nature methods

Article

https://doi.org/10.1038/s41592-024-02567-1

High-resolution, noninvasive single-cell lineage tracing in mice and humans based on DNA methylation epimutations

Received: 14 May 2024

Accepted: 19 November 2024

Published online: 16 January 2025

Check for updates

Mengyang Chen[1,2,4], Ruijiang Fu[1,2,3,4], Yiqian Chen[1,2], Li Li[1,2] ✉ & Shou-Wen Wang[1,2,3] ✉

In vivo lineage tracing holds great potential to reveal fundamental principles of tissue development and homeostasis. However, current lineage tracing in humans relies on extremely rare somatic mutations, which has limited temporal resolution and lineage accuracy. Here, we developed a generic lineage-tracing tool based on frequent epimutations on DNA methylation, enabled by our computational method MethylTree. Using single-cell genome-wide DNA methylation datasets with known lineage and phenotypic labels, MethylTree reconstructed lineage histories at nearly 100% accuracy across different cell types, developmental stages, and species. We demonstrated the epimutation-based single-cell multi-omic lineage tracing in mouse and human blood, where MethylTree recapitulated the differentiation hierarchy in hematopoiesis. Applying MethylTree to human embryos, we revealed early fate commitment at the four-cell stage. In native mouse blood, we identified ~250 clones of hematopoietic stem cells. MethylTree opens the door for high-resolution, noninvasive and multi-omic lineage tracing in humans and beyond.

Tracing lineage histories in model organisms through genetic manipulation has improved greatly over the past decade[1]. High-resolution lineage tracing can be achieved by labeling individual cells with induced and heritable DNA mutations[2–12], which can be profiled later through single-cell sequencing. We have recently developed DARLIN, a highly efficient lineage-tracing mouse model that can generate ~10^{18} distinct lineage barcodes on induction at a defined time window[13]. Applications of these recent lineage-tracing tools have revealed important insights regarding cell fate choice[2,13–19], cell migration dynamics[13], cancer evolution[20,21] and clonal memory[13,22].

In contrast, lineage tracing in humans has been much less developed as genetic manipulation is prohibited. Cell lineages in humans can be inferred from somatic mutations in our genome[23,24]. However, this approach requires whole-genome DNA sequencing of single-cell derived colonies, which is low-throughput and does not provide cell-state information[25,26]. Although mitochondrial DNA (mtDNA) mutations could be used to trace cell lineages at higher throughput[27], they undergo complex processes of inheritance and selection, which may give poor lineage accuracy[28–30]. Due to extremely slow somatic mutations (~10^{-9} per nucleotide per cell division)[31], these methods likely cannot resolve lineages at a much shorter timescale like days. It is highly desirable to develop an alternative noninvasive lineage-tracing method that provides a high 'temporal' lineage resolution, achieves nearly 100% accuracy, and is also compatible with single-cell multi-omic profiling.

In mammals, DNA methylation occurs mostly at the cytosine residue in the CpG dinucleotide, which changes over time due to epimutations that occur at a rate of ~0.001 per CpG site per division[32–35]. We recently showed in hematopoietic stem cells (HSCs) that clonal memory persists stably in DNA methylation for at least a few months, but not in chromatin accessibility or gene expression[13]. This motivates

[1]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, China. [2]School of Life Sciences, Westlake University, Hangzhou, China. [3]School of Science, Westlake University, Hangzhou, China. [4]These authors contributed equally: Mengyang Chen, Ruijiang Fu. ✉e-mail: lili@westlake.edu.cn; wangshouwen@westlake.edu.cn

us to develop a generic lineage-tracing tool based on epimutations on DNA methylation. Previously, clone-specific DNA methylation patterns have been observed in different contexts[36–42]. Bulk methylation data have been used to track the evolution of individual alleles and infer single-cell transition rates[43]. Epimutations have been explored to infer single-cell lineages[44–46], mostly in the context of cancer. However, the inferences were not convincing due to the lack of benchmark against actual lineages from neutral labels. Furthermore, cancer cells are known to have strong genomic instability and highly aberrant DNA methylation[47]. Therefore, it remains a substantial challenge to infer lineages from epimutations in normal cells.

Developing an epimutation-based generic lineage-tracing tool needs to address four critical challenges. First, along with lineage-specific epimutations, cells also adopt cell-type-specific DNA methylation changes during differentiation[48]. Correctly inferring lineages from such confounding signals is crucial for lineage tracing, and this challenge cannot be addressed by existing lineage inference methods in cancer[44–46]. Second, global DNA methylation level undergoes drastic modulation during development[48], which could disrupt existing epimutations and cripple lineage inference from DNA methylation. In addition, single-cell genome-wide DNA methylation data typically covers only ~5% of the genome, leading to a highly sparse matrix with >95% missing values, which makes lineage inference extremely challenging. Finally, how to deal with the heterogeneous measurement noise between cells is also a tricky problem. So far, there is not yet a systematic approach that addresses all these challenges to enable a generic lineage-tracing tool from sparse, genome-wide DNA methylation data. Below, we developed a computational framework called MethylTree to address these challenges and demonstrate its power in broad biological contexts.

Results

Lineage inference from sparse single-cell DNA methylation

We first considered the problem of lineage inference in a homogeneous population of the same cell type. In this case, DNA methylation differences among these cells would result from stochastic epimutations at individual CpG sites, which can be used to infer cell division histories (Fig. 1a). Genome-wide DNA methylation can be profiled in single cells with bisulfite sequencing (scBS-seq)[49,50], which results in ~5% genomic coverage at a standard sequencing depth of 1 Gigabyte of bases or 3.3 million reads per cell. This translates to only around 0.25% overlapping CpG sites that are jointly detected in two given cells. How to reliably extract lineage information from this highly sparse single-cell DNA methylation data is a considerable computational challenge.

One way to address this challenge is to aggregate signals over very large genomic bins (for example, 100,000 bp), so that each bin has enough observed values. This is a standard practice in existing feature extraction methods for DNA methylation data[51,52]. However, averaging over a large bin would erase stochastic epimutations. We used a much smaller region (for example, 500 bp or single-CpG sites) to preserve lineage signals. The resulting cell-by-feature methylation rate matrix contains mostly missing values (Fig. 1b). Imputing these missing values is required for the standard single-cell analysis workflow. However, imputation could also degrade the stochastic lineage signals.

To circumvent this problem, we directly evaluated the pairwise similarity S_{ij} between two cells i and j by computing the Pearson correlation using just the entries observed in both cells. Pearson correlation is known to bias toward zero if the raw data suffers from measurement noises[53,54]. Specifically, $S_{ij}^{*} = Z_i S_{ij} Z_j$, where S_{ij}^{*} is the noise-free correlation, S_{ij} is the observed correlation that suffers from measurement noises and $Z_i \geq 1$ is the noise damping factor. A uniform Z_i across cells simply rescales the matrix S_{ij}. However, a heterogeneous Z_i would distort the similarity matrix, which would require correction. We developed an iterative approach to search for the optimal damping factor Z_i that minimizes the variation of the bias-corrected similarity matrix S_{ij}^{*}. The corrected similarity matrix S_{ij}^{*} can be used to infer the lineage phylogeny and visualize lineage similarity in low dimensions. We combine branch support values and similarity scores to jointly identify cells from the same clone. We will refer to this computational approach as MethylTree (methylation similarity-based lineage tree inference; Fig. 1b), and the inferred clones are named methyl clones.

We use the inferred lineage ordering to rank the similarity matrix. To evaluate the lineage ordering with ground-truth lineage or clone labels, we computed for each clone the largest fraction of cells that are grouped together in the lineage ordering, subtracted the randomness background and then reported the average score across all clones as the final lineage accuracy Q (Fig. 1c). Exact lineage ordering corresponds to $Q = 1$, while randomized ordering gives $Q \approx 0$.

MethylTree recovers cell lineages in homogeneous populations

The high epimutation rate on DNA methylation should enable us to resolve the entire division histories of human cells. To test this, we simulated clonal expansion from a single human cell under realistic conditions and obtained the single-cell methylation data with low genomic coverage. MethylTree correctly infers the entire division histories at 5% genomic coverage (Fig. 1d; $Q = 1$), or even just 1% (Fig. 1e). A higher genomic coverage is needed at a lower epimutation rate (Fig. 1f). MethylTree works robustly with more complex epimutation processes (Extended Data Fig. 1a,b). Therefore, it is possible to reconstruct the full division histories in human cells from highly sparse single-cell DNA methylation data.

Next, we carried out a single-cell colony expansion experiment with human embryonic kidney (HEK) 293T cells to test MethylTree on real data. In this experiment, we expanded each 293T cell from distant lineages into a large clone. To add lineage complexity, we also seeded single cells to generate subclones. We profiled multiple cells from each clone with scBS-seq at ~5% genomic coverage (Fig. 1g,h). Applying MethylTree to this dataset, the raw similarity matrix between cells, though distorted by measurement noises, can already resolve cells by their clonal identity (Fig. 1i; $Q = 0.84$). After applying our correction algorithm, the new similarity matrix shows block-wise structure within each clone and accurately resolves the lineages (Fig. 1j; $Q = 1$). The closer lineage relationship between P9_1 and P10_1 was also inferred correctly (Fig. 1k,l). When down-sampling the sequencing reads to just 2% genomic coverage per cell, MethylTree still achieved nearly exact lineage reconstruction (Fig. 1m), similar to our observations in simulation.

We systematically evaluated the performance of MethylTree on this 293T dataset. Simply using 29 million individual CpG sites, without binning, leads to exact lineage inference ($Q = 1$; Extended Data Fig. 1c). We observed accurate performance using the faster, region-based method, which works robustly across most parameter choices (Fig. 1n and Extended Data Fig. 1d–g). In addition, our approach is robust to technical variations like the heterogeneity of CpG coverage between cells (Extended Data Fig. 1h). We also found that Pearson correlation performs better than Euclidean and cosine similarity, and unweighted pair group method with arithmetic mean (UPGMA) infers a better lineage than neighbor-joining or FastME (Extended Data Fig. 1i,j)[55,56]. Finally, MethylTree exactly reconstructed lineages from another in-house generated clonal-expansion dataset of H9 human embryonic stem cells (Extended Data Fig. 2a–c; $Q = 1$), and also from a public dataset of human colorectal cancer[38] (Extended Data Fig. 2d–f; $Q = 1$). Overall, these results confirm that DNA methylation epimutations faithfully track cellular lineage histories and MethylTree robustly reconstructs lineages from sparse single-cell DNA methylation data.

Lineage reconstruction in a heterogeneous population

During development and differentiation, changes of DNA methylation not only come from 'stochastic' epimutations that reflect lineage

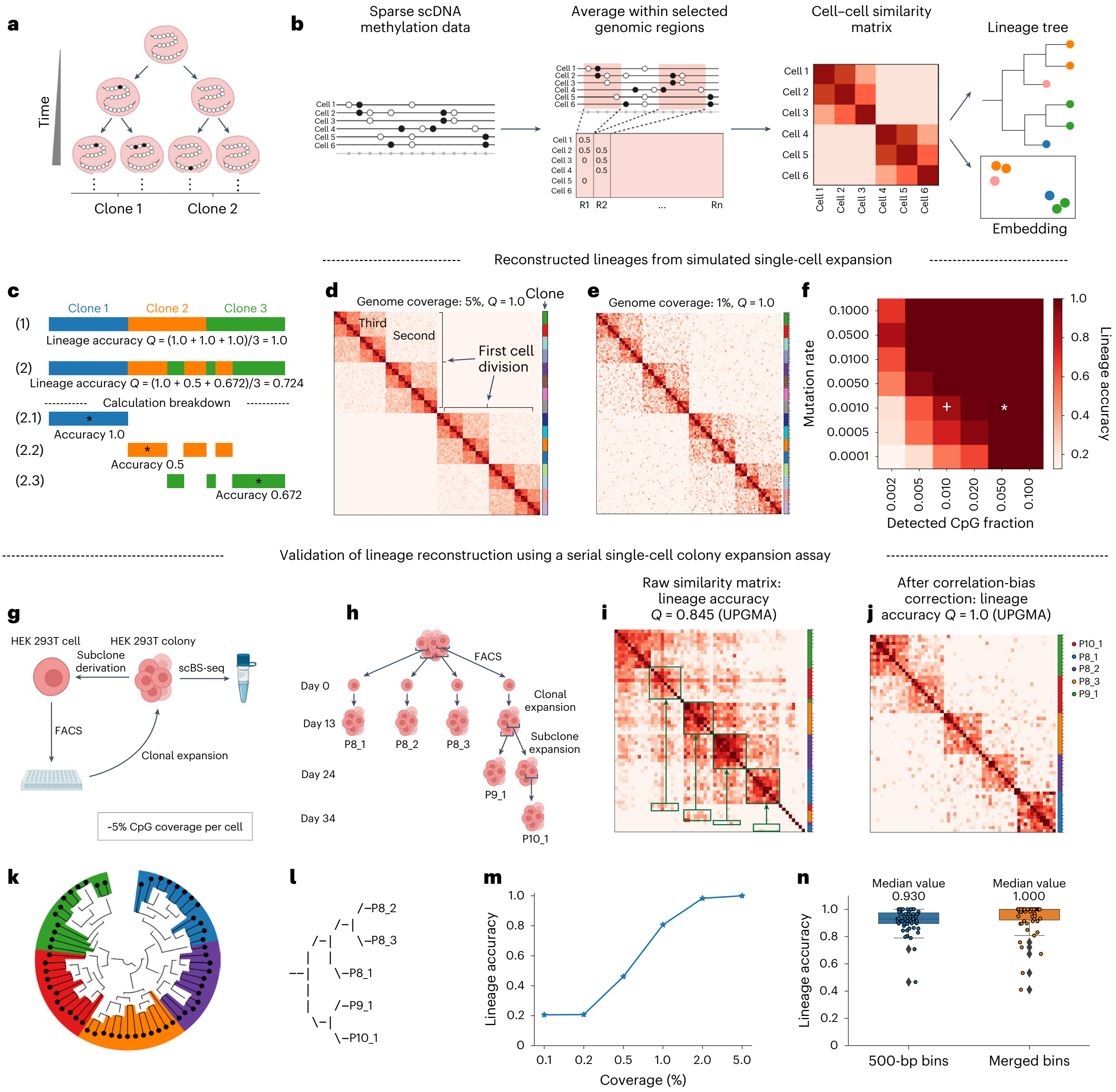


Fig. 1 | Epimutation-based lineage inference in a homogeneous population. **a**, Accumulation of epimutations during single-cell expansion. **b**, Schematic of MethylTree workflow. Sparse single-cell DNA methylation signals (left) can be aggregated in selected genomic regions and used to compute a cell–cell similarity matrix (middle), which can be used for lineage tree reconstruction and dimension reduction (right). **c**, Schematic of our lineage accuracy metric *Q*. **d**–**f**, Lineage reconstruction from simulated single-cell expansion shown in **a**. It generates 128 cells after seven divisions. **d**, Lineage-ordered heatmap of the methylation similarity for these cells, simulated at epimutation rate of 0.001 (unit, per CpG sites per division) and profiled at 5% genomic coverage. Color bar indicates synthetic clonal barcodes introduced at the 16-cell stage after the first four divisions. **e**, Same as **d**, but profiled at 1% genomic coverage. **f**, Heatmap of MethylTree lineage accuracy at various genomic coverage and different mutation rates. The results were averaged over ten independent simulations. In the heatmap, the asterisk * corresponds to **d** and the plus symbol + corresponds to **e**. **g**–**n**, Benchmark with 293T cells. **g**, Schematic of single-cell colony expansion with 293T cells. **h**, Expected lineage hierarchy among the profiled cells. **i**, Raw methylation similarity heatmap of the profiled 293T cells. Green arrows highlight correlations between cells from the same clone but not grouped together. **j**, Correlation-bias-corrected methylation similarity. **k**, Reconstructed lineage tree from the similarity matrix in **j**. In **i**–**k**, cells are colored by their clonal identity illustrated in **h**. **l**, Lineage hierarchy of all the five clones inferred from the coarse-grained methylation similarity matrix. **m**, Lineage accuracy when down-sampling to different genomic coverages. **n**, Lineage accuracy corresponding to all 55 choices of 500-bp genomic bins (Extended Data Fig. 1d) or merged bins (Extended Data Fig. 1e, f). Box plots show the median (50th percentile), the bounds of the box represent the interquartile range (25th to 75th percentile) and whiskers extend to 1.5 times the interquartile range. **a**,**b**,**g**,**h**, Created using BioRender.com.

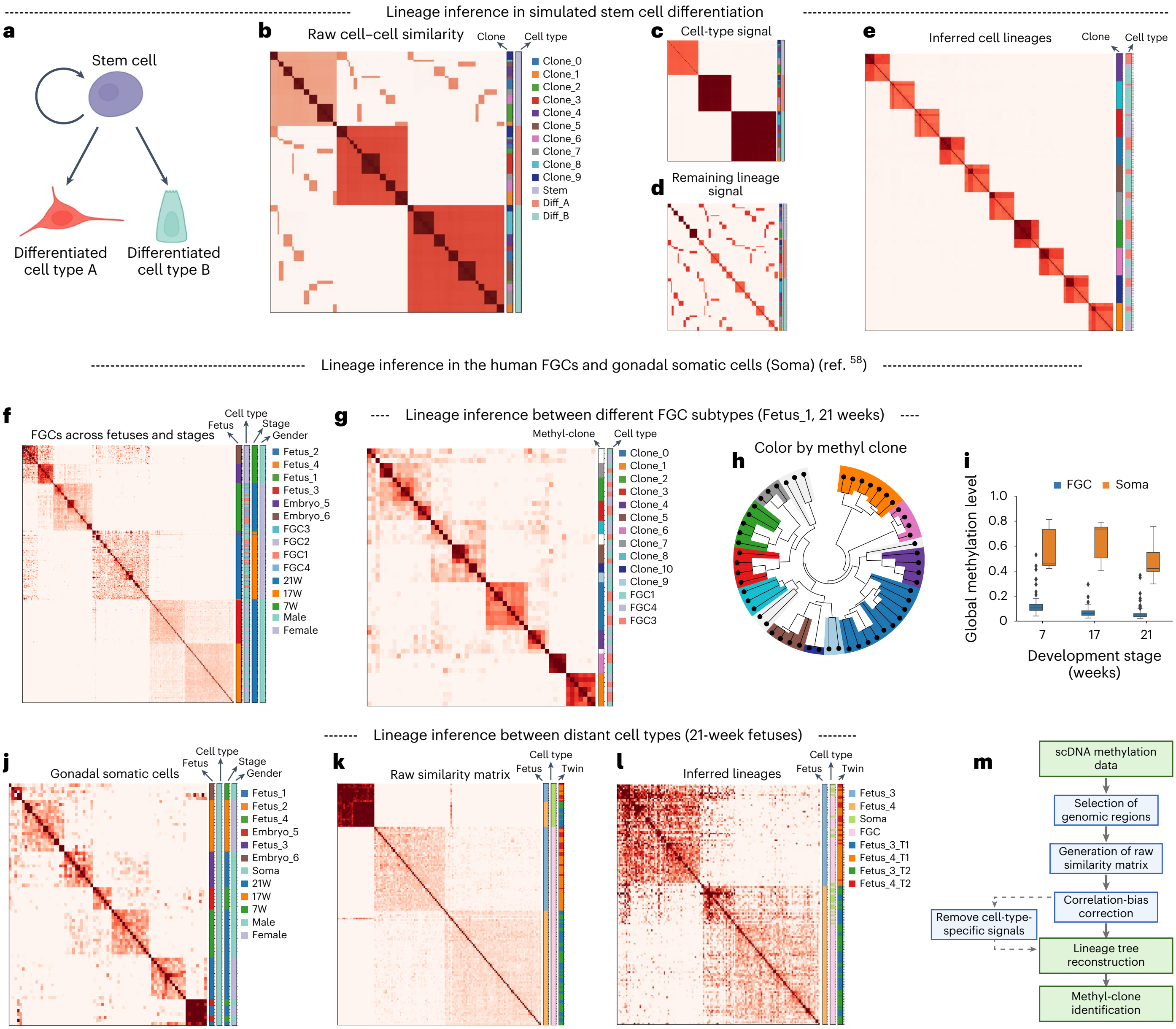


Fig. 2 | Lineage inference from DNA methylation profiles in a heterogeneous population. a, Schematic of stem-cell differentiation. **b**, Heatmap of the raw methylation similarity in this simulated data of stem-cell differentiation. The first column of the color bar indicates clone identity, and the second shows the cell type label. **c**, Similarity heatmap with just cell-type signals. **d**, Heatmap of the remaining lineage similarity after subtracting the cell-type signal **c** from the raw similarity in **b**. **e**, Heatmap of the inferred lineage similarity from the raw similarity matrix in **b**, after cell-type-aware transformation. **f**–**l**, Lineage reconstruction in the human fetal dataset generated by Li et al.[58]. **f**, Similarity heatmap of FGCs across fetuses and stages. **g**, Similarity heatmap for FGCs within fetus_1 from 21 weeks. **h**, Inferred lineages of FGCs from fetus_1, colored by methyl-clone ID. **i**, Box plot showing the average global DNA methylation level of somatic cells and FGCs over different stages. FGC cell number: 7 weeks, 97; 17 weeks, 164 and 21 weeks, 209. Somatic cell number: 7 weeks, 16; 17 weeks, 38 and 21 weeks, 53. See Fig. 1n for the box plot description. **j**, Similarity heatmap of gonadal somatic cells across fetuses and stages. **k**, Heatmap of the raw similarity matrix for cells from 21-week human embryos. **l**, Heatmap of lineage similarity after removing cell-type signal in **k**. **m**, Workflow of the MethylTree analysis. **a**, Created using BioRender.com.

histories, but also from 'deterministic' and cell-type-specific regulation. Therefore, when inferring lineages in a mixed population with cells having highly different methylation profiles, the cell-type-specific DNA methylation signal may dominate the similarity matrix and we need to remove these cell-type signals.

One approach involves identifying genomic regions or CpG sites that are lineage-specific and do not adopt cell-type-specific changes. However, this is technically highly challenging, since any given CpG site may be jointly observed in only a few cells in this sparse DNA methylation data. Besides, these lineage-specific regions or CpG sites need to be determined with ad hoc cutoffs that are hard to decide without knowing the actual lineages in advance. Here, we developed an alternative approach that first constructs the raw methylation similarity matrix S as before, but seeks to remove the cell-type signal afterward. We hypothesized that the raw matrix S is a linear combination of the cell-type similarity T and lineage similarity L, that is, $S = T + L$. The cell type similarity, by definition, only depends on the cell type identity and should be the same between two given cell types, regardless of lineage relationships of selected cells. On the other hand, the lineage signal L reports only lineage relationships and is reflected in the variations of the raw cell–cell similarity between two given cell types. Given the cell-type label, we performed cell-type-aware transformation on the raw similarity matrix to extract the lineage similarity matrix. This approach is parameter-free, and computationally efficient. However, this averaging approach may be less accurate if there are few lineages or cells in the same cell type.

To test this idea, we simulated ten clones, each starting from a stem cell that partially self-renews and also stochastically differentiates into cell type A and B (Fig. 2a). The DNA methylation profile on half of the genome is cell-type specific. As expected, the raw similarity matrix S (Fig. 2b) is indeed a superposition of the cell type signal T (Fig. 2c) and the lineage signal L (Fig. 2d). The raw similarity matrix is dominated by the cell-type signal, where cells are clustered according to their cell type identity. Applying our method, we successfully extracted the lineage similarity that groups cells according to their clonal identity (Fig. 2e). MethylTree works robustly with different proportions of lineage-specific CpG sites (Extended Data Fig. 3a–c).

Next, we tested our approach using published human fetal germ cell (FGC) datasets with different cell types. Previous studies have identified four FGC subtypes, each with distinct transcriptomic profiles[57]. Li et al. generated single-cell DNA methylation datasets for four FGC subtypes as well as gonadal somatic cells across different stages and from several human fetuses[58]. Using only FGCs, MethylTree accurately separated cells by their fetus origins, the lineage labels in this dataset (Fig. 2f; $Q = 1$). In fetus_1 of 21 weeks, the similarity matrix is not separated by FGC subtypes, but dominated by block-wise structures strongly indicative of cell lineages within this individual (Fig. 2g,h). We observed similar results in fetus_2 of 17 weeks (Extended Data Fig. 3d–f), suggesting that the cell-type differences between FGC subtypes are relatively small on DNA methylation as compared to their lineage differences.

To create a scenario where the cell-type signal would dominate the similarity matrix, we included the gonadal somatic cells jointly profiled in these datasets. These somatic cells have much higher global DNA methylation levels than those of FGCs, since FGCs undergo further global demethylation[48] (Fig. 2i). We confirmed that MethylTree also separated these somatic cells by their fetus origins (Fig. 2j; $Q = 1$). With both FGCs and somatic cells from 21 weeks, the similarity matrix is indeed dominated by cell type differences (Fig. 2k). After applying our method to remove the cell type signal, cells are clearly grouped by their fetus origins in the resulting lineage similarity matrix, despite that FGCs and somatic cells have different global DNA methylation levels (Fig. 2l and Extended Data Fig. 3g; $Q = 1$). This dataset consists of two pairs of twin fetuses. Each pair of twin fetuses are indistinguishable in the inferred lineage tree, yet these two twin pairs are clearly separated (Fig. 2l). We observed similar success of our approach for removing cell type signals when applying it to FGCs and somatic cells collected from two 7-week-old embryos (Extended Data Fig. 3h; $Q = 1$). In our updated analysis workflow, cell-type signals are removed only when it dominates the similarity matrix (Fig. 2m). Taken together, these analyses demonstrate that MethylTree can infer lineage histories even in a heterogeneous population with different cell types.

Benchmark with in vitro blood differentiation

Next, we test the feasibility of epimutation-based single-cell multi-omic lineage tracing in a complex differentiation system. We carried out an in vitro lineage-tracing experiment using blood progenitors extracted from an adult mouse and generated single-cell multi-omic readouts with ground-truth lineages. Specifically, we extracted Lin$^-$cKit$^+$Sca1$^-$ blood progenitors from a single mouse, introduced LARRY lineage barcodes in these cells by lentiviral infection[14], cultured them in vitro for 6 days in a media that supports cell expansion and pan-myeloid differentiation, and finally profiled these cells with a modified Camellia-seq protocol that we developed recently[13] to obtain LARRY lineage barcode, transcriptome and DNA methylome simultaneously in single cells (Fig. 3a,b). Using transcriptome, we identified eight cell types: megakaryocytes, erythrocytes, basophils, mast cells, eosinophils, neutrophils, neutrophil-like monocytes and dendritic-like monocytes (Fig. 3c and Extended Data Fig. 4a,b). We also observed 52 LARRY clones with ≥2 cells, and 21 were found in ≥2 cell types (Fig. 3b,d,e and Extended Data Fig. 4c).

Applying MethylTree, all 52 multi-cell LARRY clones were correctly identified in this similarity matrix (Fig. 3f; $Q = 1$), including 21 clones consisting of more than one cell type (Fig. 3g). This is even before cell-type-signal removal. Similar accuracy is achieved for both the raw similarity matrix with or without correlation-bias correction, and after removing cell-type signals (Fig. 3h and Extended Data Fig. 4d). We expected lineage signals to be enriched at small genomic bins due to sporadic nature of epimutations. To test this, we used fixed nonoverlapping bins for feature extraction and computed the lineage accuracy at different bin sizes. Indeed, the lineage accuracy is near 1 across different choices below 1,000 bp, but drops to only ~0.2 at a bin size of 100,000 bp, which is the commonly used bin size in DNA methylation data analysis[51,52] (Extended Data Fig. 4e). To test the robustness of the result over data sparsity, we also down-sampled the fastq reads to different genomic coverages and found that a median lineage accuracy of $Q = 0.92$ can be achieved at only 3.25% genomic coverage (Extended Data Fig. 4f).

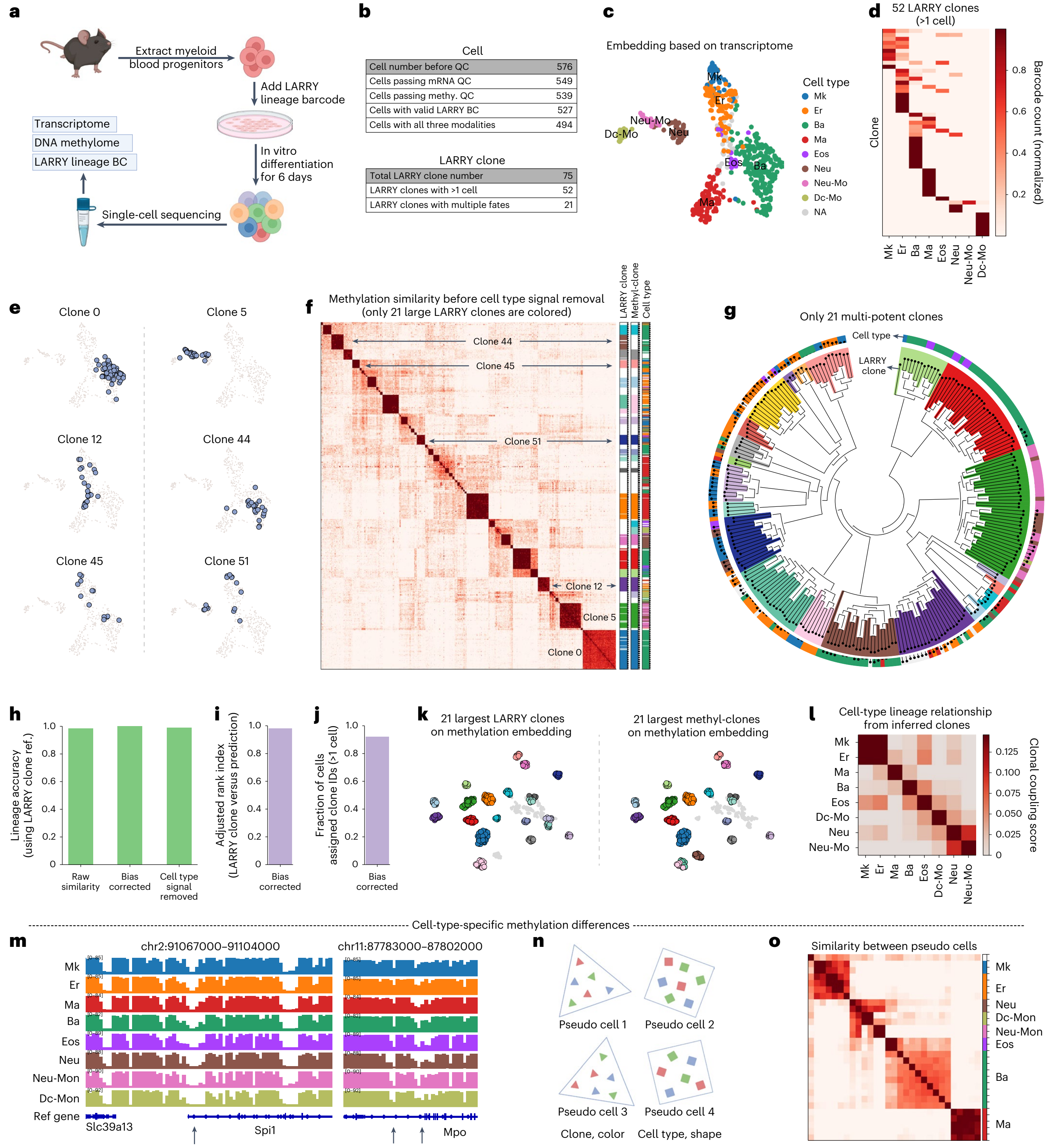
Encouraged by this striking performance, we next test whether DNA methylation can serve as reliable lineage barcodes and use them to infer the differentiation hierarchy of hematopoiesis. We identified 48 multi-cell methyl clones, which matched the observed LARRY clones (adjusted rank index of 0.98) (Fig. 3f,i,k). In total, 498 cells, or 92% of all cells, were among these multi-cell methyl clones (Fig. 3j). Applying CoSpar[59] on these 48 multi-cell methyl clones, we computed clonal coupling scores between different cell types (Fig. 3l). The resulting lineage relationships between cell types match that computed using the 52 LARRY clones (Extended Data Fig. 4g), and agree with previous reports[14,18,59].

By aggregating sparse single-cell DNA methylation measurements from each cell type into a pseudobulk dataset, we found

Fig. 3 | Single-cell multi-omic lineage tracing in in vitro blood differentiation from mouse. a, Schematic of our single-cell multi-omic lineage-tracing experiment with hematopoietic cells derived from a single mouse. **b**, Summary of data statistics. **c**, UMAP embedding with single-cell transcriptome. Mk, megakaryocytes; Er, erythrocytes; Ba, basophils; Ma, mast cells; Eos, eosinophils; Neu, neutrophils; Neu-Mon, neutrophil-like monocytes; