicle SHARE-Seq datasets, we generally found candidate pairs of priming-RFs that were distinct from those inferred with DICE and whose evidence in priming was less well validated (Figures S8 and S9). For instance, considering the MPP population in liver hematopoiesis, Palantir-only identified XBP1 as the partner of UHRF1 in defining primed-GP states, yet the TFA of XBP1 failed to convincingly discriminate primed-GPs (primedG) from primed-MEMPs (primedM), suggesting that XBP1 is not directly implicated in priming (Figure S8). In contrast, DICE had predicted IRF8 to be implicated in priming of GP cells, which was well validated (Figure S8). Similarly, in the case of the hair-follicle data, using "Palantir-only" we found Gata3 and Maml3 to be priming-IRS factors, yet their accessibility was not higher in primed-IRS cells compared to primed-HS, which is inconsistent with their hypothesized role in priming (Figure S9).

DICE helps pinpoint cell-fate transition points in real scRNA-Seq data

Finally, we applied DICE to real scRNA-Seq datasets representing differentiation time courses to assess if DICE can help pinpoint the underlying bifurcations. Eligible datasets are those profiling sufficient numbers of cells ($\geq$50 cells per timepoint) across a sufficient number of timepoints where bifurcation dynamics is evident and grounded on prior biological evidence. We first analyzed a developmental time course of liver differentiation in mice (n = 447 cells), with hepatoblasts differentiating into both hepatocytes and cholangiocytes along 7 developmental timepoints (E10.5, E11.5, E12.5, E13.5, E14.5, E15.5 & E17.5).[42] Here we focused on a set of 22 liver-specific TFs which we have previously validated as exhibiting increased differentiation activity from hepatoblasts into either hepatocytes or cholangiocytes.[22] Of note, in this earlier work we made use of liver-specific TF regulons to estimate differentiation activity for the 22 liver-specific TFs in each of the 447 cells, an approach that improves the sensitivity to detect true dynamic changes of TFs, as compared to TF-expression.[22] Applying Diffusion Maps[35] to the differentiation activity matrix defined over the 22 TF and 447 cells confirmed the known bifurcation into the two main liver epithelial subtypes, although the precise timing of the bifurcation point is unclear due to low sampling-sparsity in this region (Figure 6A). We note that the imprecision in the timing of the bifurcation is also unclear had we derived the diffusion map using all variable genes.[22] Although pseudotime and developmental timepoint were strongly correlated (Figure S10), the substantial asynchrony displayed by cells (Figures 6B andS10) motivated us to study the covariation of TFs as a function of pseudotime (STAR Methods). We identified 4 high-confidence hepatocyte (*Foxa2, Nr1i3, Nr1i2, Trim15*) and cholangiocyte (*Lsr, Elf3, Bgn, Irf6*) specific factors (Figures S11 and6C), and computing their DICE as a function of pseudotime revealed a transition following E13.5 (Figure 6D), consistent with previous knowledge.[42] Importantly, DICE displayed a clear transition-like behavior, unlike the individual TF variances which increased but in a mostly asynchronous manner (Figure 6E). We also compared DICE to BioTIP,[19] a tool designed to detect cell-fate transitions using the concept of a dynamical network biomarker (DNB),[17,18]which also relies on covariation patterns, albeit not just of TFs but of specific subsets of all genes. In line with this, BioTIP's criticality index also displayed a clear transition, although at an earlier timepoint compared to DICE, whilst also displaying larger fluctuations during the timecourse (Figures S12A and S12B). This demonstrates that covariation of TFs and the derived DICE measure displays smoother behavior that can help pinpoint bifurcation points more precisely than what is possible from the inferred diffusion map bifurcation diagram, from consideration of TF variances only, or from DNB-based criticality indices.

As a second example, we considered differentiation of human fetal retinal progenitor cells into either neuronal or epithelial subtypes, another well-known bifurcation system.[43] For this dataset, we did not have an *a-priori* set of tissue-specific TFs, hence we inferred relevant TFs from the scRNA-Seq dataset itself using the measured TF-expression levels (STAR Methods). We identified 17 TFs (10 neuronal and 7 epithelial specific) displaying increased expression during the timecourse and clear antagonistic behavior between the two differentiation branches. Application of Diffusion Maps to the 17 TF-expression matrix confirmed the bifurcation into the two retinal subtypes, although, as before, the precise timing of the bifurcation is unclear (Figures S13A–S13C). By computing DICE over the top 7 TF from each lineage, we could readily identify the developmental timepoint 13W as the critical transition point (Figures S13D and S13E),

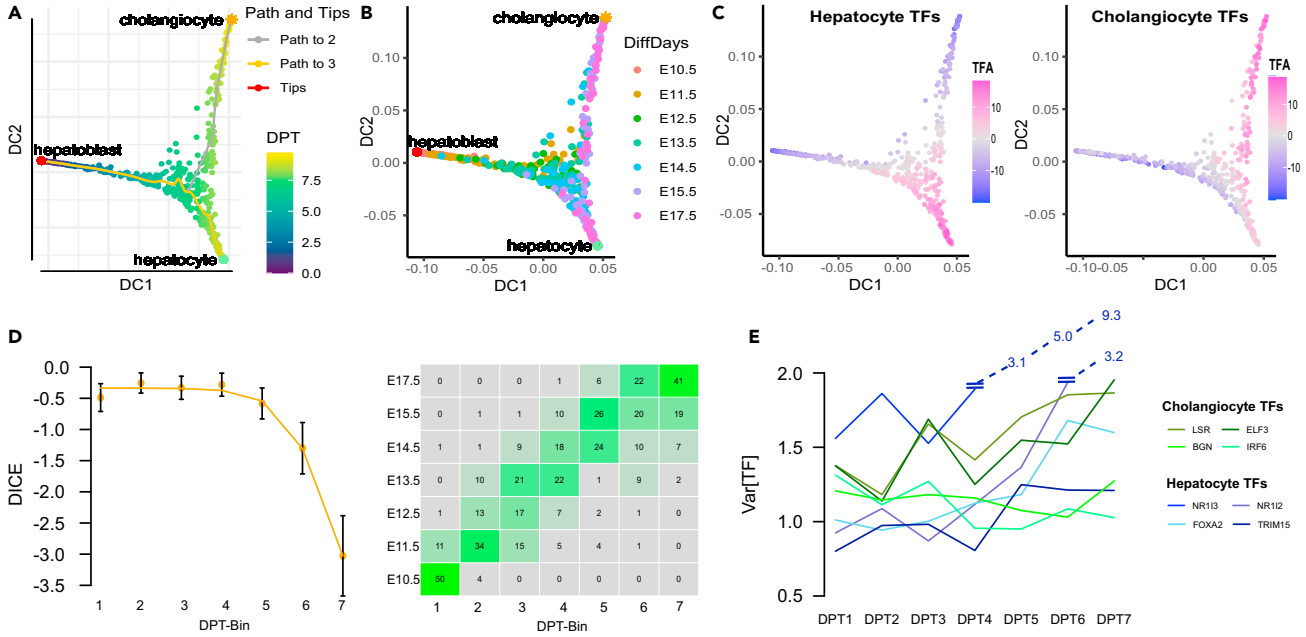


Figure6. DICE analysis in liver differentiation

(A) Diffusion Map bifurcation diagram as inferred over the TFA-matrix of 22 liver-specific TFs and 447 cells, with cells colored by diffusion pseudotime (DPT).

(B) As (A), but will cells colored by developmental timepoint.

(C) As (A), but with cells colored by the average TFA of hepatocyte (Foxa2, Nr1i3, Nr1i2, Trim15) and cholangiocyte factors (Lsr, Elf3, Bgn, Irf6).

(D) Left: DICE as a function of DPT-bin. Error bars indicate 95% CIs as estimated from 1000 Bootstraps. Right: overlap cell count matrix between developmental timepoint and DPT-bin.

(E) Variance of the hepatocyte and cholangiocyte TFs (estimated using TFA-matrix) as a function of DPT-bin. Data on y-axis has been capped at 2, with outlier values displayed in text.

consistent with previous knowledge.[43] Importantly, inspection of the individual TF variances or their average did not exhibit a clear transition-like behavior (Figures S13F and S13G), while BioTIP's criticality index displayed large fluctuations, not allowing precise identification of a cell-fate transition point (Figures S12C and S12D).

