Title page

Reinforcement learning guides single-cell sequencing in decoding lineage and cell fate decisions

Zeyu Fu[1#], Song Wang[1#], Kangfu Sun[1], Baichuan Xu[1], Xianpeng Ye[1], Zhaoyang Wen[1], Mingqiang Shen[1], Mo Chen[1], Fang Chen[1], Yang Xu[1], Shilei Chen[1]*, Junping Wang[1]*

[1]State Key Laboratory of Trauma and Chemical Poisoning, Institute of Combined Injury, Chongqing Engineering Research Center for Nanomedicine, College of Preventive Medicine, Third Military Medical University, Chongqing, 400038, China

*Correspondence author

Dr. & Prof. Junping Wang

Tel: +86-023-68752283; Fax: +86-02368752009

Email: wangjunping@tmmu.edu.cn

Dr. & Prof. Shilei Chen

Tel: +86-023-68752154; Fax: +86-02368752154

Email: chen.shilei@foxmail.com

[#]: Z. F. and S. W. contributed equally to this work.

Running title: Single-cell Reinforcement Learning Decodes Cell Fate Decisions

Abstract

The rapid development of single-cell sequencing provides unparalleled opportunity for delineating the heterogeneous characteristics of individual cells, while current methodologies fall short in accurately deciphering the origins of cellular heterogeneity and pinpointing the states of cell fate decisions. In this study, we introduce a novel approach termed Single-cell Reinforcement Learning (scRL) by integrating a framework of reinforcement learning into single-cell data analysis with the assistance of an actor-critic architecture. scRL reveals that the intensity of fate decisions precedes the maximum potential of lineage differentiation and actual gene expression, which enables the forecasting and perturbation of cellular fate in an earlier stage. The superiority of scRL in assessing fate decisions is confirmed in several different patterns of single-cell datasets. By using scRL, we disclose a distinct role of *Dapp1* in myeloid lineage decision during hematopoietic stem cells (HSCs) differentiation and uncover the priority of erythroid fate decision occurring post irradiation exposure.

Introduction

Single-cell sequencing has emerged as a transformative technology for the exploration of cellular heterogeneity at an unprecedented resolution of individual cells[1-3]. A key advancement related to single-cell sequencing is pseudotime analysis, as it can unravel the sequential order of cellular events and the transition states in a virtual timeline. Cell state transitions manifest as branching differentiation trajectories in the two-dimensional reduced space of the data[4,5]. Embedding techniques such as PCA,

t-SNE, and UMAP have proven exceptionally useful for single-cell data analysis[6-8]. Particularly, UMAP can adeptly preserve both local and global structures, so that advanced applications like PAGA, Monocle3, Slingshot, and CAPITAL leverage UMAP to enhance representations of the differentiation manifold for trajectory inference[9-12].

Despite widely served for revealing the differentiation trajectories and temporal ordering, the current methodologies fall short in addressing where and when cell fate decisions occur during differentiation, as well as in providing a framework to evaluate these decisions. In fact, cellular differentiation is not merely a linear pathway but a complex landscape where progenitors become lineage-restricted at various points and times influenced by a myriad of factors[13]. This insight challenges the traditional hierarchical model of differentiation and highlights the importance of identifying specific decision points where cells commit to a particular lineage or fate. Given that single-cell analysis has highlighted the dynamic nature of the differentiation process, it should consider both spatial and temporal dimensions when studying cell fate decisions.

Reinforcement Learning (RL) is a branch of machine learning with distinct characteristic that an employed agent learns to make decisions by interacting with the environments. Normally, the agent seeks to achieve a goal by performing actions and receiving feedback in the form of rewards or penalties. In recent years, RL has excelled in addressing sequential decision-making challenges and been applied in multiple areas such as robot control, gaming intelligence, and navigation[14-16]. The

power of reinforcement learning mainly lies in its adaptability and its ability to iteratively improve its understanding of the process of a certain event. Given that the process of cellular differentiation within the single-cell landscape is inherently sequential and decision-centric, it is achievable to utilize RL for single-cell sequencing analysis of cellular differentiation. By leveraging its ability to explore new paths and exploit known pathways efficiently, RL is expected to pinpoint the key differentiation nodes and pathways, so as to facilitate a deep decoding of cell fate decisions.

In this study, we introduce a novel approach termed Single-cell Reinforcement Learning (scRL) by integrating reinforcement learning into single-cell data analysis to resolve the challenges in decoding lineage and cell fate decisions. Briefly, scRL utilizes a grid world created from a UMAP two-dimensional embedding of high-dimensional data, followed by an actor-critic architecture to optimize differentiation strategies and assess fate decision strengths. The effectiveness of scRL is demonstrated through its ability to closely align pseudotime with distance trends in the two-dimensional manifold and to correlate lineage potential with pseudotime trends. By being applied to the single-cell datasets of acute myeloid leukemia (AML), human hematopoiesis and mouse endocrinogenesis, scRL is shown to have prominent capability to accurately pinpoint fate decision states and pre-expression states in complex multilineage systems. The practicality and superiority of scRL are further confirmed with forecasting and dynamic gene perturbation of lineage and cell fate decisions during hematopoietic stem cells (HSCs) differentiation under

physiopathological conditions.

Results

The architecture of scRL

As shown in Figure 1, the process of scRL begins by constructing a grid embedding with a two-dimensional reduction space of the cell data managed by t-SNE, PCA, or UMAP, to put up a structured environment for navigating cell differentiation. In the implementation, edge detection algorithms identify boundary points on this grid, which help in defining a fully connected graph structure, so as to ensure connectivity across the grid from the most primitive to the most mature cell states. For the reinforcement learning setup, scRL utilizes an actor-critic architecture[17] working in the grid environment filled with predefined reward signals. These rewards are configured to either decay or increase exponentially with calculated pseudotime, reflecting the maturity of cell states. The value network evaluates the potential returns from each state, while the policy network decides on the next action, guiding the agent through the grid to explore different cell fates.

The core functionalities of scRL are divided into three modules: gene, lineage, and trajectory. The gene module uses gene expression data to train networks that predict gene decision strength and potential. The lineage module focuses on subgroup classification to determine fate decision strength and contribution strength of cell differentiation. The trajectory module samples and analyzes paths through the grid, representing differentiation trajectories.

Finally, an autoencoder[18] is trained to predict changes in cell features like gene expression and pseudo-time, based on the differentiation trajectories identified by the policy network. This setup allows scRL to effectively map and predict cell differentiation processes, providing insights into the dynamics of cell fate decisions without the need to delve deeply into the technical specifics of the method.

Pseudotime and lineage potential functions of scRL

First, by employing a tabular Q-learning approach as an initial validation of the reinforcement learning environment, human hematopoietic $CD34^+$ cell dataset with well-separated differentiation branches and mouse endocrinogenesis dataset with a stage-separated process underwent scRL's grid embedding and pseudotime alignment functions within a UMAP 2-D space. The cluster information was subsequently projected onto the grid space (Figure S1A). Meanwhile, specific rewards were generated for erythroid, myeloid, and lymphoid lineages within the grid embedding in the hematopoiesis dataset (Figure S1B). For endocrinogenesis dataset, the rewards were set for early and late differentiation stages (Figure S1B). Starting points were designated at HSC clusters for the hematopoiesis dataset and EP clusters for the endocrinogenesis dataset, with rewards setting at myeloid, erythroid, lymphoid, Fev^+ cells, and mature cell clusters. Following the implementation of tabular Q-learning, we obtained and projected the final state values onto the original embedding (Figure S1C). The results demonstrated that the training process was effective, evidenced by the convergence of both the cumulative reward obtained by the policy network and the maximum state value estimated by the critic network over time (Figure S1D-E).

To evaluate the decoding capability of lineage decision states, we specifically used a dataset of human hematopoietic progenitor cells and a dataset of human $CD34^+$ bone marrow cells (Figure 2A, Figure S2A). After aligning the pseudotime, we first conducted a comparative analysis between scRL and other prevalent methods such as Palantir, Wishbone, DPT, Monocle2 and Monocle3, based on the correlation coefficient with the Euclidean distance from the starting point. Notably, scRL was shown to achieve the highest score (Figure 2B-C). Furthermore, by employing scRL's contribution reward mode, we assessed the commitment potentials of the primary MEP and GMP lineages properly (Figure 2D-F). A similar trend was found in erythroid and myeloid lineages of $CD34^+$ bone marrow cells (Figure S2B-F).

To evaluate the lineage decision state values, we performed an analysis on MEP and GMP lineages. The lineage decision values along pseudotime suggested that the decision values for MEP may peak during the intermediate stage of differentiation. Additionally, the contribution values reached their maximum at the most mature stage of differentiation. By comparing the weighted pseudotime for lineage decisions and the contribution strengths of MEP, we observed that the pseudotime weighted by decision value preceded the pseudotime weighted by contribution value (Figure 2G-J). The trend is similar in GMP lineage and lineages of $CD34^+$ bone marrow cells (Figure 2K-N, Figure S2G-N). These results indicated that the lineage decision values inferred by scRL represented a state prior to actual lineage commitment, highlighting the predictive capability of scRL in decoding lineage decisions. This finding underscored the utility of scRL in anticipating developmental outcomes in

hematopoietic lineage commitment.

Simulation function of scRL

HSCs had the capability to differentiate into restricted progenitors including the erythroid and myeloid lineage[19]. To simulate the differentiation of primary lineages in HSCs, we selected the HSC-MEP and HSC-GMP lineages from human phenotypic hematopoietic stem and progenitor cell (HSPC) dataset for differentiation simulation (Figure 3A, Figure S3A). During the exploratory training phase, an autoencoder was initialized at various starting points and progressed through 50 steps, inputting a single step and outputting the state difference. The simulation results demonstrated a correlation between the predicted pseudotime and branch information, revealing a gradual increase in the proportion of MEP and GMP lineages as the simulation advanced (Figure 3B, Figure S3B).

To validate the accuracy of the simulated paths, we compared the state space of the path points with the initial HSC targeting MEP and GMP lineages in the phenotypic HSC dataset. In the original space, the correlation with HSC decreased over pseudotime, while the correlation with MEP increased (Figure 3C). Conversely, in the simulated space, the correlation coefficient between path points and HSC state spaces steadily declined as the paths progressed, while the correlation with MEP increased (Figure 3D). Cluster heatmaps revealed that the correlation scores of HSC with MEP were higher in the simulated space than in the original space (Figure 3E, F). Regarding gene expression levels, we observed an increase in correlation with MEP associated genes (Figure 3G), while the correlation with HSC-associated genes

decreased as pseudotime progressed. The trend was similarly observed in the GMP lineage (Figure S3C-G). In the simulated gene expression space, the patterns mirrored those observed in the simulated lineage cells (Figure 3H-J, Figure S3H-J). We ranked the simulated genes based on their time correlation within the MEP lineage (Figure 3K), and found that these genes exhibited lower time correlations in the original expression space for HSC-MEP (Figure 3L). Moreover, the trends in simulated gene expression were more distinct for the HSC and MEP, compared to those observed in the original expression space (Figure 3M-N) . The results were consistent in the GMP lineage (Figure S3K-N). In summary, these dynamic trends of correlation with HSC and various lineages illustrated the distinctive switching characteristics of fate decision in differentiation. These findings highlighted the complexity of gene expression dynamics and underscored the importance of simulation in understanding the nuanced changes across different hematopoietic lineages.

Gene decision function of scRL

scRL introduced a novel functionality in gene expression-based reward systems by enabling the identification of a pre-expression state for genes. Here, we utilized the datasets derived from human hematopoiesis and mouse endocrinogenesis to evaluate the effectiveness and convergence of scRL for gene expression analysis. We specifically selected marker genes indicative of distinct lineages: *GATA1* for the erythroid lineage[20], *IRF8* for the myeloid lineage[21], and *EBF1* for the lymphoid lineage[22]. Additionally, we chose *Ngn3* as a marker for the early phases of endocrinogenesis and *Fev* to represent its intermediate stages[23,24] (Figure S4A). By

employing these gene-specific markers, we generated rewards for grid embedding during the process of scRL (Figure S4B).

Leveraging the tabular Q-learning algorithm, we explored the environment using an epsilon-greedy strategy to maximize gene rewards, while utilizing the Q-table to assess state values. The outputs were subsequently mapped back to the original embedding space (Figure S4C). The meaningfulness of the training procedures was evidenced by the convergence of both episode rewards and peak state values over time (Figure S4D, S4E). This convergence underscored the efficacy and robustness of our gene expression-driven reward system within the scRL framework, demonstrating its potential to enhance understanding of gene regulation dynamics in complex biological systems.

GATA1 is a crucial transcription factor in MEP differentiation, while *CEBPA* is involved in the differentiation of the GMP lineage. To validate the pre-expression states predicted by scRL, we first observed that *GATA1* expression peaked at the mature stage in MEP lineage (Figure 4A-B), and CEBPA expression in the GMP lineage exhibited a similar pattern (Figure 4C-D). Following the scRL fate determination analysis, both *GATA1* and *CEBPA* displayed more primitive regions of high decision value intensity in the UMAP two-dimensional space compared to the contribution value (Figure 4E-H).

As for *GATA1* in the erythroid lineage, the decision state value was shown to peak at the medium stage of differentiation, preceded the expression peak (Figure 4I-J). The contribution state value for *GATA1* peaked at the mature stage of differentiation and

demonstrated a stronger correlation with pseudotime in the MEP lineage than the original expression (Figure 4K-L). Similarly, the decision state value and contribution state value of CEBPA also showed a similar trend in HSC-GMP lineage (Figure 4M-P). These results suggested that the analysis of decision and contribution values provided a nuanced view of gene regulation during cell differentiation, offering insights into the timing and regulation of gene expression critical for effective lineage specification.

The lineage and gene decision state value of scRL decode a common differentiation stage

To perform a detailed analysis of both gene and lineage decision state values, pseudotime analysis of scRL was conducted for a AML dataset[25]. We first identified four distinct branches representing lineage decision values (Figure 5A-C). Then, we selected 10 marker genes as indicators of gene decision values for each of these branches, followed by gene decision analysis on these branch-specific genes and lineage decision analysis on four identified branches (Figure 5D-F). Notably, a strong correlation between lineage and gene decision values was revealed within each branch (Figure 5G). Additionally, it was also observed that the weighted average pseudotime of lineage decision values consistently preceded that of gene expression across all four branches (Figure 5H).

Examination of the lineage decision values across the four branches, when mapped along binned pseudotime, revealed more distinct trajectory dynamics compared to the mean Pearson correlation values with cells specific to each branch (Figure 5I-J).

Similarly, the analysis of gene decision values along binned pseudotime provided a clearer depiction of the dynamics than the Pearson correlation values with branch-specific genes (Figure 5K-L). This enhanced clarity in the data facilitated a deeper understanding of the developmental trajectories in the context of both lineage and gene decision processes. Our findings suggested that lineage decision values serve as early indicators of gene expression changes, providing valuable insights into the dynamics of gene regulation within the context of AML progression. And the approach underscored the utility of integrating pseudotime analysis with decision values to elucidate complex biological patterns and regulatory mechanisms in developmental processes.

scRL facilitates to discover the key regulators for lineage and cell fate decisions during HSCs differentiation

To further illustrate the versatility of scRL, we tried to use it to discover the key regulators for lineage and cell fate decisions during hematopoietic stem cells (HSCs) differentiation. We initially mapped the natural differentiation trajectory of LSK ($Lin^{-}Sca1^{+}ckit^{+}$) cells using PALANTIR diffusion and t-SNE embedding spaces. Then, we applied scRL to create a grid embedding, anchored by the primitive genes *Hlf* and *Mllt3*, and analyzed the dynamic process of cell differentiation through pseudotime and velocity vector fields. This approach enabled us to identify rapidly upregulated genes during early lineage differentiation, including *Dapp1*, *Adgrl4*, *Lmo2*, *Cbfa2t3*, *Malat1*, *Ctla2a* and *Hlf*, with *Dapp1* showing a strong temporal association (Figure 6A). Further analysis using DDRTree traced the differentiation trajectory to predict

the lineage fate decision statuses (Figure S5A). By intersecting dynamically expressed genes with those identified as significant by BEAM[26], we focused on 280 genes enriched in hematopoietic differentiation-related Gene Ontology Biological Process (GOBP) terms (Figure S5B-C). Additionally, we conducted a gene association analysis using LSK datasets with representative lineage markers, *Dach1* for myeloid lineage differentiation and *Dntt* for lymphoid lineage differentiation[27,28]. Intriguingly, *Dapp1* emerged as highly correlated with *Dach1* and exhibited a lesser correlation with *Dntt* (Figure S5D, S5E). The intersecting genes from these datasets further substantiate the close association of *Dapp1* with myeloid differentiation within the LSK population (Figure S5F).

To validate whether scRL can effectively capture the impact of perturbation in key differentiation dynamical gene on hematopoietic stem cell differentiation states, a single-cell atlas was constructed from the bone marrow LSK cells with conditional *Dapp1* knockout in the hematopoietic compartent (Figure 6B). The lineages within this atlas were annotated based on the expression patterns of key transcription factors (Figure S6). This mapping revealed a significant disruption in lineage commitment homeostasis, characterized by a marked reduction in the myeloid lineage (Figure 6C). Additionally, pseudotime and expression of marker genes *Cd34*, *Mpo* and *Ctsg* suggesting an impediment to the cellular maturation process (Figure 6D).

Utilizing the myeloid, lymphoid, and erythroid clusters as a basis for reinforcement learning, we observed shifts in lineage decision values (Figure S7). Analysis of lineage decision trajectories along pseudotime revealed that *Dapp1* deficiency led to

significant perturbations (Figure 6F). Specifically, the decision values for myeloid and erythroid lineages decreased, while those for the lymphoid lineage increase, indicating an imbalance in fate decision after *Dapp1* was conditional knockout. However, the differentiation trajectories learned were not sufficiently distinct (Figure S8). To further validate these findings, we analyzed gene decision values specific to each lineage: *Mpo* for myeloid, *Itga2b* for erythroid, and *Dntt* for lymphoid (Figure S9). The trajectories of these gene decision values underwent notable changes (Figure 6G). Moreover, *DAPP1* deletion in hematopoiesis resulted in a decrease in the decision values for the myeloid gene *Mpo* and the erythroid gene *Itga2b*, while the decision value for the lymphoid gene *Dntt* increased (Figure 6H).

Focusing on the earliest cluster (Figure S10A), we observed that both the decision trajectory for the myeloid gene *Mpo* and the myeloid lineage decision trajectory are disrupted if *Dapp1* was conditional knockout (Figure 6I). Specifically, both gene and lineage decision values for myeloid decrease after the knockout (Figure 6K). In contrast, the lymphoid lineage decision trajectory and the *Dntt* gene decision value did not decrease (Figure S10B). While the erythroid lineage and *Itga2b* gene decision values also decreased, this change was less pronounced compared to that observed in the myeloid lineage and *Mpo* (Figure S10C). In the earliest cluster, lymphoid potential remained largely unchanged. However, within the multipotent cluster (Figure S11A), both lymphoid and *Dntt* decision values are enhanced (Figure 6L), along with their decision trajectories along pseudotime (Figure S11B). Additionally, we note a decrease in *Itga2b* and erythroid decision potential (Figure S11C), as well as in *Mpo*

and myeloid decision values (Figure S11D). The application of scRL successfully addressed the challenge of assessing lineage fate decisions in single-cell data. Moreover, the integration of scRL with single-cell subpopulation differentiation provided a comparative analysis of lineage fate decisions across different differentiation levels, our results were confirmed with single-cell data from key differentiation gene perturbations.

scRL decodes the fate decision atlas of HSCs after irradiation

After validating the effectiveness of scRL, we sought to apply it to analyze the fate decision atlas of HSCs under pathological conditions. As known, hematopoietic cells are sensitive to ionizing radiation, while the cell fate transitions of HSCs after radiation injury remain largely unknown. Here, we first collected LSK cells from mice inflicted with total body ionizing radiation (TBI) at different days after radiation and performed single-cell sequencing. Then, a comprehensive fate decision atlas of HSCs post radiation exposure was constructed with scRL. Eleven biologically significant clusters were marked by critical gene expressions, consistent with previous studies[29-33] (Figure S12 A-D). The number of HSPC, especially HSCs, sharply reduced after radiation and reached the bottom around day 5, followed by stepwise recovery. Meanwhile, an erythroid commitment bias emerged after radiation, peaking at day 8, and persisted until day 21(Figure 7B).

Utilizing pseudotime as a metric for cellular primitivity, the recovery process was interpreted as HSCs regaining their primitive state (Figure 7C). As shown, HSPC components decreased during the injury phase but recovered by day 30, correlating

with the expression patterns of *Hlf*, *Meis1*, and *Satb1*[34-34] (Figure 7D, Figure S13A). However, both of the erythroid and myeloid components increased before day 8 after radiation (Figure 7E), which corresponded to the expression patterns of *Klf1* and *Gata1*, as well as the expression patterns of *Gfi1*, *Cebpa*, and *Cebpe* respectively[37-40] (Figure 7F, Figure S13B-C).

To decipher the fate decisions of erythroid and myeloid lineages following irradiation, we devised lineage-specific reward environments and computed pseudotime for each day post-irradiation (Figure S14A). By employing lineage rewards, we utilized the critic network to assess lineage decision values for both erythroid and myeloid lineages (Figure 7G-H). Additionally, we determined the lineage contribution values for both erythroid and myeloid lineages (Figure S14B, S15A-B). We then conducted a comparative analysis of erythroid (colored in red) and myeloid (colored in blue) lineage decision values across pseudotime within selected HSPC subpopulations (Figure 7I). Although the fate decision potential of erythroid-biased stem cells decreased dramatically after radiation, it rapidly returned to near normal level at day8. Conversely, the proportion of myeloid-biased stem cells increased following irradiation injury and subsequently decreased as the erythroid-biased stem cells increased (Figure 7I-J). Interestingly, the erythroid potential of HSPCs showed the strongest correlation with the overall proportion of HSPCs, whereas the myeloid potential was most closely associated with the proportion of myeloid cells during the early days after radiation (Figure 7K). These findings indicate that the surviving HSCs after radiation preferentially differentiate into erythroid lineage cells, and the

increase of myeloid potential is possibly ascribed to the radiation resistance property of myeloid progenitor cells (MPP2/3). In fact, the priority of erythroid fate decision occurs not only post irradiation exposure but also under physiological status (Figure 7K), probably due to the indispensable role of erythrocytes in the body. Therefore, integrating reinforcement learning with single-cell analysis can reveal deeper insights into HSCs fate decisions under physiopathological conditions.

Discussion

The adaptation of reinforcement learning methodologies to biological systems highlights the versatility and potential of these techniques in addressing complex challenges in biological research.[35,36]In this study, our novel framework leverages the power of reinforcement learning to model and analyze cellular differentiation pathways within the high-dimensional landscape of single-cell data. By efficiently exploring and exploiting differentiation pathways, scRL identifies key nodes and pathways that influence cell fate decisions, providing valuable insights into cellular dynamics and potential avenues for manipulating cell fate in various biological applications.

Compared to traditional methods that rely on deterministic models or predefined genetic markers, such as CEFCON[37], scRL offers a more dynamic and adaptive approach to analyzing cell fate decisions. The integration of reinforcement learning into single-cell data provides a nuanced analysis of differentiation pathways, including learning which pathways lead to optimal outcomes and identifying critical decision points within those pathways. This capability sets scRL apart from clustering

algorithms or trajectory inference techniques that may not effectively pinpoint exact decision nodes[38]. The reward mechanism in scRL, based on lineage-specific and gene-specific targets, allows for continuous learning and adjustment of predictions. However, further improvements can be made to scRL's reward modeling by integrating multi-modal data sources to enhance the granularity and accuracy of the reward signals.

In terms of trajectory inference, scRL's module can be compared to other methods such as Monocle, TSCAN, and Slingshot[39]. Although these algorithms have their strengths, they may struggle with complex trajectory structures, miss transitional states, or be sensitive to noise and parameter settings. By integrating adaptive learning rates and incorporating feedback mechanisms based on experimental validations, we achieved the goal of improving the accuracy and robustness of scRL's trajectory inference.

The gene trend module in scRL is preferable to algorithms like Wanderlust or Wishbone, which can also track the changes in gene expression but may be limited to specific types of trajectories or lineage progressions[40]. By adaptively selecting relevant genes based on their impact on trajectory decisions and integrating external datasets for validation, scRL can provide more nuanced insights and help understand the broader biological implications of these trends.

The intrinsic logic behind scRL's critic network lies in its comprehensive evaluation, including integrating lineage or gene expression reward signals with two-dimensional trajectory information constrained by the grid-embedded environment. Moreover, the

evaluation represents an expected reward estimation that decays with the gamma parameter, indicating the expected lineage or gene rewards associated with a particular state under a specific decay rate. The estimation enables the assessment of lineage decision and pre-gene expression states, which exhibit strong correlations within the same lineage, mutually reinforcing and revealing cellular fate determination. However, it should be acknowledged that scRL's state value estimation relies on the recovery of differentiation trajectories by dimensionality reduction methods, which may lead to variations in lineage fate decsion information due to the chosen technique. The definition of starting and target subpopulations of a lineage plays a pivotal role in the final estimation results, as lineage information is influenced by the nearest neighbor graph constructed from the principal component space and dimensionality reduction methods like UMAP and DIFFMAP.

In conclusion, this study presents a transformative tool in the field of computational biology, offering a novel approach to decoding the lineage and cell fate decisions based on single-cell sequencing.

Methods

Grids embedding function

The $Grids\ from\ Embedding$ function (Figure S16 Algorithm 1) processes a 2-D embedding space X to produce a grid-based representation. This function is parameterized to allow flexibility in grid generation, boundary detection, and computational efficiency through parallel processing.

Grid and spine generation

Initially, the function invokes $Generate\ Grids$ (Figure S16 Algorithm 2), which computes grid points across the embedding space. This involves determining the extremal coordinates of the space (right, left, top, bottom) and using these to create a mesh grid of points. Additionally, the function constructs spines at the boundaries of the grid, which are essential for subsequent boundary detection processes.

Boundary detection

Using parallel processing, the function calculates boundary grids by evaluating the arc distances from each spine point to all points in the embedding space (Figure S16 Algorithm 4). This is achieved through the $Get\ Arc\ Dist$ function (Figure S16 Algorithm 3), which computes arc tangent values and distances, grouping them to identify the maximum distances that define the boundaries.

Mask generation

Depending on the observer number j, the function either recalculates spines or reuses the existing ones to determine grids that should be masked out from analysis. This is done to focus the analysis on regions of interest, improving both the efficiency and

relevance of the results (Figure S16 Algorithm 5).

Adjacency and connectivity

Post mask generation, the function establishes adjacency relationships between the remaining grid points using the $Get\ Adjacent$ function (Figure S16 Algorithm 6). This step is crucial for understanding the connectivity and layout of the space, which aids in further analyses like clustering or region-based statistics.

Boundary refinement

Finally, the function refines the detected boundaries using the $Check\ Boundary$ function (Figure S16 Algorithm 7). This ensures that all boundary grids are accurately identified, incorporating any adjacent grids that may have been initially overlooked.

Annotation projecting function

Utilizing the information from the nearest point in original embedding to the grid as annotation for that specific grid point.

Pseudotime alignment function

As depicted in Figure S17 Algorithm, to generate a pseudotime ordering, we select grid points annotated as belonging to the starting subpopulation. From these, we randomly sample n points and calculate the average Dijkstra shortest path length from these n points to all remaining grid points, which serves as the pseudotime. In cases where a matched grid world encompasses two or more components, the component with the highest number of grids is considered the primary one and utilize for initial pseudotime computation. For the remaining components, their starting points are

determined as the points that have the minimum distance to any point in the primary component. The pseudotime values for these remaining points are then calculated based on the shortest path lengths from their respective starting points, using the pseudotime of the nearest point in the primary component as the starting value. This process is repeated for all components, ultimately yielding pseudotime values for the entire dataset.

Value projection function

As depicted in Figure S18, the value projection function calculates the weights W as a normalized exponential function of the squared distances D^2 from a point to its nearest N grids. The distances are scaled by the variance σ^2, which modulates the influence of each grid based on its proximity. This results in a weighted average where closer grids have a higher impact on the outcome.

$$W = \frac{exp\left(\frac{-D^2}{2\sigma^2}\right)}{\sum exp\left(\frac{-D^2}{2\sigma^2}\right)}$$

The value projection function that calculates a weighted average of the nearest N grids, where the weights are determined by a Gaussian kernel based on the distance vector and standard error. This function allows for smooth interpolation of values across the state space.

Lineage reward function

The targeting cluster grids are selected, and t represents the scaled pseudotime value. β is the decaying factor, R_d is the decision mode reward, and R_c is the contribution mode reward.

$$R_d = e^{-\beta t}$$

$$R_c = 1 - e^{-\beta t}$$

Gene reward function

Let r be the reward value, p be the punishing value, and t be the entire pseudotime. β is the decaying factor, R_d is the decision mode reward, and R_c is the contribution mode reward.

$$R_d = e^{-\beta t}(r - p)$$

$$R_c = \left(1 - e^{-\beta t}\right)(r - p)$$

Tabular Q learning

The Tabular Q-Learning algorithm is a form of reinforcement learning that updates the Q-values, which are estimates of the expected future rewards for a given state-action pair. The Q-values are stored in a table, with each entry corresponding to a state-action pair. The update rule for the Q-values is as follows:

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a'} Q\,(s', a') - Q(s, a)]$$

The update rule modifies the Q-value of the current state-action pair by incorporating the immediate reward plus the discounted maximum future reward, adjusted by the learning rate. The difference between the updated Q-value and the current Q-value is scaled by α to ensure smooth convergence.

Actor-critic architecture

Here, δ_t is the temporal difference error at time t, r_t is the reward received after taking an action at state s_t, γ is the discount factor, $V_\omega(s_t)$ is the value function parameterized by ω at state s.

$$\delta_t = r_t + \gamma V_\omega(s_{t+1}) - V_\omega(s_t)$$

δ_t is computed as the difference between the estimated value of the next state s_{t+1} discounted by γ and added to the reward r_t, and the estimated value of the current state s_t. This error measures the difference between the predicted value and the actual return received, guiding the updates for both the actor and the critic.

ω is updated using the learning rate α_ω, the sum of the product of the TD error δ_t and the gradient of the value function with respect to its parameters ω across all timesteps t.

$$\omega = \omega + \alpha_\omega \sum_t \delta_t \nabla_\omega V_\omega(s_t)$$

θ represents the parameters of the policy π_θ, which is updated using the learning rate α_θ, the TD error δ_t, and the gradient of the logarithm of the policy probability with respect to θ.

$$\theta = \theta + \alpha_\theta \sum_t \delta_t \nabla_\theta \, log \, \pi_\theta \, (a_t | s_t)$$

Autoencoder architecture for simulating

The encoder component of the autoencoder, denoted as E, processes these inputs through individual neural network layers, which are then concatenated to form a unified latent representation, z. This is mathematically represented as $E(x) = Concat(h1, h2, h3, h4)$, where x includes $z_{latent}, g_{expr}, c_{coord}, t_{pseudo}$. The decoder, D ,then takes this concatenated latent space as input and aims to reconstruct the normal distribution of one-step differences for the encoder's inputs, outputting $\Delta x \sim N(\mu, \sigma^2)$. This approach allows the model to capture the dynamics of

the input data by focusing on the changes rather than the absolute values.

The primary objective of our training process is to minimize the reconstruction error of the one-step differences, thereby effectively capturing the dynamic changes in the input spaces. This setup is particularly advantageous for simulating and understanding complex biological processes, such as cellular development or changes in gene expression over time, where accurate modeling of stepwise trajectories is crucial.

The superiorities of scRL

As depicted in Figure S19, the functionalities of current pseudotime analysis tools could be broadly categorized into three areas: pseudotime and differentiation potential analysis, trajectory inference, and gene trends analysis. Most available packages estimated global pseudotime, with exceptions such as FateID[41], which primarily focused on fate commitment value estimation, and Slingshot[11], which emphasized pseudotime and trajectory inference based on principal curves. While pseudotime alignment was not unique to scRL, our method demonstrated superior performance under specific conditions (Figure S20C), positioning it as a competitive option in the rapidly evolving landscape of single cell pseudotime analysis tools. In the realm of lineage potential analysis, only a few methods, such as Palantir, FateID, and scVelo, could perform such estimations. Notably, scVelo required both spliced and unspliced transcriptomic data, which limited its applicability. Palantir and FateID utilized supervised learning approaches, while scRL introduced a novel deep reinforcement learning-based method. Importantly, scRL's lineage potential estimation strongly correlated with the embedding's manifold (Figure S21B), effectively leveraging

trajectory information. When assessing differentiation potential, scRL adopted a biologically meaningful approach by designating early-stage or stem cells as both reward lineages and starting points, simulating the fate decision process of stem cells. This approach contrasts with methods like Palantir and FateID, which rely on the entropy of already estimated lineage potentials. Regarding trajectory inference, the Monocle series preferred the principal trees method, whereas Slingshot and FateID favored the principal curves approach. Both methods provided representations of the original data's 2-D embedding space, showcasing the data generation process. However, by returning to the essence of the trajectory itself and employing the policy network to sample paths depicting the differentiation process, scRL introduced a fresh perspective to these techniques and a comparative analysis between our method and existing techniques had been conducted (Figure S22). Innovatively, scRL could simulate the differentiation process beyond the original data scope. For gene trends analysis, nearly all methods could perform comprehensive pseudotime alignment and predict gene trends through a general additive linear model (GAM). Similarly, scRL set gene expression as the reward and utilized the critic network to predict the gene decision state value, which should be defined as the intensity of the pre-gene expression state.

Application of grid embedding and superiority of scRL pseudotime

Different embeddings offer distinct perspectives on the differentiation process. Here, we explore the versatility of our grid embedding function by applying it to six diverse dimensional reduction methods: PCA, diffusion map, UMAP, t-SNE, FA (force atlas),

and Phate. Remarkably, each of these methods achieves impressive results when combined with grid embedding (Figure S20A). Furthermore, the pseudotime alignment is seamlessly integrated with each grid embedding, providing a comprehensive view of the data (Figure S20 B). To validate the effectiveness of scRL's pseudotime, we conducted correlation analyses using Pearson, Spearman, and Kendall metrics. The results demonstrate that scRL's pseudotime exhibits the highest correlation with the distance to the starting point on each embedding, which was established as ground truth (Figure S20C). And the superior could be interpreted as smoother pseudotime is estimated regarding each embedding.

Comparison of lineage potential between methods

In the realm of lineage potential calculation, several methods have emerged, including FateID - a random forest-based supervised learning approach, and Palantir - a diffusion map-based unsupervised method. scRL, however, leveraging deep reinforcement learning and a critic neural network to estimate potential. To assess the temporal association of lineage potential, we selected four well-distinguished hematopoietic branch embedding spaces: DC, UMAP, FA, and PHATE (Figure S21A). Correlation analysis revealed that scRL's lineage potential consistently achieves superior temporal correlation scores which indicates that the potential changing smoothly with time, while Palantir and FateID approaches are negative associated with time indicating more sketchy estimation (Figure S21B). The comparison highlighted scRL’s effectiveness in capturing temporal dynamics.

Distinctive Trajectory Representations Across Methods

scRL's trajectory sampling is influenced by the gamma parameter, allowing for adjustable path shrinkage or divergence (Figure S23, 24). Compared to other methods, scRL's trajectory is based on grid embedding paths, showcasing all possible routes from source to target in a 2-D embedding. In contrast, Monocle2 relies on a reverse graph embedding of the original data, resulting in a potentially unitary path. Monocle3's path is derived from the UMAP embedding's principal tree, which lacks diversity. Slingshot outlines the backbone of differentiation using principal curves of any 2-D embedding, resembling Monocle2's approach. PAGA, a cluster-based method, presents a coarse-grained neighborhood structure among clusters. While each method offers valuable insights, scRL's adaptability to different embeddings and its ability to simulate authentic differentiation paths from start to maturity constitute its innovative advantage as scRL is designed for optimizing a policy neural network and sampling a real trajectory on any embedding.

Parameters

The parameter n represents the density of the background grid to be generated in the preprocessing step of scRL. Specifically, it generates an n-by-n grid embedding based on the circumscribed rectangle of the input two-dimensional embedding of single cells.

The parameter j denotes the number of observation points for grid embedding in the preprocessing step of scRL. The grid embedding function of scRL generates j observation points equidistantly on the four boundaries of the circumscribed rectangle of the original two-dimensional data embedding, to scan the edges of the original data.

More observation points indicate finer edges.

The parameter γ is a factor in the reinforcement learning training process of scRL, representing the discount degree of future rewards during the policy iteration process. Its value ranges from 0 to 1. A higher value indicates greater emphasis on future rewards, making the agent more forward-looking. However, too low a value may prevent the agent from learning the optimal differentiation strategy, resulting in significant deviations in the evaluation of state value.

The parameter α represents the learning rate during the policy iteration process, indicating the acceptance of new observations. The default learning rate for the policy net optimizer is set to 0.005, while the default learning rate for the critic net optimizer is set to 0.0001.

The parameter *n neighbor* represents the number of neighbors considered when mapping between grid embedding and original embedding.

Parameters robustness

We have conducted experiments to examine the robustness of scRL's key parameters. Primarily, we focused on parameters *n* and *j* within the grid embedding function. Here, *n* represents the width of the anticipated grid world; a higher value of n indicates a denser grid environment. Meanwhile, j can be interpreted as the number of observers surrounding the grid. As the number of observers increases, the generated boundary becomes more precise (Figure S25, 26). It is important to note that excessively high j value could potentially introduce inconsistencies within the grid world, which may make certain island areas inaccessible to the agent, thereby rendering their state

values unavailable. To assess the stability of pseudotime estimations, we tested various values of the parameter n for each value of the parameter j and found that the results remained stable across different groups of parameters j (Figure S27A-B).

Calculation of correlation coefficient

As for the pseudotime correlation with distance-based ground truth, we calculate the Pearson, Spearman and Kendall correlation coefficient with Pandas package after collecting the pseudotime and distance into one data frame. As for the time correlation of genes expression value or lineage contribution value, the Pearson correlation coefficient was calculated.

Calculation of distance-based ground truth

The distance-based ground truth for cells is the Euclidean distance from a pre-defined starting cell in a certain 2-D embedding space.

Calculation of weighted average time

The pseudotime is weighted by the expression value of genes or lineage decision value which the sum of weight contribution is 1.

Availability of codes

To ensure the reproducibility and transparency of our study, the software tool we developed has been made publicly available on GitHub Readers can access the source code and related documentation via the following link: https://github.com/PeterPonyu/scRL

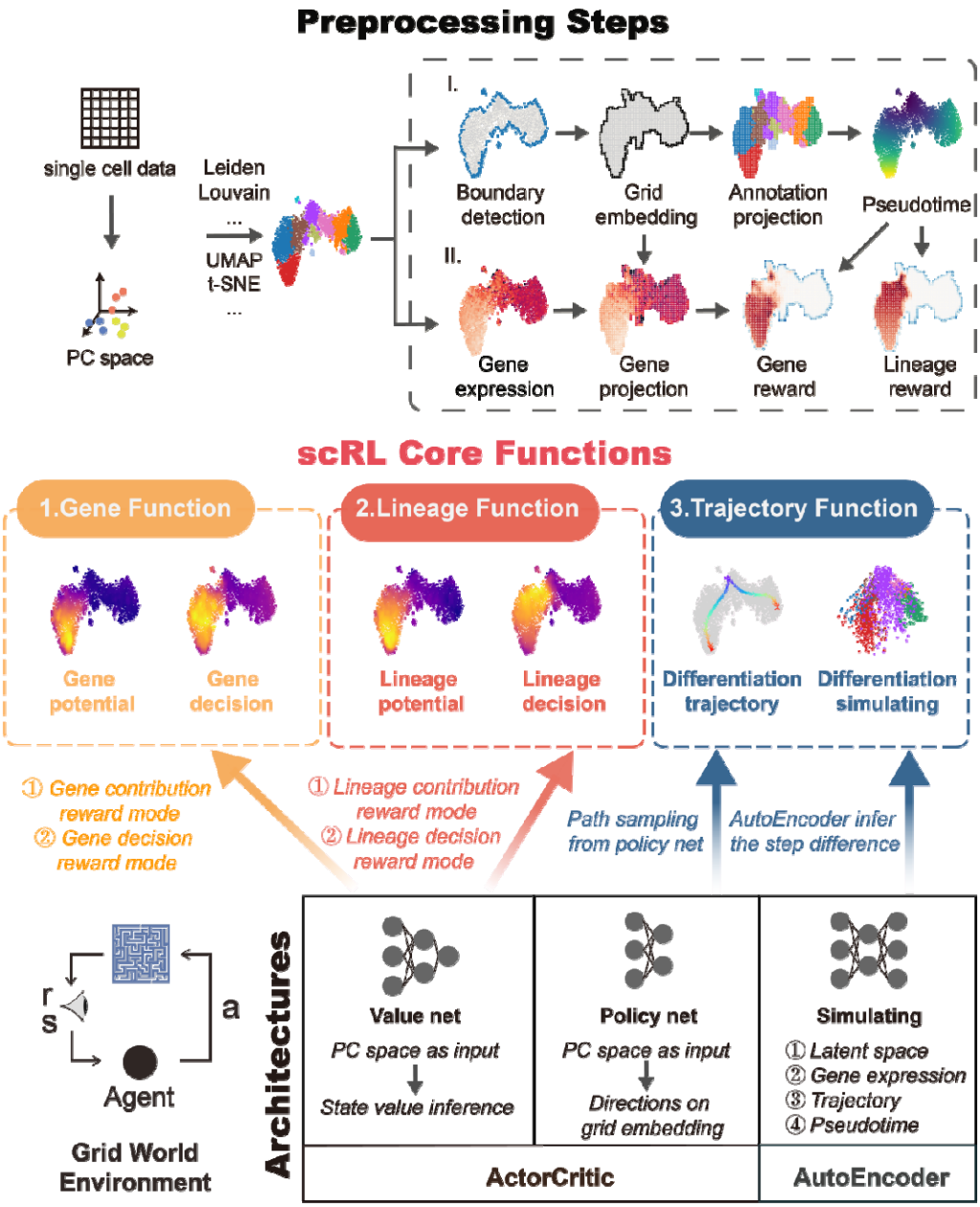
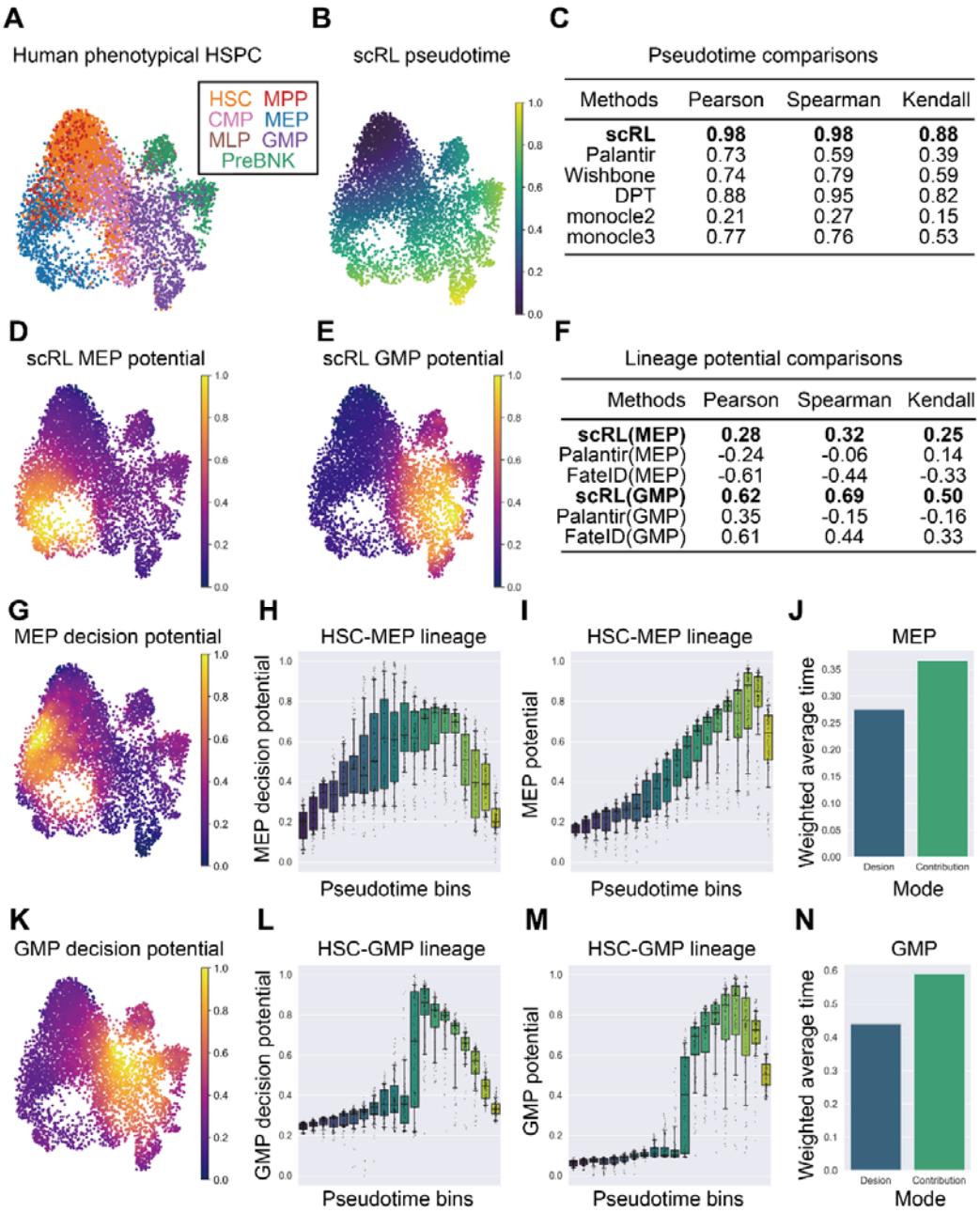


Figure 1: The architecture of scRL

The top panel of the figure outlines the preprocessing steps. It commences with Principal Component Analysis, progressing to dimensionality reduction and clustering techniques rooted in the principal component space. Through an edge detection algorithm, a grid embedding representing the original space emerges within the derived two-dimensional framework. Next, subpopulation information is mapped onto

this grid embedding, and a pseudotime is computed from a designated starting subpopulation. Parallel to this, projecting gene expression data from the initial space onto the grid embedding gives rise to a gene reward environment. These dual reward environments, governed by their dynamic relationship with pseudotime shifts, yield a decision mode that diminishes along the pseudotime and a contribution mode that escalates in tandem. The central portion of the figure showcases scRL's core functional modules, primarily segmented into three primary categories: gene, lineage, and trajectory modules. The gene module incorporates gene potential, derived from the contribution reward mode, alongside the gene decision value, emanating from the decision reward mode. The lineage module encompasses lineage potential from the contribution reward mode and lineage decision value from its decision counterpart. Lastly, the trajectory module encompasses both the differentiation trajectory and its simulation. The figure's lower panel reveals the underlying structure of scRL. It commences with a pattern of the interplay between a reinforcement learning agent and its surroundings. This prelude segues into the Actor-Critic algorithm architecture. Here, the value net processes the principal component space, outputting the state value for each node. The policy net, operating on the same input space, selects differentiation strategies. An AutoEncoder is then utilized to simulate differentiation, processing the latent space from the preceding step, gene expression data, differentiation pathway, and pseudotime, subsequently generating predictions for the ensuing transition.



Methods	Pearson	Spearman	Kendall
scRL	**0.98**	**0.98**	**0.88**
Palantir	0.73	0.59	0.39
Wishbone	0.74	0.79	0.59
DPT	0.88	0.95	0.82
monocle2	0.21	0.27	0.15
monocle3	0.77	0.76	0.53

Methods	Pearson	Spearman	Kendall
scRL(MEP)	**0.28**	**0.32**	**0.25**
Palantir(MEP)	-0.24	-0.06	0.14
FateID(MEP)	-0.61	-0.44	-0.33
scRL(GMP)	**0.62**	**0.69**	**0.50**
Palantir(GMP)	0.35	-0.15	-0.16
FateID(GMP)	0.61	0.44	0.33

Figure 2: Pseudotime and lineage potential function of scRL

(A)UMAP projection of phenotypes of HSPC (hematopoietic stem and progenitor cell), including HSC (hematopoietic stem cell), MPP (multipotent progenitor), CMP (common myeloid progenitor), MLP (multipotent lymphoid progenitor), MEP (Megakaryocyte erythrocyte progenitor), GMP (Granulocyte monocyte progenitor),

and PreBNK (Pre-B cell and Pre-natural killer cell).(B)Pseudotime aligned by scRL.(C) Comparisons between scRL and other methods including Palantir, Wishbone, DPT (diffusion pseudotime), Monocle2, Monocle3 with highest score in bold format. Pearson, Spearman and Kendall correlations are computed between pseudotime and the distance-based ground truth. The ground truth is defined as Euclidean distance from starting point in UMAP space. (D)MEP potential inferred by scRL with MEP cluster as reward and HSC cluster as start. (E)GMP potential inferred by scRL with GMP cluster as reward and HSC cluster as start.(F) Comparisons between scRL and other methods including Palantir and FateID with highest score in bold format. Pearson, Spearman and Kendall correlations are computed between MEP as well as GMP potential and the distance-based ground truth. The ground truth is defined as Euclidean distance from starting point in UMAP space. (G)MEP decision potential inferred by scRL with MEP cluster as reward and HSC cluster as start.(H-I)MEP potential and decision potential within HSC-MEP lineage along binned pseudotime.(J)Weighted average of pseudotime by MEP decision potential and lineage(contribution) potential.(K)GMP decision potential inferred by scRL with GMP cluster as reward and HSC cluster as start.(L-M)GMP potential and decision potential within HSC-GMP lineage along binned pseudotime.(N)Weighted average of pseudotime by GMP decision potential and lineage(contribution) potential.

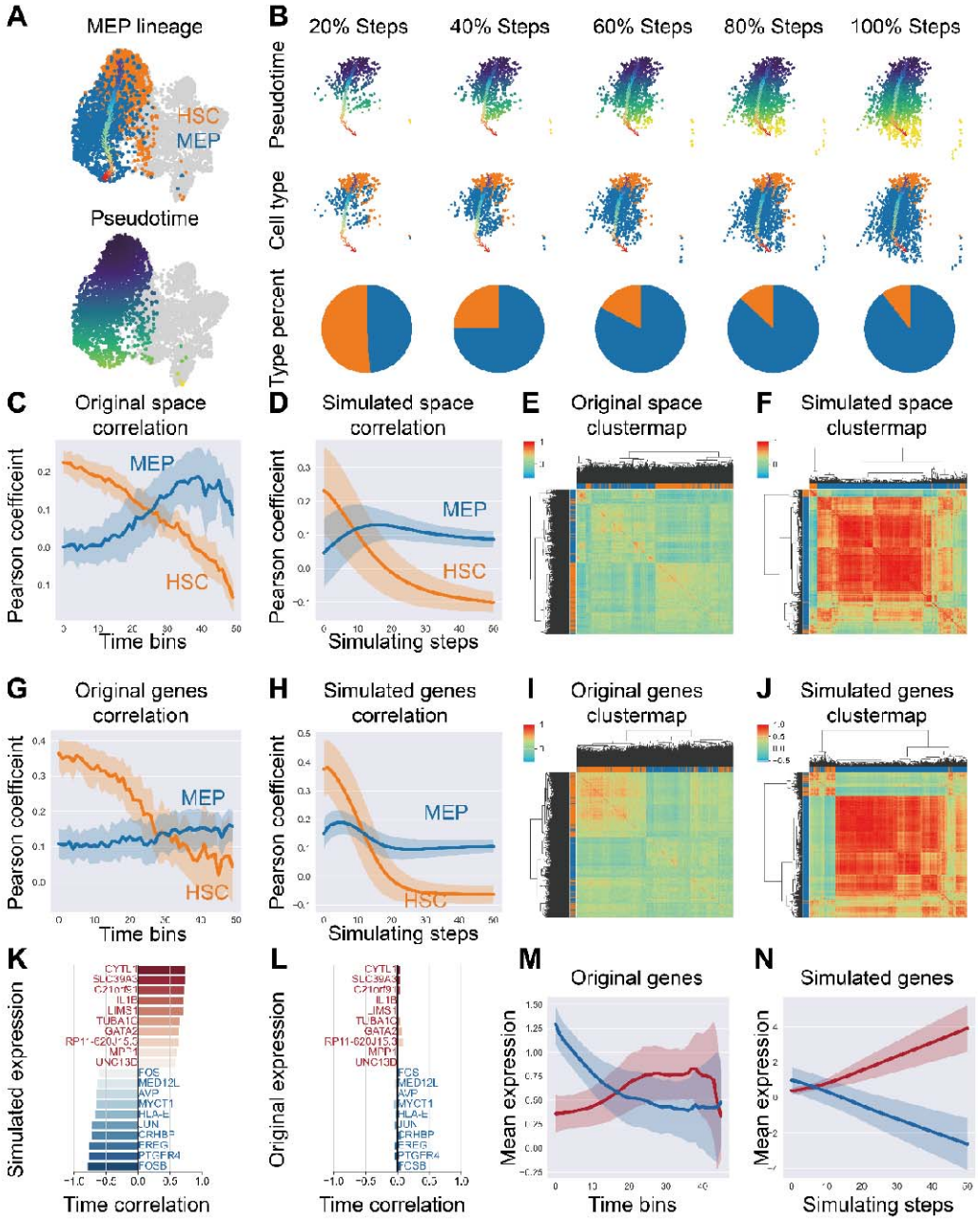


Figure 3: Trajectory simulating function of scRL

(A)MEP lineage and pseudotime aligned by scRL (B) Simulated MEP trajectory with 20%, 40%, 60%, 80% steps and the whole trajectory. A single trajectory steps 50 times and 100 trajectories in total. Cell types are colored as shown in (A) and black dots referred to the starting points. Pie plots show the dynamic changes of cell type

components. (C) Pearson correlation coefficient of the original space of cells along uniformly 50-binned pseudotime, the mean value is represented by a line and the standard error is indicated by the colored filled area, examining the relationship between all cells in relation to MEP and HSC. (D)The same as (C) but in simulated space. (E)Hierarchical clustering of Pearson correlation between all the cells in original space, cells are colored as in (A). (F)The same as (E) but in simulated space. (G)The same as (C) but the coefficient is calculated in gene expression space. (H)The same as (C) but the coefficient is calculated in simulated gene expression space. (I)The same as (E) but in gene expression space. (J)The same as (E) but in simulated gene expression space. (K)The correlation score with pseudotime of 10 most and 10 least relevant genes in simulated space. (L)The same as (K) but in gene expression space. (M)The moving average of original mean expression along binned pseudotime of 10 most time relevant genes are colored as MEP, and 10 least relevant genes are colored as HSC, with line represents the mean value and filled area indicates the standard error. (N)The same as (M) but in simulated space.

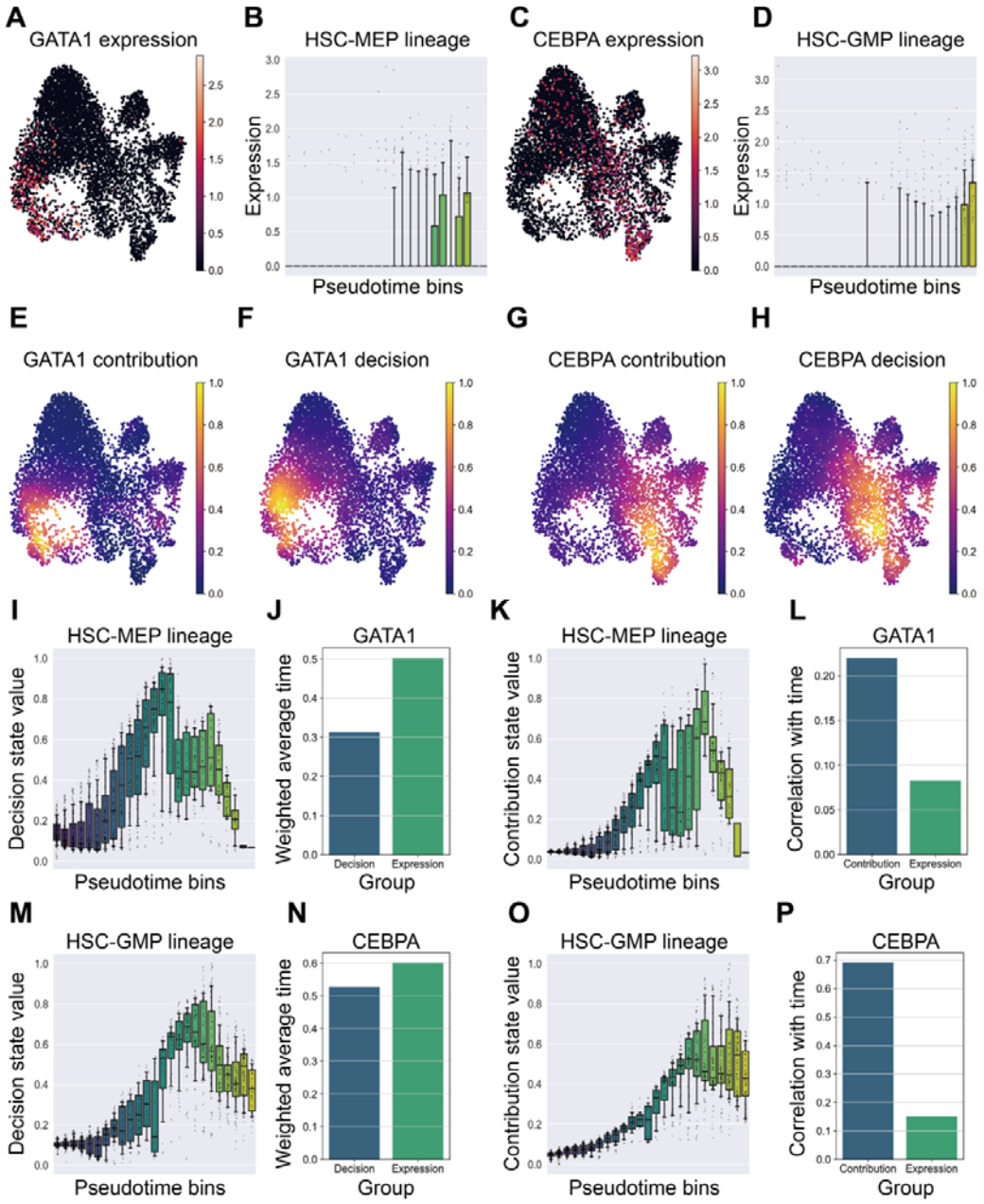


Figure 4: Gene decision function of scRL

(A)GATA1 expression on UMAP marking the erythroid lineage. (B)GATA1 expression on HSC-MEP (erythroid)lineage along binned pseudotime. (C)CEBPA expression on UMAP marking the myeloid lineage. (D)CEBPA expression on HSC-GMP (myeloid)lineage along binned pseudotime. (E)GATA1 gene contribution value on UMAP. (F)GATA1 gene decision value on UMAP (G)CEBPA

gene contribution value on UMAP. (H)CEBPA gene decision value on UMAP. (I)The HSC-MEP lineage decision value along binned pseudotime. (J)The average pseudotime in HSC-MEP lineage weighted by GATA1 decision value and expression. (K)The HSC-MEP lineage contribution value along binned pseudotime. (L)The average correlation with pseudotime of GATA1 contribution value and its expression. (M)The HSC-GMP lineage decision value along binned pseudotime. (N)The average pseudotime in HSC-GMP lineage weighted by CEBPA decision value and expression. (O)The HSC-GMP lineage contribution value along binned pseudotime. (P)The average correlation with pseudotime of CEBPA contribution value and its expression.

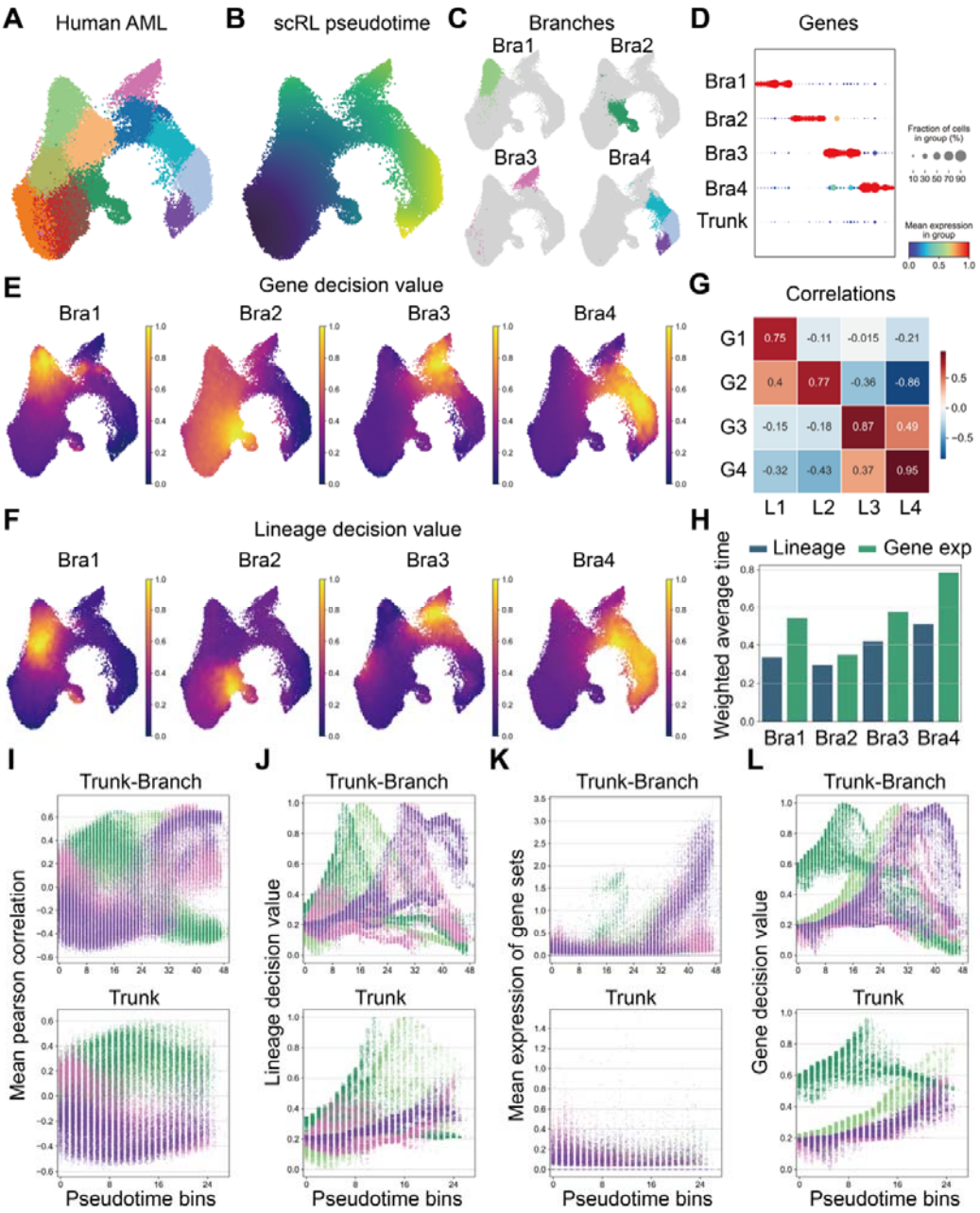


Figure 5: Combining lineage and gene information to decode fate decision states

(A)UMAP of human acute myeloid leukemia (AML), colored with clusters. (B)Pseudotime inferred by scRL. (C)4 Branches of the UMAP space. (D)10 marker genes sorted by log-foldchanges for each branch, with at least 25% percent cells express but less than 25% cells express in reference cell. (E)Gene set decision value of each branch. (F)Lineage decision value of each branch. (G)Pearson correlation

heatmap of each branches' gene set decision value represented by G1-4 and lineage decision value represented by L1-4. (H)Average pseudotime of the dataset weighted by the lineage decision value as well as the gene expression of each branch. (I)Mean Pearson correlation with branch specific cells along uniformly 50-binned pseudotime of the whole dataset at top panel and the trunk part at bottom panel. (J)Lineage decision value of each branch along uniformly 50-binned pseudotime of the whole dataset at top panel and the trunk part at bottom panel. (K)Mean expression of branch specific gene sets along uniformly 50-binned pseudotime of the whole dataset at top panel and the trunk part at bottom panel. (L)Gene decision value of branch specific gene sets along uniformly 50-binned pseudotime of the whole dataset at top panel and the trunk part at bottom panel.

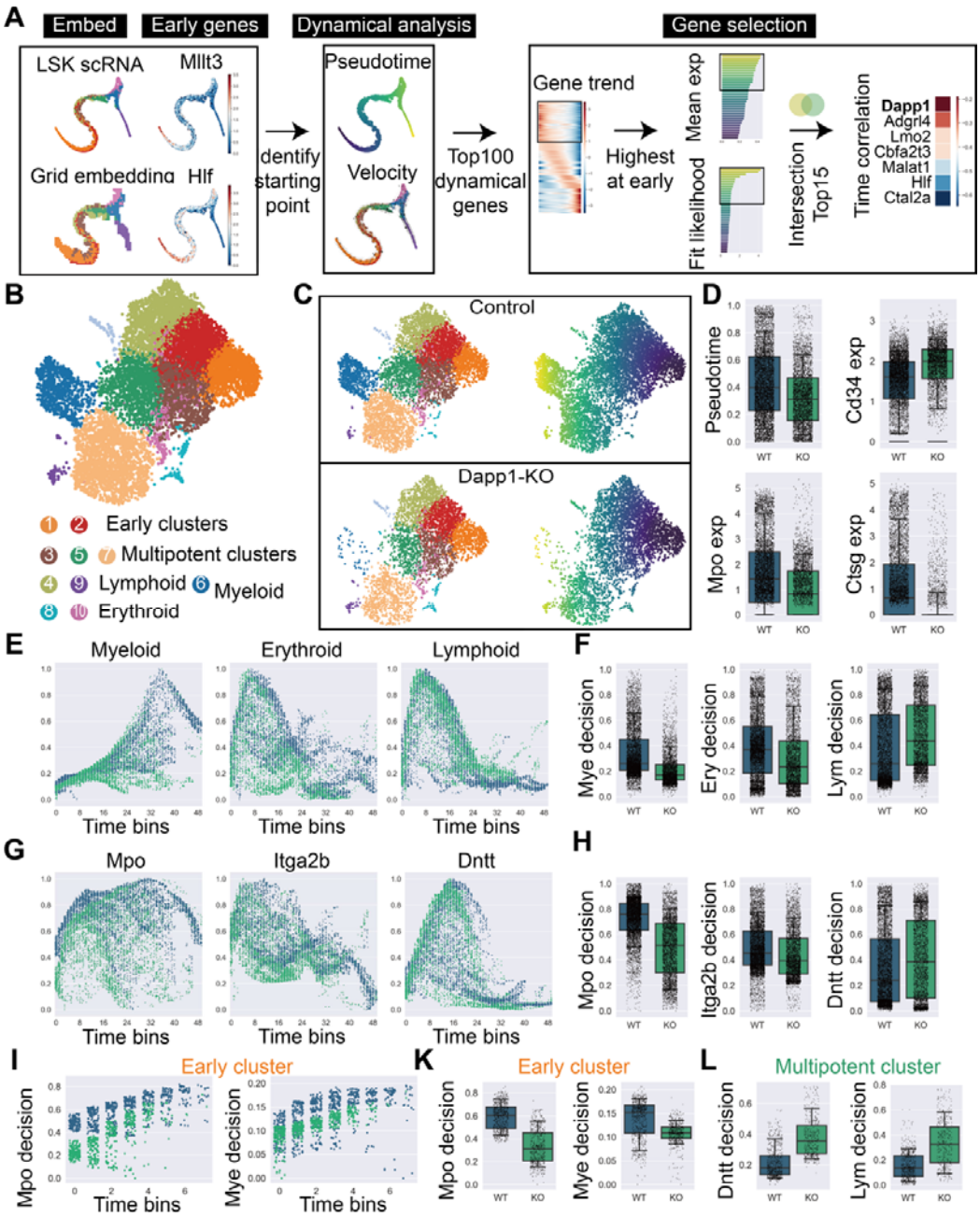


Figure 6: Lineage commitment of HSCs at the earliest primitive stage perturbed by knockout of critical dynamical gene.

(A)The workflow for identifying critical dynamical gene Dapp1 in the differentiation of HSCs begins from left panel with embeddings of LSK single-cell data to pinpoint the differentiation starting point. It then employs velocity analysis to identify the top

dynamical genes involved in differentiation, meanwhile pseudotime is used to select the earliest expressing genes, from which the highest expressed and most dynamic genes are intersected. Finally, the right panel displays a time correlation score to rank the intersected genes. (B)Dapp1 knockout LSK atlas with clusters 1,2 annotated as Early clusters, 3,5,7 as multipotent clusters, 4, 9 as lymphoid clusters, 6 as myeloid cluster and 8, 10 as erythroid clusters. (C)Cluster distributions and pseudotime for control group and knockout group. (D)Box plots for comparisons of pseudotime and the expression of Cd34, Mpo, Ctsg. (E)Lineage decision value of myeloid, erythroid and lymphoid along uniformly 50-binned pseudotime. (F)Box plots for comparison of lineage decision value of myeloid, erythroid and lymphoid lineages. (G)Gene decision value of Mpo, Itga2b and Dntt along uniformly 50-binned pseudotime. (H)Box plots for comparison of gene decision value of Mpo, Itga2b and Dntt. (I)The myeloid lineage, Mpo decision value within early cluster 1 along binned pseudotime. (J)Box plots for comparison of myeloid lineage and Mpo decision value within early cluster 1. (K)Box plots for comparison of lymphoid lineage and Dntt decision value within multipotent cluster 5.

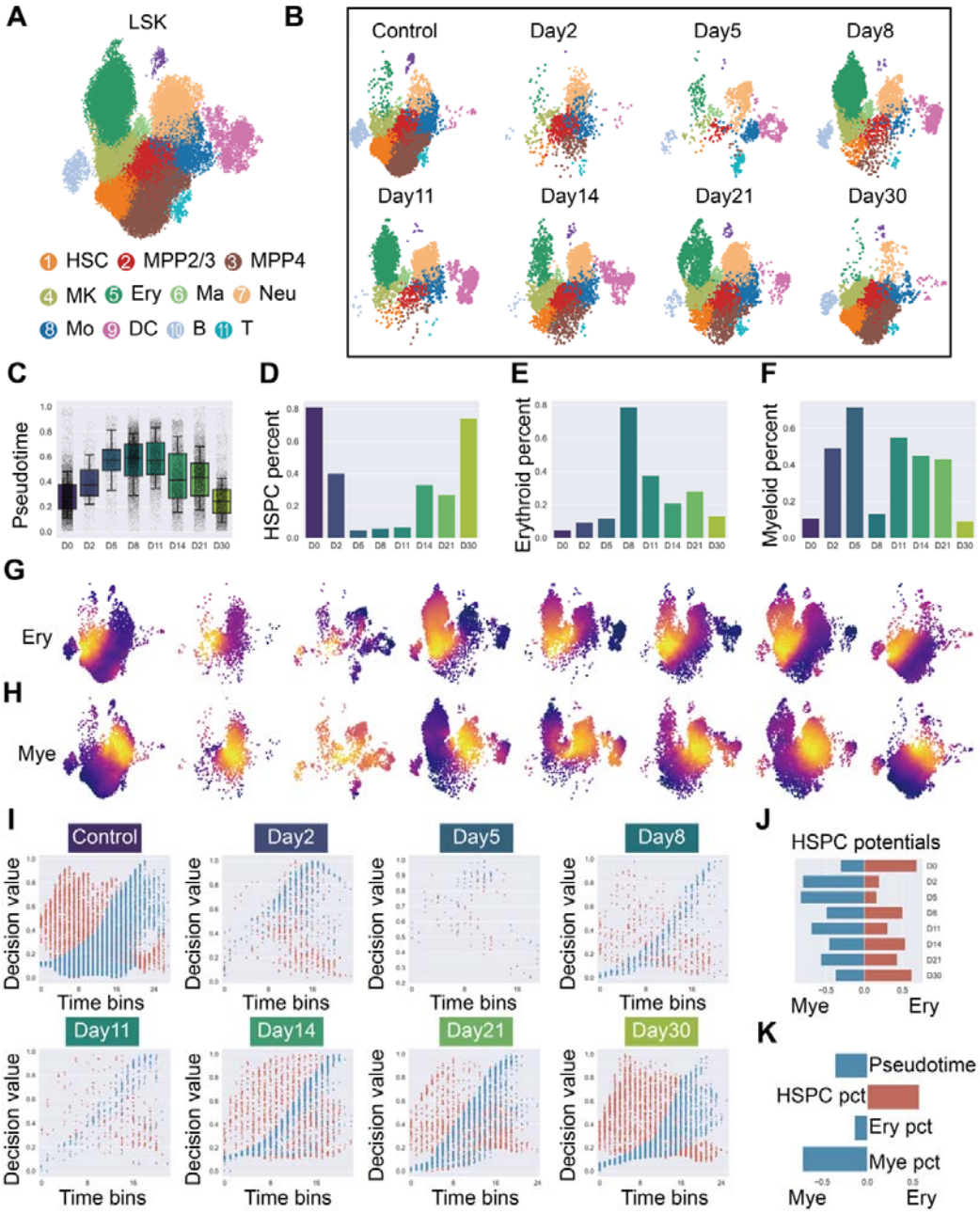


Figure 7: Erythroid bias of primitive HSCs correlated with their proportion in time series of irradiation recovery

(A)LSK (lin-Sca1+ckit+) phenotype bone marrow cells, clusters were annotated as HSC (Hematopoietic stem cell), MPP2/3(Multipotent progenitors 2/3), MPP4(Multipotent progenitors 4), MK(Megakaryocyte), Ery(Erythrocyte), Ma(Mast cell), Neu(Neutrophil), Mo(Monocyte), DC(Dendritic cell), B(B cell), T(T cell)

(B)Single cell atlas of irradiation injured LSKs at day points 2, 5, 8, 11, 14, 21, and 30, colored with annotated cell types in (A). (C)Box plot of pseudotime at each day points. (D)HSPC (HSC, MPP2/3, MPP4) proportion at each day points. (E)Erythroid cells (MK, Ery, Ma) proportion at each day points. (F)Myeloid cells (Gra, Mo, DC) proportion at each day points. (G)Erythroid lineage decision value projected on UMAP space at different time points. (H)Myeloid lineage decision value projected on UMAP space at different time points. (I)Lineage decision atlas for each time points within HSPC, erythroid decision value (dots colored in red) and myeloid decision value (dots colored in blue) along uniformly 50-binned pseudotime. (J)Horizonal bar plot of erythroid biased proportion (colored in red) and myeloid biased proportion (colored in blue) of HSPC at different time points. (K)Horizonal bar plot of the Pearson correlation of erythroid biased HSPC proportion, myeloid biased HSPC proportion at each time points with respect to the average pseudotime, HSPC proportion, erythroid proportion and myeloid proportion at each time points

References

1 Auerbach, B. J., Hu, J., Reilly, M. P. & Li, M. Applications of single-cell genomics and computational strategies to study common disease and population-level variation. *Genome Res* **31**, 1728-1741 (2021).

2 Yang, X. H. *et al.* Detecting critical transition signals from single-cell transcriptomes to infer lineage-determining transcription factors. *Nucleic Acids Res* **50**, e91 (2022).

3 Hou, W. *et al.* A statistical framework for differential pseudotime analysis with multiple single-cell RNA-seq samples. *Nat Commun* **14**, 7286 (2023).

4 Pandey, K. & Zafar, H. Inference of cell state transitions and cell fate plasticity from single-cell with MARGARET. *Nucleic Acids Res* **50**, e86 (2022).

5 Pillai, M., Chen, Z., Jolly, M. K. & Li, C. Quantitative landscapes reveal trajectories of cell-state transitions associated with drug resistance in melanoma. *iScience* **25**, 105499 (2022).

6 Sun, S., Zhu, J., Ma, Y. & Zhou, X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biol* **20**, 269 (2019).

7 Wang, K. *et al.* Comparative analysis of dimension reduction methods for cytometry by time-of-flight data. *Nat Commun* **14**, 1836 (2023).

8 McInnes, L. & Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv* **abs/1802.03426** (2018).

9 Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* **20**, 59 (2019).

10 Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).

11 Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics* **19**, 477 (2018).

12 Sugihara, R., Kato, Y., Mori, T. & Kawahara, Y. Alignment of single-cell trajectory trees with CAPITAL. *Nat Commun* **13**, 5972 (2022).

13 Sagar & Grün, D. Deciphering Cell Fate Decision by Integrated Single-Cell Sequencing Analysis. *Annual review of biomedical data science* **3**, 1-22 (2020).

14 Kempka, M., Wydmuch, M., Runc, G., Toczek, J. & Jaśkowski, W. in *2016 IEEE Conference on Computational Intelligence and Games (CIG).* 1-8.

15 Mnih, V. *et al.* Playing Atari with Deep Reinforcement Learning. *ArXiv* **abs/1312.5602** (2013).

16 Brahmbhatt, S. & Hays, J. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 3087-3096.

17 Jia, Y. & Zhou, X. Y. Policy Gradient and Actor-Critic Learning in Continuous Time and Space: Theory and Algorithms. *J. Mach. Learn. Res.* **23**, 275:271-275:250 (2021).

18 Bank, D., Koenigstein, N. & Giryes, R. Autoencoders. *ArXiv* **abs/2003.05991**

(2020).
19 Grinenko, T. *et al.* Hematopoietic stem cells can differentiate into restricted myeloid progenitors before cell division in mice. *Nat Commun* **9**, 1898 (2018).
20 Wakabayashi, A. *et al.* Insight into GATA1 transcriptional activity through interrogation of cis elements disrupted in human erythroid disorders. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 4434-4439 (2016).
21 Cytlak, U. *et al.* Differential IRF8 Transcription Factor Requirement Defines Two Pathways of Dendritic Cell Development in Humans. *Immunity* **53**, 353-370.e358 (2020).
22 Nechanitzky, R. *et al.* Transcription factor EBF1 is essential for the maintenance of B cell identity and prevention of alternative fates in committed cells. *Nature immunology* **14**, 867-875 (2013).
23 Magenheim, J. *et al.* Ngn3(+) endocrine progenitor cells control the fate and morphogenesis of pancreatic ductal epithelium. *Developmental biology* **359**, 26-36 (2011).
24 Yu, X.-X. *et al.* Sequential progenitor states mark the generation of pancreatic endocrine lineages in mice and humans. *Cell research* **31**, 886-903 (2021).
25 Naldini, M. M. *et al.* Longitudinal single-cell profiling of chemotherapy response in acute myeloid leukemia. *Nat Commun* **14**, 1285 (2023).
26 Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979-982 (2017).
27 Amann-Zalcenstein, D. *et al.* A new lymphoid-primed progenitor marked by Dach1 downregulation identified with single cell multi-omics. *Nature immunology* **21**, 1574-1584 (2020).
28 Klein, F. *et al.* Dntt expression reveals developmental hierarchy and lineage specification of hematopoietic progenitors. *Nature immunology* **23**, 505-517 (2022).
29 Challen, G. A., Boles, N., Lin, K. K.-Y. & Goodell, M. A. Mouse hematopoietic stem cell identification and analysis. *Cytometry. Part A : the journal of the International Society for Analytical Cytology* **75**, 14-24 (2009).
30 Laurenti, E. & Göttgens, B. From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553**, 418-426 (2018).
31 Giladi, A. *et al.* Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nature cell biology* **20**, 836-846 (2018).
32 Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e1516 (2018).
33 Gao, S. *et al.* Identification of HSC/MPP expansion units in fetal liver by single-cell spatiotemporal transcriptomics. *Cell research* **32**, 38-53 (2022).
34 Doi, Y. *et al.* Variable SATB1 Levels Regulate Hematopoietic Stem Cell Heterogeneity with Distinct Lineage Fate. *Cell reports* **23**, 3223-3235 (2018).
35 Neftci, E. O. & Averbeck, B. B. Reinforcement learning in artificial and

biological systems. *Nature Machine Intelligence* **1**, 133-143, doi:10.1038/s42256-019-0025-4 (2019).

36 Mahmud, M., Kaiser, M. S., Hussain, A. & Vassanelli, S. Applications of Deep Learning and Reinforcement Learning to Biological Data. *IEEE Transactions on Neural Networks and Learning Systems* **29**, 2063-2079 (2017).

37 Wang, P. *et al*. Deciphering driver regulators of cell fate decisions from single-cell transcriptomics data with CEFCON. *Nat Commun* **14**, 8459 (2023).

38 Torregrosa, G. & Garca-Ojalvo, J. Mechanistic models of cell-fate transitions from single-cell data. *Current Opinion in Systems Biology* (2021).

39 Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat Biotechnol* **37**, 547-554 (2019).

40 Setty, M. *et al.* Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol* **34**, 637-645 (2016).

41 Herman, J. S., Sagar & Grün, D. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat Methods* **15**, 379-386 (2018).