DISCUSSION

Here we have proposed the concept of distance covariance entropy to quantify priming and bifurcation dynamics in scRNA-Seq data. As shown, DICE can be a useful measure for identifying bifurcation points which otherwise could be hard to pinpoint when only using lineage-trajectory inference algorithms. This is because sampling of cells in the vicinity of bifurcation points is generally quite sparse which leads to high variance and blurring of the lineage-trajectory landscape near such transition points. Our DICE strategy circumvents this problem by identifying the relevant RFs that display the typical antagonistic switch-like behavior during differentiation, a task which only requires identification of the terminal branch points and which can therefore be easily accomplished with standard lineage trajectory or clustering algorithms. By subsequently computing the DICE of these RFs as a function of differentiation-or-pseudotime, we can thus track the covariation behavior of these RFs, to more accurately identify the specific timepoint at which DICE changes, reflecting the transition point. In this regard, it is worth pointing out that DICE displayed improvements over competing methods. For instance, although in the real datasets considered here we did not observe the covariation patterns between RFs to deviate strongly from monotonic linear dependencies, DICE did display a higher dynamic range in the vicinity of the transition points compared to the ordinary covariance entropy. We also observed improvements of DICE over DNB-based methods such as BioTIP, in the sense that DICE displayed much smoother patterns away from the cell-fate transition points. With BioTIP we often observed wilder fluctuations on either sides of the transition point. We attribute the improved smoothness of DICE to the fact that it is anchored on the expression or differentiation activity levels of RFs. In other words, although DNBs derive from expression covariation patterns, these are

inferred from much larger pools of genes, and hence are more likely to be purely associative and not causal for the cell-fate commitment process. It will be interesting if future work were to perform a more comprehensive comparison to include other tipping point algorithms such as scGET.[18]

Using DICE we have also obtained clear evidence of priming, i.e. non-stochastic covariation of lineage specific RFs within multipotent cell populations. We observed such priming within a hematopoietic MPP cell population during fetal liver hematopoiesis and within an MPP population driving hair-follicle regeneration, suggesting that this is a broad phenomenon. Although we were able to validate the primed states and RFs using scATAC-Seq data, it is important to observe that the identification of such primed states and their RFs did not require the scATAC-Seq itself. Indeed, DICE is able to infer the RFs defining the primed cells using only the scRNA-Seq data as input, without the need for any epigenome data. This is consistent with observations from Velten et al.,[25] but appears to contradict the claim made by others[27,44] that such primed states are only detectable using scATAC-Seq data. In our opinion, primed states should be detectable from scRNA-Seq data, because chromatin accessibility is determined by RFs, including epigenetic modifiers and pioneer TFs, and the expression levels of these factors must change first to facilitate subsequent changes in chromatin accessibility. In support of this, in the case of MPPs primed to MEMPs, the main identified RF-pair was GATA1-KLF1, the latter being a well-known pioneer TF that facilitates open chromatin and recruitment of GATA1.[45] Similarly, for MPPs primed to GPs, our candidate priming pair IRF8-UHRF1 involves an epigenetic modifier (UHRF1) involved in DNA methylation maintenance.[46] Consistent with our data, UHRF1 has been shown to be critical for granulocyte differentiation with UHRF1 KO cells being biased toward erythroid differentiation.[47] These findings are consistent with priming being associated with a particular chromatin modification state ("chromatin potential") that precedes cell-fate commitment,[44] and that such priming is detectable by subtle expression covariation patterns of specific RFs. Indeed, we note that one important reason why previous studies may have been unable to identify primed states from scRNA-Seq data,[27,44] is that they did not consider the covariation patterns of RFs.

Although we did not explore the application of DICE to a disease context, we envisage that DICE could also be useful as a means of inferring which differentiation networks are altered in diseases like cancer. Indeed, cell-lineage has recently been shown to be a critical determinant of oncogenesis,[48] and lineage-specific TFs are preferentially silenced in the corresponding cancer type.[22,49] Thus, DICE could help identify the specific combinatorial patterns of RFs whose covariation is disrupted, promoting oncogenic transformation.

In summary, the DICE framework and metric presented here is a novel useful tool to help quantify and identify primed states in multipotent cell populations, as well as to help pinpoint the exact timing of bifurcation events in differentiation and development.

Limitations of the study

Ideally, we would have had SHARE-Seq profiles (i.e., joint scRNA-Seq & scATAC-Seq) for the liver hematopoiesis study, yet this data was not available. As a result, to validate the primed states as identified from the scRNA-Seq data, we had to invent a method to assign corresponding primed states in the scATAC-Seq data. Although the method is rigorous, it is inevitably also an approximation. Ideally, one would also perform experimental work to demonstrate the importance of the identified RFs in priming. Instead, we used the generated scATAC-Seq data to validate the role of the RFs in priming.

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
- METHOD DETAILS
  - Single-cell omic datasets analyzed
  - Selection of lineage specific regulatory factors
- THE DISTANCE COVARIANCE ENTROPY (DICE) METRIC

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Quantification of priming in pluri-and-multipotent cell populations
  - Dimensionality reduction of scRNA-Seq data and scATAC-Seq data
  - Cell fate probability calculation
  - Identification and validation of primed cells in scATAC-seq data
  - Identification of primed states using variance in hair-follicle regeneration data
  - Identification of cell-fate transition points from differentiation time course scRNA-Seq data
  - BioTIP analysis
  - Palantir-only analysis
  - Description of GRNs and simulation of scRNA-Seq from GRNs

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105709.

ACKNOWLEDGMENTS

We thank the scientific community who support open access and deposit data without restrictions in public databases for the benefit of mankind. Funding: This project was funded by the National Natural Science Foundation of China (grant numbers 31970632 and 32170652). We are also grateful to Michael Nelson and Ana Cvejic for stimulating discussions.

AUTHOR CONTRIBUTIONS

Study was conceived by A.E.T. Analyses were performed by Q.L., A.M., and A.E.T. MS was written by A.E.T. with contributions from Q.L. and A.M.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: July 20, 2022
Revised: November 8, 2022
Accepted: November 29, 2022
Published: December 22, 2022

REFERENCES

1. Graf, T., and Enver, T. (2009). Forcing cells to change lineages. Nature *462*, 587–594. https://doi.org/10.1038/nature08533.

2. Moris, N., Pina, C., and Arias, A.M. (2016). Transition states and cell fate decisions in epigenetic landscapes. Nat. Rev. Genet. *17*, 693–703. https://doi.org/10.1038/nrg.2016.98.

3. Heinäniemi, M., Nykter, M., Kramer, R., Wienecke-Baldacchino, A., Sinkkonen, L., Zhou, J.X., Kreisberg, R., Kauffman, S.A., Huang, S., and Shmulevich, I. (2013). Gene-pair expression signatures reveal lineage control. Nat. Methods *10*, 577–583. https://doi.org/10.1038/nmeth.2445.

4. Huang, S., Guo, Y.P., May, G., and Enver, T. (2007). Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. Dev. Biol. *305*, 695–713. https://doi.org/10.1016/j.ydbio.2007.02.036.

5. Bargaje, R., Trachana, K., Shelton, M.N., McGinnis, C.S., Zhou, J.X., Chadick, C., Cook, S., Cavanaugh, C., Huang, S., and Hood, L. (2017). Cell population structure prior to bifurcation predicts efficiency of directed differentiation in human induced pluripotent cells. Proc. Natl. Acad. Sci. USA *114*, 2271–2276. https://doi.org/10.1073/pnas.1621412114.

6. Teschendorff, A.E., and Feinberg, A.P. (2021). Statistical mechanics meets single-cell biology. Nat. Rev. Genet. *22*, 459–476. https://doi.org/10.1038/s41576-021-00341-z.

7. Ferrell, J.E., Jr. (2012). Bistability, bifurcations, and Waddington's epigenetic landscape. Curr. Biol. *22*, R458–R466. https://doi.org/10.1016/j.cub.2012.03.045.

8. Wang, J., Zhang, K., Xu, L., and Wang, E. (2011). Quantifying the Waddington landscape and biological paths for development and differentiation. Proc. Natl. Acad. Sci. USA *108*, 8257–8262. https://doi.org/10.1073/pnas.1017017108.

9. Flouriot, G., Jehanno, C., Le Page, Y., Le Goff, P., Boutin, B., and Michel, D. (2020). The basal level of gene expression associated with chromatin loosening shapes Waddington landscapes and controls cell differentiation. J. Mol. Biol. *432*, 2253–2270. https://doi.org/10.1016/j.jmb.2020.02.016.

10. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods *6*, 377–382. https://doi.org/10.1038/nmeth.1315.

11. Stuart, T., and Satija, R. (2019). Integrative single-cell analysis. Nat. Rev. Genet. *20*, 257–272. https://doi.org/10.1038/s41576-019-0093-7.

12. Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. Nat. Biotechnol. *37*, 547–554. https://doi.org/10.1038/s41587-019-0071-9.

13. Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–386. https://doi.org/10.1038/nbt.2859.

14. Grün, D. (2020). Revealing dynamics of gene expression variability in cell state space. Nat. Methods *17*, 45–49. https://doi.org/10.1038/s41592-019-0632-3.

15. Richard, A., Boullu, L., Herbach, U., Bonnafoux, A., Morin, V., Vallin, E., Guillemin, A., Papili Gao, N., Gunawan, R., Cosette, J., et al. (2016). Single-cell-based analysis highlights a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process. PLoS Biol. *14*. e1002585. https://doi.org/10.1371/journal.pbio.1002585.

16. Mojtahedi, M., Skupin, A., Zhou, J., Castaño, I.G., Leong-Quong, R.Y.Y., Chang, H., Trachana, K., Giuliani, A., and Huang, S. (2016). Cell fate decision as high-dimensional critical state transition. PLoS Biol. *14*. e2000640. https://doi.org/10.1371/journal.pbio.2000640.

17. Chen, L., Liu, R., Liu, Z.P., Li, M., and Aihara, K. (2012). Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. Sci. Rep. *2*, 342. https://doi.org/10.1038/srep00342.

18. Zhong, J., Han, C., Zhang, X., Chen, P., and Liu, R. (2021). scGET: predicting cell fate transition during early embryonic development by single-cell graph entropy. Genom. Proteom. Bioinform. *19*, 461–474. https://doi.org/10.1016/j.gpb.2020.11.008.

19. Yang, X.H., Goldstein, A., Sun, Y., Wang, Z., Wei, M., Moskowitz, I.P., and Cunningham, J.M. (2022). Detecting critical transition signals from single-cell transcriptomes to infer lineage-determining transcription factors. Nucleic Acids Res. *50*, e91. https://doi.org/10.1093/nar/gkac452.

20. Liu, J., Ding, D., Zhong, J., and Liu, R. (2022). Identifying the critical states and dynamic network biomarkers of cancers based on network entropy. J. Transl. Med. *20*, 254. https://doi.org/10.1186/s12967-022-03445-0.

21. Peng, H., Zhong, J., Chen, P., and Liu, R. (2022). Identifying the critical states of complex diseases by the dynamic change of multivariate distribution. Brief. Bioinform. *23*. bbac177. https://doi.org/10.1093/bib/bbac177.

22. Teschendorff, A.E., and Wang, N. (2020). Improved detection of tumor suppressor events in single-cell RNA-Seq data. NPJ Genom. Med. *5*, 43. https://doi.org/10.1038/s41525-020-00151-y.

23. Holland, C.H., Tanevski, J., Perales-Patón, J., Gleixner, J., Kumar, M.P., Mereu, E., Joughin, B.A., Stegle, O., Lauffenburger, D.A., Heyn, H., et al. (2020). Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. Genome Biol. *21*, 36. https://doi.org/10.1186/s13059-020-1949-z.

24. Chen, S., and Mar, J.C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinf. *19*, 232. https://doi.org/10.1186/s12859-018-2217-z.

25. Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. Nat. Cell Biol. *19*, 271–281. https://doi.org/10.1038/ncb3493.

26. Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., et al. (2020). Chromatin potential identified by shared single-cell profiling of RNA and chromatin. Cell *183*, 1103–1116.e20. https://doi.org/10.1016/j.cell.2020.09.056.

27. Ranzoni, A.M., Tangherloni, A., Berest, I., Riva, S.G., Myers, B., Strzelecka, P.M., Xu, J., Panada, E., Mohorianu, I., Zaugg, J.B., and Cvejic, A. (2021). Integrative single-cell RNA-seq and ATAC-seq analysis of human developmental hematopoiesis. Cell Stem Cell *28*, 472–487.e7. https://doi.org/10.1016/j.stem.2020.11.015.

28. Teschendorff, A.E., and Relton, C.L. (2018). Statistical and integrative system-level analysis of DNA methylation data. Nat. Rev. Genet. *19*, 129–147. https://doi.org/10.1038/nrg.2017.86.

29. Ghanbari, M., Lasserre, J., and Vingron, M. (2019). The Distance Precision Matrix: computing networks from non-linear relationships. Bioinformatics *35*, 1009–1017. https://doi.org/10.1093/bioinformatics/bty724.

30. Szkely, G.J., and Rizzo, M.L. (2017). The energy of data. Annu. Rev. Stat. Appl. *4*, 447–479.

31. van Wieringen, W.N., and van der Vaart, A.W. (2011). Statistical analysis of the cancer cell's molecular entropy using high-throughput data. Bioinformatics *27*, 556–563. https://doi.org/10.1093/bioinformatics/btq704.

32. Ye, Y., Kang, X., Bailey, J., Li, C., and Hong, T. (2019). An enriched network motif family regulates multistep cell fate transitions with restricted reversibility. PLoS Comput. Biol. *15*. e1006855. https://doi.org/10.1371/journal.pcbi.1006855.

33. Chang, R., Shoemaker, R., and Wang, W. (2011). Systematic search for recipes to generate induced pluripotent stem cells. PLoS Comput. Biol. *7*. e1002300. https://doi.org/10.1371/journal.pcbi.1002300.

34. Li, C., and Wang, J. (2013). Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: landscape and biological paths. PLoS Comput. Biol. *9*. e1003165. https://doi.org/10.1371/journal.pcbi.1003165.

35. Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F., and Theis, F.J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. Nat. Methods *13*, 845–848. https://doi.org/10.1038/nmeth.3971.

36. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.

37. Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L., and Pe'er, D. (2019). Characterization of cell fate probabilities in single-cell data with Palantir. Nat. Biotechnol. *37*, 451–460. https://doi.org/10.1038/s41587-019-0068-4.

38. Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. *37*, 38–44. https://doi.org/10.1038/nbt.4314.

39. Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods *14*, 975–978. https://doi.org/10.1038/nmeth.4401.

40. Angerer, P., Haghverdi, L., Büttner, M., Theis, F.J., Marr, C., and Buettner, F. (2016). destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics *32*, 1241–1243. https://doi.org/10.1093/bioinformatics/btv715.

41. Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J., and Aerts, S. (2019). cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. Nat. Methods *16*, 397–400. https://doi.org/10.1038/s41592-019-0367-1.

42. Yang, L., Wang, W.H., Qiu, W.L., Guo, Z., Bi, E., and Xu, C.R. (2017). A single-cell transcriptomic analysis reveals precise pathways and regulatory mechanisms underlying hepatoblast differentiation. Hepatology *66*, 1387–1401. https://doi.org/10.1002/hep.29353.

43. Hu, Y., Wang, X., Hu, B., Mao, Y., Chen, Y., Yan, L., Yong, J., Dong, J., Wei, Y., Wang, W., et al. (2019). Dissecting the transcriptome landscape of the human fetal neural retina and retinal pigment epithelium by single-cell RNA-seq analysis. PLoS Biol. *17*. e3000365. https://doi.org/10.1371/journal.pbio.3000365.

44. Buenrostro, J.D., Corces, M.R., Lareau, C.A., Wu, B., Schep, A.N., Aryee, M.J., Majeti, R., Chang, H.Y., and Greenleaf, W.J. (2018). Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. Cell *173*, 1535–1548.e16. https://doi.org/10.1016/j.cell.2018.03.074.

45. Gillinder, K.R., Magor, G., Bell, C., Ilsley, M.D., Huang, S., and Perkins, A. (2018). KLF1 acts as a pioneer transcription factor to open

chromatin and facilitate recruitment of GATA1. Blood *132*, 501.

46. Bostick, M., Kim, J.K., Estève, P.O., Clark, A., Pradhan, S., and Jacobsen, S.E. (2007). UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science *317*, 1760–1764. https://doi.org/10.1126/science.1147939.

47. Zhao, J., Chen, X., Song, G., Zhang, J., Liu, H., and Liu, X. (2017). Uhrf1 controls the self-renewal versus differentiation of hematopoietic stem cells by epigenetically regulating the cell-division modes. Proc. Natl. Acad. Sci. USA *114*, E142–E151. https://doi.org/10.1073/pnas.1612967114.

48. Sahu, B., Pihlajamaa, P., Zhang, K., Palin, K., Ahonen, S., Cervera, A., Ristimäki, A., Aaltonen, L.A., Hautaniemi, S., and Taipale, J. (2021). Human cell transformation by combined lineage conversion and oncogene expression. Oncogene *40*, 5533–5547. https://doi.org/10.1038/s41388-021-01940-0.

49. Liu, T., Zhao, X., Lin, Y., Luo, Q., Zhang, S., Xi, Y., Chen, Y., Lin, L., Fan, W., Yang, J., et al. (2022). Computational identification of preneoplastic cells displaying high stemness and risk of cancer progression. Cancer Res. *82*, 2520–2537. https://doi.org/10.1158/0008-5472.CAN-22-0668.

50. Rainer, J., Gatto, L., and Weichenberger, C.X. (2019). ensembldb: an R package to create and use Ensembl-based annotation resources. Bioinformatics *35*, 3151–3153. https://doi.org/10.1093/bioinformatics/btz031.

51. Fornes, O., Castro-Mondragon, J.A., Khan, A., van der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranašić, D., et al. (2020). Jaspar 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *48*, D87–D92. https://doi.org/10.1093/nar/gkz1001.

52. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420. https://doi.org/10.1038/nbt.4096.

53. Pineda-Krch, M. (2008). Implementing the stochastic simulation algorithm in R. J. Stat. Softw. *25*, 1–18.

54. Soetaert, K. (2012). Solving Differential Equations in R (Springer).

55. Stuart, T., Srivastava, A., Madad, S., Lareau, C.A., and Satija, R. (2021). Single-cell chromatin state analysis with Signac. Nat. Methods *18*, 1333–1341. https://doi.org/10.1038/s41592-021-01282-5.

56. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296. https://doi.org/10.1038/s41592-019-0619-0.

57. Teschendorff, A.E., Zhuang, J., and Widschwendter, M. (2011). Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. Bioinformatics *27*, 1496–1505. https://doi.org/10.1093/bioinformatics/btr171.

58. Miyamoto, T., Furusawa, C., and Kaneko, K. (2015). Pluripotency, differentiation, and reprogramming: a gene expression dynamics model with epigenetic feedback regulation. PLoS Comput. Biol. *11*. e1004476. https://doi.org/10.1371/journal.pcbi.1004476.

59. Alagha, A., and Zaikin, A. (2013). Asymmetry in erythroid-myeloid differentiation switch and the role of timing in a binary cell-fate decision. Front. Immunol. *4*, 426. https://doi.org/10.3389/fimmu.2013.00426.

60. Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. *81*, 2340–2361. https://doi.org/10.1021/j100540a008.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Fetal liver hematopoiesis scRNA-Seq data	Ranzoni et al.[27]	https://gitlab.com/cvejic-group/integrative-scrna-scatac-human-foetal/-/tree/master/Data/ScanpyObjets
Fetal liver hematopoiesis scATAC-Seq data	Ranzoni et al.[27]	https://u.pcloud.link/publink/show?code=XZdMm1XZssJXAvpOTVyH183w4QwCN4PtroPk
Fetal liver hematopoiesis metadata and TF activity data	Ranzoni et al.[27]	https://gitlab.com/cvejic-group/integrative-scrna-scatac-human-foetal/-/tree/master/Data/scATAC_CSV_file_for_Scanpy
Mouse skin SHARE-Seq data	Ma et al.[26]	GEO: GSE140203
Mouse liver differentiation	Yang et al.[42]	GEO: GSE90047
Human fetal retina differentiation	Hu et al.[43]	GEO: GSE107618
Molecular Signature database	Subramanian A et al.[36]	https://www.gsea-msigdb.org/gsea/msigdb
DOROTHEA database	Holland CH et al.[23]	https://saezlab.github.io/DoRothEA/
EnsDb.Mmusculus.v79	Rainer J et al.[50]	https://bioconductor.org/packages/EnsDb.Mmusculus.v79/
BSgenome.Mmusculus.UCSC.mm10	The Bioconductor Dev Team	https://bioconductor.org/packages/BSgenome.Mmusculus.UCSC.mm10/
JASPAR2020	Fornes O et al.[51]	http://jaspar.genereg.net/
Software and algorithms		
Seurat v4.0.4	Butler et al.[52]	https://cran.r-project.org/web/packages/Seurat/
Destiny v3.11.0	Angerer et al.[40]	https://bioconductor.org/packages/release/bioc/html/destiny.html
DICE v0.9.1	This paper	https://github.com/aet21/DICE
BioTIP v1.11.0	Yang et al.[19]	https://bioconductor.org/packages/release/bioc/vignettes/BioTIP/inst/doc/BioTIP.html
Scira v1.0.3	Teschendorff et al.[22]	https://github.com/aet21/scira
GillespieSSA v0.6.2	Pineda-Krch[53]	https://github.com/rcannood/GillespieSSA
deSolve v1.34	Soetaert[54]	https://cran.r-project.org/web/packages/deSolve/
Signac v1.4.0	Stuart T et al.[55]	https://cloud.r-project.org/package=Signac
harmony v0.1.0	Korsunsky I et al.[56]	https://github.com/immunogenomics/harmony
cisTopic v0.3.0	Bravo Gonzalez-Blas C et al.[41]	http://github.com/aertslab/cistopic
Palantir v0.2.1	Setty M et al.[37]	https://github.com/dpeerlab/Palantir/
isva v1.9	Teschendorff AE et al.[57]	https://CRAN.R-project.org/package=isva
chromVAR v1.16.0	Schep AN et al.[39]	http://www.github.com/GreenleafLab/chromVAR

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Andrew Teschendorff (andrew@sinh.ac.cn).

Materials availability

This study does not generate novel data.

Data and code availability

SHARE-Seq data analyzed in this study is publicly available from the Gene Expression Omnibus (GEO) under accession number GSE140203. Single-cell RNA-Seq data for mouse liver and human fetal retina

analyzed in this study is publicly available from the GEO under accession numbers, GSE90047 and GSE107618.

The main functions implementing the DICE framework are available as part of the DICE R-package, freely available from https://github.com/aet21/DICE. The R-package comes with a user-friendly tutorial vignette.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Single-cell omic datasets analyzed

We analyzed a total of four single-cell RNA-seq and two single-cell ATAC-Seq datasets, respectively.

scRNA-Seq and scATAC-Seq data of fetal hematopoiesis in liver

scRNA-Seq and scATAC-Seq data was derived from Ranzoni et al.[27] QC and normalization for scRNA-Seq data were done with *filter_cells* and *normalize_per_cell* functions (normalize_per_cell function with scaling factor 10,000 and log1p function) from Python (v.3.6.9) package SCANPY (v.1.4.5.1). After QC, 4504 cells remained. Their annotation was given by Ranzoni et al. as 1200 MPPs, 169 cycling MPPs (MPPs-Cycle), 156 lymphoid-myeloid progenitors (LMPs), 192 MK-erythroid-mast progenitors (MEMPs), 255 cycling MEMPs (MEMPs-Cycle), 177 granulocytic progenitors (GPs), 265 granulocytes, 254 erythroid cells, 200 mast cells, 96 megakaryocytes, 135 plasmacytoid dendritic cells (pDCs), 42 cycling pDCs, 476 monocytes, 756 B cells, 41 endothelial cells, 46 NK cells and 44 unspecified cells. The Scanpy Object containing the normalized scRNA-Seq data for cells passing QC and annotation information was downloaded from https://gitlab.com/cvejic-group/integrative-scrna-scatac-human-foetal/-/tree/master/Data/ScanpyObjets. The Seurat object of scATAC-Seq data containing a peak assay for 3611 cells passing QC was downloaded from https://u.pcloud.link/publink/show?code=XZdMm1XZssJXAvpOTVyH183w4QwCN4PtroPk. For scATAC-Seq data we also used the annotation provided by the authors: there were 2264 MPPs, 313 cycling MPPs, 133 MEMPs, 265 cycling MEMPs, 186 GP, 185 LMPs and 265 unspecified cells. The metadata and TF activity calculated with chromVAR files for scATAC-Seq data were downloaded from https://gitlab.com/cvejic-group/integrative-scrna-scatac-human-foetal/-/tree/master/Data/scATAC_CSV_file_for_Scanpy.

SHARE-seq data of mouse skin

SHARE-seq data, containing joint profiles of single cell gene expression and chromatin accessibility, was downloaded from GEO: GSE140203[27]. We used 34,774 joint profiles from mouse skin that passed QC with cell-type annotations, encompassing 4378 transit-amplifying cells (TACs), 3433 $CD34^{+}$ bulge cells, 7787 basal cells, 1121 Dermal Fibroblast cells, 766 dermal papilla cells, 398 dermal sheath cells, 927 endothelial cells, 291 granular cells, 1166 hair shaft cuticle cortex cells, 4139 infundibulum cells, 672 inner root sheath (IRS) cells, 689 isthmus cells, 514 K6+ bulge companion layer cells, 263 macrophage DCs, 981 medulla cells, 187 melanocytes, 1029 outer root sheath cells (ORS), 163 Schwann cells, 181 sebaceous gland cells, 3146 spinous cells. scRNA-Seq data was normalized with Seurat.[52] For our analysis, we focused on the joint scRNA-Seq scATAC-Seq profiles of 4378 TACs, 672 IRS, 981 medulla and 1166 cuticle cortex cells.

scRNA-Seq data of liver differentiation

This Fluidigm C1 dataset was derived from Yang et al.,[42] a study of differentiation of mouse hepatoblasts into hepatocytes and cholangiocytes. Normalized (TPM) data was downloaded from GEO:GSE90047 (file: GSE90047-Singlecell-RNA-seq-TPM.txt). Data was further transformed using a log2 transformation adding a pseudocount of 1. After quality control, 447 single-cells remained, with 54 single cells at embryonic day 10.5 (E10.5), 70 at E11.5, 41 at E12.5, 65 at E13.5, 70 at E14.5, 77 at E15.5 and 70 at E17.5.

scRNA-Seq data of human retina differentiation

This STRT scRNA-Seq dataset derives from Hu et al.,[43] representing a time-course differentiation of human fetal retinal progenitor cells into retinal neuronal and retinal epithelial cells (GEO:GSE107618). After quality control (following Hu et al.), 2421 single-cells remained encompassing 10 developmental stages (5W–24W).

Selection of lineage specific regulatory factors

For each dataset analyzed, the RFs used in DICE were selected as follows:

Fetal hematopoiesis in liver

Starting out from a combined list of 1994 RFs derived from the Molecular Signature (MSigDB) and DOROTHEA databases,[23,36] we performed differential expression analysis on the log-normalized scRNA-seq data from[27] encompassing 1569 non-cycling cells annotated as MPP (n = 1200), megakaryocyte-erythroid-mast progenitors (MEMPs, n = 192) and granulocyte progenitors (GPs, n = 177). MEMP specific RFs were selected by performing Wilcoxon rank sum tests comparing expression profiles between MEMPs and MPPs, and separately also between MEMPs and GPs, so as to find RFs with significantly higher expression values in MEMPs in both comparisons (MEMP-RFs). We corrected for multiple testing using a Benjamini-Hochberg (BH) significance threshold of 0.05. An analogous procedure was implemented for selecting GP specific RFs (GP-RFs). We observed that both MEMP-RFs and GP-RFs displayed two different types of overexpression depending on the frequency of expression in the MPP population. For each MEMP and GP specific RF we computed the expression frequency across all MPP cells, and those with an expression frequency lower than the mean expression frequency were classified as "type-1", while the rest fell into "type-2". This was done separately for MEMP and GP specific RFs, leading to type-1 MEMP-RFs, type-2 MEMP-RFs, type-1 GP-RFs and type-2 GP-RFs. Of note, only type-1 RFs were used to assess priming, since priming is not a frequent event among multipotent cells.

Hair-follicle regeneration

Using the log-normalized scRNA-Seq data encompassing TACs, IRS, medulla and hair shaft cuticle cortex cells, we performed a similar differential overexpression analysis as described above to select type-1 lineage specific RFs for IRS and hair shaft (HS) lineages. For the HS, we combined medulla and cuticle cortex cells as one broad HS group. We selected the IRS and HS specific RFs from the 1994 RFs mentioned above. The IRS specific RFs were selected by performing Wilcoxon rank sum tests comparing expression profiles between TACs and IRS, and separately also between IRS and HS, so as to find RFs with significantly higher expression values in IRS in both comparisons (IRS-RFs). We corrected for multiple testing using a Benjamini-Hochberg (BH) significance threshold of 0.05. HS specific RFs were selected with an analogous procedure (HS-RFs). Similarly, IRS-RFs and HS-RFs displayed two different types of overexpression in relation to the frequency of expression in the TAC population. For each IRS and HS specific RF we computed the expression frequency across all TACs, and those with an expression frequency lower than the mean expression frequency were classified as "type-1", while the rest as "type-2". By doing this separately for IRS and HS specific RFs, we obtained type-1 IRS-RFs, type-2 IRS-RFs, type-1 HS-RFs and type-2 HS-RFs, and only type-1 RFs were used to assess priming.

Mouse liver differentiation

In this dataset, we used a prior list of 22 liver-specific TFs derived and validated by us previously.[22] This list included well-known hepatocyte and cholangiocyte factors Hnf4a, Foxa2, Hnf4g, Hnf1a, Elf3, Bcl3, Lhx2, Trim15, Lsr, Irf6, Bgn. Because for liver we have corresponding liver-specific regulons for these 22 TF,[22] in this dataset we don't use TF-expression but instead estimate the transcription factor differentiation activity (TFA) of the 22 TF. Briefly, the estimation of TFA proceeds by performing a linear regression of the log-normalized expression profile of a cell against the regulon binding profile, with the inferred t-statistic of the regression defining the TFA-value.[22] The regulon binding profile is a vector with entries equal to 1 indicating positively regulated targets of the TF, −1 indicating repressive interactions and 0 indicating no regulation. Next, from the estimated TFA profiles, we selected the TFs displaying increased TFA with differentiation timepoint, and which also displayed significant differential TFA between cholangiocyte and hepatocyte branches. Finally, DICE was computed over the top-4 selected hepatocyte and top-4 selected cholangiocyte factors (4 was chosen because this was the minimum number of selected TFs per branch).

Human retina differentiation

In this dataset we started out with the list of TFs from MSigDB.[36] Because cells were not annotated, we first used Seurat (4.0.4)[52] to perform normalization and cluster analysis, defining 11 cell clusters. To annotate them, we identified corresponding marker genes using the Seurat function FindMarkers(only.pos = TRUE, min.pct = 0.25, logfc.threshold = 1). For further analyses we discarded all clusters representing immune cells, fibroblasts and microglia, because they are unrelated to the retinal differentiation process, only

keeping the neuronal and epithelial cells. We identified specific clusters representing the terminal differentiation state (23–24W) of retinal neuronal cells and others associated with the terminal differentiation state (23–24W) of retinal epithelial cells. Next, we identified a total 17 TF displaying increased expression during the whole developmental process (5W–24W) and which also displayed antagonistic differential expression between the terminal neuronal (n = 10) and epithelial branches (n = 7). Finally, for the DICE computation we selected the top 7 retinal neuronal (NR2E3, RCVRN, NRL, NEUROD4, CRX, NEUROD1, VSX1) and top 7 retinal epithelial (TTLL4, BMP4, NFIX, EPAS1, NFE2L1, PDLIM5, LITAF) specific TFs.

THE DISTANCE COVARIANCE ENTROPY (DICE) METRIC

Here we describe the DICE metric that we use to quantify priming in pluri-or-multipotent cell populations or to detect cell-fate bifurcation events from scRNA-Seq data. In both applications, DICE is computed over a set of carefully selected regulatory factors (RFs), using either their measured expression values or using differentiation regulatory activity estimates derived from a regulon-based approach. See other subsections below for how these RFs are selected. Here we shall assume that these RFs have already been selected. The aim of the DICE metric is to capture the degree of covariation between these RFs in a multi-or-pluripotent cell population or as a function of time/pseudotime, and where due to the complexity of real GRNs, this covariation could be non-linear or non-monotonic. Conveniently, DICE quantifies the overall degree of covariation using entropy. In effect, the DICE metric calculation is based on the following two concepts:

(i) Because of the complexity and non-linearity of GRNs, associations between TFs can be non-monotonic and non-linear. In order to capture these complex dependencies, we use the recently proposed distance correlation/covariance measures.[29,30] The beauty of these distance-based measures is that they allow definition of ordinary linear covariance matrices, but in the space of pairwise distances between data points, which allows any non-linear/non-monotonic form of dependency between observables to be captured. In more detail, given two random variables X and Y, which could represent the differentiation activity (TFA) or expression levels of two TFs, and given n sample draws from each $x_i = X(x_i)$, $y_i = Y(y_i)$, $i = 1, \ldots, n$ (these could represent n cells collected at a given timepoint or within a pseudotime bin), the distance correlation is obtained by first computing the distance matrices $a_{ij} = \|x_i - x_j\|_2$ and $b_{ij} = \|y_i - y_j\|_2$, where we have here used an Euclidean distance norm. These distance matrices are then double centered to yield A_{ij} and B_{ij} matrices as follows:

$$A_{ij} = a_{ij} - \frac{1}{n}\sum_{k=1}^{n} a_{ik} - \frac{1}{n}\sum_{k=1}^{n} a_{kj} + \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}$$

$$B_{ij} = b_{ij} - \frac{1}{n}\sum_{k=1}^{n} b_{ik} - \frac{1}{n}\sum_{k=1}^{n} b_{kj} + \frac{1}{n^2}\sum_{k,l=1}^{n} b_{kl}$$

The distance covariance $dCov(X,Y)$ and distance correlation $dCor(X,Y)$ are then defined as

$$dCov(X,Y) = \frac{1}{n^2}\sum_{i,j=1}^{n} A_{ij}B_{ij}$$

$$dCor(X,Y) = \frac{dCov(X,Y)}{\sqrt{dCov(X,X)dCov(Y,Y)}}$$

Of note, the distance correlation/covariance can be computed for a set of d RFs/TFs ($p \geq 2$), and the expression or TFA values for the RFs/TFs can always be scaled so that $dCor = dCov$. Here we always scale the profiles in such a way that $dCor = dCov$, and hence we interchangeably use the terms distance correlation and distance covariance. Thus, given a set of d TFs with $\alpha = 1 \ldots d$ and X_α denoting the expression or TFA profile of TF α, we can construct a data covariance matrix Σ with entries $\Sigma_{\alpha\beta} = dCov(X_\alpha, X_\beta)$ where $dCov$ is as defined above.

(ii) Assuming we have d RFs, where d is typically less than n (the number of cells), we quantify the overall degree of dependency of the RFs using entropy. Assuming a multivariate Gaussian variable $X = (X_1, \ldots, X_d)$ of mean $\mu \in \mathbb{R}^d$ and $d \times d$ data covariance Σ, its entropy is defined by

$$H(X) = -\int dX\, p(X) \log p(X)$$

where $p(X) = G(\mu, \Sigma)$ is the multivariate Gaussian. Our key insight is to replace the covariance matrix Σ with the distance covariance matrix, as defined in point (i) above. Henceforth we refer to this entropy as "DICE" (DIstance Covariance Entropy).

In order to compute DICE or PCC, we note that the entropy expression can be rewritten as

$$H(X) = \frac{1}{2}\log|\Sigma| + \frac{d}{2}(1 + \log 2\pi)$$

The determinant of any covariance matrix can be expressed in terms of the product of its eigenvalues, and thus the above equation can be simplified to[31]:

$$H(X) = \frac{1}{2}\sum_{i=1}^{d} \log \lambda_i + \frac{d}{2}(1 + \log 2\pi)$$

where λ_i is the i-th eigenvalue of Σ. We note that the second term in the expression above is a constant, since d is generally fixed for any given system. Thus, the entropy varies as a function of the eigenvalues of the covariance matrix. For convenience we shall be using the simplified definition $H^*(X) = \frac{1}{2}\sum_{i=1}^{d} \log \lambda_i$ which is bounded above by zero. This is because for d *i.i.d* RVs X_i, i.e for a random $d \times n$ matrix X, its entropy will be maximal and zero.

Several notes with the above definition are in order: First of all, it only requires strong dependency for any 2 of the d RFs for the DICE value to drop significantly relative to the random null distribution. This feature makes DICE very appealing for detecting the earliest regulatory changes underlying cell-fate commitment, as is the case with priming. A corollary of this, is that we often need to determine the specific pairs or sets of RFs that "drive" the deviations from the random null distribution. How this is done is described in the next subsection. Second, DICE makes distributional assumptions which strictly speaking are not satisfied for expression derived from scRNA-Seq data, although they are satisfied for transcription factor activity (TFA)-values (see later). Extensive simulations we have performed however demonstrate that for all practical purposes the *dCor(X,Y)* measure above can be replaced with a corresponding non-parametric Spearman rank correlation coefficient computed over the vectorized distance matrices A and B. In practice, we use pairwise Spearman correlations to confirm the specific predictions of DICE as to which RFs drive priming (see further below).

QUANTIFICATION AND STATISTICAL ANALYSIS

Quantification of priming in pluri-and-multipotent cell populations

Statistical significance of priming

Given a selected set of RFs and a pool of multi-or-pluripotent cells, we first compute the DICE-metric of the RFs, using either their mRNA expression values, or their TFA profiles. To assess the statistical significance of the obtained DICE value, we derive a null distribution for DICE by randomly permuting the values for each RF across cells, using a different permutation for each RF. This randomization destroys any correlative structure that might exist in the data. We typically perform on the order of 1000 to 10,000 Monte-Carlo randomizations, and derive a P-value by calculating the fraction of randomizations that lead to a lower DICE value than the observed one. If the P-value is significant, this means that at least two RFs display non-random associations across the pluri-or-multipotent cells.

Identification of RFs driving primed states

To identify the key RFs displaying these non-random associations, i.e the RFs defining the putative primed states, we devised a dual perturbation DICE and Spearman correlation strategy. In the perturbation approach, and assuming d RFs, we recomputed DICE a total of d times, each time excluding one of the RFs. The difference between the new DICE values and the original one can be viewed as a measure of how much a given RF contributes to the original DICE value. For instance, if the DICE value computed over the d TFs is very low (indicating strong dependency), and if removing a given TF x leads to a substantial increase in DICE, then we can conclude that TF x drives the lower entropy and is a candidate "priming TF". The perturbation analysis allows ranking of the d RFs in order of decreasing importance. Because the

distributional assumptions underlying DICE may not be satisfied, we supplement the perturbation analysis with the computation of the Spearman correlation coefficient between every pair of RFs. The pair of RFs displaying the largest increases in DICE (from the perturbation analysis) and which also display a highly significant Spearman correlation were declared the RFs driving primed states.

Identification of primed cells from scRNA-Seq data

Finally, we identify the multi-or-pluripotent cells defining the primed states as those that co-express the RFs responsible for the priming. This definition is sensible as long as the RFs are only expressed in a relatively low fraction of the multi-or-pluripotent cells. We note that if necessary this requirement of a low frequency of expression in the multi-or-pluripotent cells is imposed when selecting the RFs prior to DICE computation.

Dimensionality reduction of scRNA-Seq data and scATAC-Seq data

Fetal liver hematopoiesis

For scRNA-Seq data from Ranzoni et al., we applied *RunUMAP* function from *Seurat* on the expression data of primed-MEMP and primed-GP marker genes across MEMPs, GPs, primed-MEMP and primed-GP. For scATAC-Seq data from Ranzoni et al., we did dimensionality reduction for MEMPs, GPs, primed-MEMP and primed-GP in two different ways: 1) We found all the peaks within the region of 3 kb upstream and the gene body of each marker gene for primed-MEMP and primed-GP, binarized the peak data, turned the peak matrix into a TF-IDF matrix by *RunTFIDF* function from *Signac*,[55] did SVD, used *harmony*[56] on PC2-50 to reduce effects of lanes, samples and organs, and applied *RunUMAP* on the 1–50 harmony spaces. 2) We selected 50% of peaks with highest total accessible counts across MPPs, GPs and MEMPs, binarized the data for these peaks, applied *RunTFIDF*, did SVD, followed by harmony on PC2-50, and applied *RunUMAP* on 1–20 harmony spaces.

Hair follicle generation

For scRNA-Seq data from Ma et al., we selected genes specifically highly expressed in IRS and hair shaft lineages respectively by doing Wilcoxon rank sum tests between IRS, hair shaft cells and TACs, and constructed a diffusion map on expression of these genes across TACs, IRS and hair shaft with package *destiny*.[40] For the scATAC-Seq data, we applied *cisTopic*[41] on the peak data for the above 3 cell types, using functions *runWarpLDAModels* (alpha = 20, topics = 15) and *runUmap*.

Cell fate probability calculation

Fetal liver hematopoiesis

To compute the cell fate probabilities, we used the Palantir algorithm.[37] We performed PCA on the scRNA-Seq data of MPPs, GPs, MEMPs from Ranzoni et al., subsequently applying Palantir on the top 56 PCs. The number 56 was selected by using the RMT function of the isva R-package.[57] Diffusion map coordinates and pseudotime were calculated with *run_diffusion_maps* and *run_palantir* functions. The *run_palantir* function was run 100 times with parameter *num_waypoints* set to 400 to get a more stable estimate of the cell-fate probabilities (i.e the probabilities to differentiate into the GP and MEMP lineages). The average cell fate probabilities toward one lineage of 100 Palantir runs for each MPP was used to compare the probabilities between different subpopulations.the cell fate probabilities for MPPs, we used the Palantir algorithm.[37] After doing PCA, we applied Palantir on 1–56 PCs of the log-normalized scRNA-seq data for MPPs. The number 56 was selected by using RMT as implemented in isva R-package[57] to determine the significant number of PCs. Diffusion map coordinates and pseudotime were calculated with *run_diffusion_maps* and *run_palantir* functions. The *run_palantir* function was run 100 times with parameter *num_waypoints* set to 400 to get a more stable estimate of the cell-fate probabilities. The final cell fate probabilities for each MPP were obtained by averaging the probabilities over the 100 runs.

Hair follicle generation (scRNA-Seq data)

We constructed a diffusion map from the normalized expression data encompassing all up-regulated genes for IRS and HS lineages, selected as described previously. We then applied Palantir on the top diffusion components to predict IRS and HS differentiation probabilities for each of the TAC cells, setting the *num_waypoints* parameter to 400. The final cell fate probabilities for each TAC was obtained by averaging over 50 separate Palantir runs.

Hair follicle generation (scATAC-Seq data)

We applied *cis-Topic*[41] on the scATAC-Seq data of TACs, IRS, medulla and hair shaft cortex from Ma et al. Palantir was subsequently applied to the latent representation of the cells defined over 15 regulatory topics. The final cell fate probabilities toward for each TAC was obtained by averaging over 50 separate Palantir runs, with *num_waypoints* set to 400.

Identification and validation of primed cells in scATAC-seq data

Fetal liver hematopoiesis

Having identified the primed-MEMP and primed-GP cell populations from the scRNA-Seq data, we aimed to identify the corresponding primed states in the scATAC-Seq data. To this end, we selected marker genes for primed-MEMP and primed-GP cells by performing Wilcoxon rank sum tests comparing expression of primed MPPs of the given lineage to all other MPPs. We selected marker genes displaying significantly increased expression (BH adjusted $p < 0.05$) in either primed-MEMP or primed-GP cells. Markers genes significantly overexpressed in both primed-MEMP and primed-GP cells were removed. Having thus identified specific marker genes for the primed-MEMP and primed-GP cell populations, we next turned to the scATAC-Seq peak data. We first generated a gene accessibility count matrix by aggregating peak counts within a 3 kb region upstream of each gene including the gene-body. We log-normalized this gene-cell accessibility matrix with the *NormalizeData* function from Seurat package. For each MPP cell (annotation of cells with scATAC-Seq profiles provided by authors), we next calculated the average of the normalized peak counts overall selected marker genes for a given lineage (primed-MEMP or primed-GP). This allows us to rank MPP cells according to an overall accessibility score over the primed-MEMP marker genes, and separately another ranking derived over the primed-GP marker genes. MPPs in the top-10% of these ranked lists were declared putative primed-MEMP and primed-GP cells, respectively. MPP cells appearing in both of the top-10% ranked lists were not specific and thus were excluded and categorized as "unprimed" alongside all other MPPs. To validate these assignments as well as to validate the role of the identified RFs in defining these primed states, we applied *chromVar*[39] on the scATAC-seq data to derive transcription factor regulatory activity (TFA) values for each of the MPP cells. In the case of priming into the MEMP-lineage, this was done for GATA1 and KLF1, the two TFs identified from the scRNA-Seq data as driving the primed states. In the case of the GP-lineage, this was only done for IRF8, since the other implicated RF (UHRF1) is not a TF. For the validation, we then performed one-sided Wilcoxon rank sum tests comparing the average TFA of GATA1 and KLF1 (or the TFA of IRF8), across the different subpopulations of MPPs (primed MEMPs, primed GPs and unprimed) as identified earlier from the scATAC-Seq data. To check the robustness of our findings to the choice of top-10% threshold considered earlier when defining primed subpopulations, we supplemented the validation analysis by performing Spearman correlation analysis between the TFA values and the marker accessibility scores obtained earlier.

Hair-follicle generation

Since the scATAC-Seq data from Ma et al. was generated with SHARE-Seq, this data is matched to scRNA-Seq profiles, hence the primed cells can be identified by direct matching of barcodes between the joint profiles. To validate the RFs driving the primed states in the scATAC-Seq data, we first generated a gene level accessible matrix by aggregating peak counts over each RF, and summed the gene level accessible counts for the RFs, comparing these values between primed and unprimed cells. Peaks were matched to the nearest gene if the peaks were located within a 3 kb region upstream of a gene and the gene body by using Bioconductor package EnsDb.Mmusculus.v79 [55]. We also validated the priming-status of the RFs by calculating for each RF/TF with available target information, an activity score with chromVAR. This was done by running function *RunChromVAR* from *Seurat* on the peak level data, setting the genome to BSgenome.Mmusculus.UCSC.mm10. The TF motif information derived from JASPAR2020.[51]

Identification of primed states using variance in hair-follicle regeneration data

To benchmark DICE on the hair-follicle regeneration data, we compared it to a similar strategy that only uses variability in RF expression across the multipotent cells to select candidate priming RFs. Thus, the DE-analyses were performed exactly as with DICE to identify initial pools of candidate priming IRS and HS factors. However, one difference with DICE is that instead of selecting type-1 RFs (i.e. those exhibiting low frequency of expression among the TACs), we considered the full pool of selected RFs from the DE-analyses. This is because setting a threshold on the frequency of expression across TACs would automatically favor less variable RFs, which would be counter to using variance as a criterion for selection. Thus, as

priming RFs for a given lineage, we selected from the pool passing the DE-analyses, the two RFs displaying the highest variance across TACs. To find the corresponding primed cells, we first selected cells co-expressing the two RFs. Then, for each RF we calculated the variance across TACs after removing one candidate primed cell, subsequently computing the difference between the new and original variances. We then ranked the cells in descending order of the contribution to the variance. The average over the ranks from the two priming RFs was then taken as the final rank. Finally, we declared the top n cells to be primed for the given lineage, where n was matched to the number of primed cells from the DICE analysis. Using the SHARE-Seq data (joint expression and chromatin accessibility profiles), we could then define primed cells in the scATAC-Seq dataset. The gene level accessibility for each RF was obtained by aggregating the accessible peak counts within a 3 kb range upstream of the gene and gene body. The average gene accessibility of the two priming RFs was compared between primed-IRS, primed-HS and unprimed TACs with one-tailed Wilcoxon rank sum tests. For those RFs with TF-motif information, we also compared the TF activity obtained with chromVAR between primed-IRS, primed-HS and unprimed TACs with one-tailed Wilcoxon rank sum tests.

Identification of cell-fate transition points from differentiation time course scRNA-Seq data

Having identified the interesting sets of TFs, we then compute DICE over these TFs and for each timepoint. Alternatively, if diffusion map analysis[35] reveals substantial asynchrony of cells between collected timepoints, we use pseudotime bins instead, with the number of bins matched to the number of timepoints, and with the same number of cells per pseudotime bin. This procedure generates a profile for how DICE changes with developmental timepoint/pseudotime bin. Abrupt changes in DICE between successive timepoints/bins indicate potential cell-fate transition points. To quantify the uncertainty in the measured DICE values, we use a bootstrapping approach where we resample with replacement an equal number of cells from each timepoint/bin (n = 100 bootstraps). From these bootstraps we compute an SD and derive a 95% confidence interval.

BioTIP analysis

BioTIP is an algorithm designed to detect tipping points from gene-expression data.[19] It works by inferring dynamic network biomarker (DNB) modules, computing a score for each of these modules in each differentiation stage and then identifying those that display transition-like behavior. We applied BioTIP to the log-normalized gene expression data of the two timecourse scRNA-Seq dataset (liver and retina differentiation). Each data matrix was split into submatrices defined by physical timepoints or pseudotime bins representing differentiation stages. Default parameter values were used to construct a graphical representation of genes of interest based on Pearson correlation matric and identify clusters defining dynamic network biomarker (DNB) modules. The maximum statistical score among the biomodules of each differentiation stage is quantified as DNB score. BioTIP returns a criticality index (Ic.shrink) to allow identification of the cell-fate transition.

Palantir-only analysis

Fetal liver hematopoiesis

We applied Palantir as described before to compute cell-fate probabilities for all MPP cells, and used a probability threshold of 0.6 to assign MPP cells to primed/unprimed states. With the MPP cells divided up into primed-GP, primed-MEMP and unprimed categories, we then used Wilcoxon rank sum tests to identify RFs overexpressed in the primed states relative to unprimed cells and the primed ones from the competing lineage. This was done using a BH-adjusted p value < 0.05. Among the selected RFs, the top 2 RFs with largest logFC were selected as the priming RFs for each lineage. To define the primed cells in scATAC-Seq data from Ranzoni et al., we identified marker genes displaying significantly increased expression (BH adjusted $p < 0.05$) in either primed-MEMP or primed-GP cells, and computed accessibility scores over these marker-genes for each MPP and each lineage, as before. MPPs in the top-10% of these ranked lists were declared putative primed-MEMP and primed-GP cells for the Palantir-only method, respectively. MPP cells appearing in both of the top-10% ranked lists were excluded from the primed-MEMPs and primed-GPs. Finally, we used chromVAR to compute transcription factor regulatory activity (TFA) values for the selected RFs with PWM information in each of the MPP cells (GATA1 for MEMPs and XBP1 for GPs, since the other priming RFs ZBTB16 for MEMPs and UHRF1 for GPs lacked the PWM information). TFA-values between the primed and unprimed cell categories were compared using one-sided Wilcoxon rank sum tests.

Hair-follicle generation

We used the same procedure for the hair-follicle dataset except that for this set the Palantir cell-fate probabilities for the TACs displayed more extreme values. Hence, we declared TACs with probabilities toward IRS greater than 0.97 to be "primed-IRS" cells whilst TACs with probabilities toward HS greater than 0.9 to be "primed-HS cells". Since the scATAC-Seq data from Ma et al. was generated with SHARE-Seq, this data is matched to scRNA-Seq profiles, hence the primed cells in the scATAC-Seq data can be identified by direct matching of barcodes between the joint profiles. To validate the RFs driving the primed states in the scATAC-Seq data, we first generated a gene level accessible matrix by aggregating peak counts over each RF, and summed the gene level accessible counts for the RFs, comparing these values between primed and unprimed cells. We also used chromVAR to derive TFA values for the RFs with PWM info, and to then compare the TFAs between the primed and unprimed cell categories. For priming to IRS and HS, TFA was calculated for Gata3 and Lef1, respectively, whilst for the other priming RFs (Maml3 for IRS and Trps1 for HS) no PWM info was available. We also compared the cell fate probabilities between different subgroups of TACs with the previously obtained Palantir results for scATAC-Seq data from Ma dataset, as described in the cell fate probability calculation section.

Description of GRNs and simulation of scRNA-Seq from GRNs

Simple 4-gene GRN describing transition between pluripotent and differentiated cells

We considered a simple four-gene regulatory network model derived from experimentally observed gene-gene regulatory interactions in pluripotent and differentiated cells.[58] This model was chosen to illustrate the importance of considering the distance correlation *dCor* measure, since in this GRN model, pluripotency is associated with oscillatory (and hence non-monotonic) dynamics of the TFs. In detail, the GRN consists of two pluripotent genes (Nanog (G_1) and Oct4 (G_2)) and two differentiation genes (Gata6 (G_3) and Gata4 (G_4)), with seven mutually activating and inhibitory interactions. During differentiation, the expression of pluripotent genes gradually decreases whereas expression of differentiation genes increases. Cellular states are described by the combined expression pattern of the four genes, G_1, G_2, G_3 and G_4. The gene expression dynamics in single cells is described by the four coupled Langevin equations.

$$\frac{dG_1}{dt} = -G_1 + \frac{\left(\frac{G_1}{K_{11}}\right)^n}{1+\left(\frac{G_1}{K_{11}}\right)^n} \cdot \frac{1}{1+\left(\frac{G_3}{K_{13}}\right)^n} + \eta_1\sqrt{G_1} \qquad \text{(Equation 1)}$$

$$\frac{dG_2}{dt} = -G_2 + \frac{\left(\frac{G_1}{K_{21}}\right)^n}{1+\left(\frac{G_1}{K_{21}}\right)^n} + \eta_2\sqrt{G_2} \qquad \text{(Equation 2)}$$

$$\frac{dG_3}{dt} = -G_3 + \frac{1}{1+\left(\frac{G_1}{K_{31}}\right)^n} \cdot \frac{1}{1+\left(\frac{G_4}{K_{34}}\right)^n} + \eta_3\sqrt{G_3} \qquad \text{(Equation 3)}$$

$$\frac{dG_4}{dt} = -G_4 + \frac{\left(\frac{G_3}{K_{43}}\right)^n}{1+\left(\frac{G_3}{K_{43}}\right)^n} \cdot \frac{1}{1+\left(\frac{G_2}{K_{42}}\right)^n} + \eta_4\sqrt{G_4} \qquad \text{(Equation 4)}$$

The first term in the right-hand side of the above equations represent degradation of gene G_i and rest of the terms describe different types of regulatory interactions (activation/inhibition function). The expression of pluripotent gene G_1 is controlled by the interaction parameters K_{11} and K_{13} that are crucial for defining the cellular state. In the above model, the pluripotent state can be characterized by oscillatory expression dynamics of the four genes, and such pluripotency can be lost due to a decrease in the ability of the expressed genes to uphold the oscillatory dynamics. By changing the interaction parameter values, the gene expression dynamics switches between fixed-point (differentiated) and oscillatory (pluripotency) states. We opted for a parameter set that allows for the oscillatory state in which all expression levels show temporal cycles. In the above equations, $\eta_i (i = 1, \ldots, 4)$ is a Gaussian white noise term with noise strength ($\sigma = 0.01$) to represent stochastic gene expression. All parameter values are summarized in Table S1, and are identical to those used previously.[58] We used the *de-Solve* R-package[54] to numerically simulate the above differential equations for 1000 cells, each with different initial values of gene expression,

randomly sampled from a uniform distribution $U(0.1,1)$. Final expression values of the four genes/TFs for each of the 1000 cells were obtained in the long-time limit (i.e. steady state). This resulted in a 4 TF x 1000 cell expression data matrix, from which we then estimated the Distance Correlation *dCor* and Pearson Correlation (PCC) for the 6 TF-TF pairs.

GRN describing two mutually repressing TFs with auto-activation (GATA1-PU.1)

A well-known dynamical system is the one describing the binary cell-fate decision of bipotent myeloid progenitors into either erythroid or myeloid lineages, and which is controlled by the antagonistic pair of TFs PU1 and GATA1.[4,59] The mathematical equations that describe the dynamics are:

$$\frac{dx}{dt} = -kx + \frac{ax^n}{\theta^n + x^n} + \frac{b\theta^n}{\theta^n + y^n} \quad \text{(Equation 5)}$$

$$\frac{dy}{dt} = -ky + \frac{ay^n}{\theta^n + y^n} + \frac{b\theta^n}{\theta^n + x^n} \quad \text{(Equation 6)}$$

Depending on the value of the auto-activation and decay rate constants, this GRN can display two distinct types of pitchfork bifurcations, known as supercritical and subcritical. For the numerical simulation of single-cell dynamics, we adopted the exact stochastic simulation algorithm (SSA) formulated by Gillespie.[60] Used parameter values are shown in Table S2 and were taken directly from.[4,59] These parameter values were used for the SSA simulation (using *GillespieSSA* R-package[53]) with the expression values of the two transcription factors (TFs) obtained in the long time limit. Of note, we applied a scale factor (sf = 100) to convert the arbitrary units (A.U.) into molecular units, since the Gillespie simulation is run at the level of reaction molecule numbers. The SSA simulation was run for each individual cell, where for each cell we used a unique initial value of TF expression, a generated random integer within the range between 20 and 80 reactant molecules. Finally, we scaled the simulated TFs expression value by the scale factor for downstream analysis. In these analyses, we considered the mutual repression rate constant (b) as the bifurcation parameter whose value ranged from 0.05 to 1.25.

GRN describing T cell development

The GRN circuit consists of 4 TF (TGF1, PU1, GATA3 and BLC11B) that together with external Notch signaling determines the stage of T cell development.[32] The ODEs describing the system are given below

$$\frac{dF}{dt} = -fF + k_{0,F} + k_{N,F}N + K_F \frac{\left(\frac{F}{K_{F,F}}\right)^{n_{F,F}}}{1+\left(\frac{F}{K_{F,F}}\right)^{n_{F,F}}} \cdot \frac{\left(\frac{G}{K_{G,F}}\right)^{n_{G,F}}}{1+\left(\frac{G}{K_{G,F}}\right)^{n_{G,F}}} \cdot \frac{1}{1+\left(\frac{P}{K_{P,F}}\right)^{n_{P,F}}} \quad \text{(Equation 7)}$$

$$\frac{dP}{dt} = -fP + k_{0,P} + K_P \frac{\left(\frac{P}{K_{P,P}}\right)^{n_{P,P}}}{1+\left(\frac{P}{K_{P,P}}\right)^{n_{P,P}}} \cdot \frac{1}{1+\left(\frac{G}{K_{G,P}}\right)^{n_{G,P}}} \cdot \frac{1}{1+\left(\frac{B}{K_{B,P}}\right)^{n_{B,P}}} \cdot \frac{1}{1+\left(\frac{F}{K_{F,P}}\right)^{n_{F,P}}} \quad \text{(Equation 8)}$$

$$\frac{dG}{dt} = -fG + k_{0,G} + k_{N,G}N + K_G \frac{\left(\frac{F}{K_{F,G}}\right)^{n_{F,G}}}{1+\left(\frac{F}{K_{F,G}}\right)^{n_{F,G}}} \cdot \frac{1}{1+\left(\frac{P}{K_{P,G}}\right)^{n_{P,G}}} \quad \text{(Equation 9)}$$

$$\frac{dB}{dt} = -fB + k_{0,B} + k_{N,B}N + K_B \frac{\left(\frac{F}{K_{F,B}}\right)^{n_{F,B}}}{1+\left(\frac{F}{K_{F,B}}\right)^{n_{F,B}}} \cdot \frac{\left(\frac{G}{K_{G,B}}\right)^{n_{G,B}}}{1+\left(\frac{G}{K_{G,B}}\right)^{n_{G,B}}} \quad \text{(Equation 10)}$$

where F, P, G and B stand for TGF1, PU1, GATA3 and BLC11B concentrations, respectively. See Table S3 for specific parameter values used. In the above, N denotes the Notch signaling, which varies between 0.05 and 0.5. For low Notch signaling, i.e. when N is less than the critical value 0.25, there are four distinct states (high potency regime). All four states can converge into a single-attractor system (low potency) at high Notch signal (i.e. when $N > 0.25$). To generate the expression patterns of the four TFs across single-cells, we applied the SSA simulation scheme (using GillespieSSA R-package[53]) quantifying the expression values of the four TFs in the long time limit. We converted A.U. into molecular unit by a scale factor (sf = 50) for Gillespie simulation. For each cell, the SSA simulation was initiated with unique initial values of TF expression, obtained by sampling random integer within the range between 5 and 175 reactant molecules. We

scaled the simulated TFs expression values by the scale factor for downstream analysis. Of note, when computing DICE and the ordinary PCC-based entropy, only 3 of the 4 TF were used (TCF1, GATA3 and BLC11B) because only these 3 represent differentiation factors (their expression goes up with increased commitment).

52-Gene GRN describing transition to induced pluripotency

Finally, we decided to consider a more complex GRN composed of 52 genes, that has been proposed to model the transition of somatic cells to induced pluripotency.[33,34] Among the 52 genes, 11 mark the pluripotent stem cell state, another 11 mark the differentiation state and the rest (30 genes) are regulated by these 22 marker genes. The network consists of 183 regulatory interactions (84 activating and 39 inhibitory). The stem cell state is characterized by high NANOG & low GATA6 levels, whilst the differentiation state is defined by low NANOG & high GATA6. The ODEs take the form

$$\frac{dx_i}{dt} = -fx_i + \sum_{j=1} A_{ij}\frac{a * x_j^n}{S^n + x_j^n} + \sum_{j=1} B_{ij}\frac{a * S^n}{S^n + x_j^n} \qquad \text{(Equation 11)}$$

where A_{ij} and B_{ij} are the network adjacency matrices for activating and inhibitory interactions, respectively, and where $i = 1, \ldots, 52$. The first term on the right-hand side of the ODE represents degradation of gene x_i, whilst the second and third terms describe positive and negative regulatory interactions, respectively. Table S4 lists the parameter values that were used for the SSA simulation (using GillespieSSA R-package) and where expression values of the 52 genes were extracted in the long time limit. In this analysis, regulatory interaction (activation/inhibition) parameter a ranges from 0.25 to 0.5 and is the parameter driving cell-fate decision with $a = 0.35$ representing the critical value. We converted A.U. into molecular unit by a scale factor (sf = 100) for Gillespie simulation. Individual cells were run with a unique initial value of gene expression, generated as a random integer within the range between 50 and 120 reactant molecules. We ran the SSA simulation for 1000 cells and evaluated steady state expression of 52 genes, followed by scaling for downstream analysis.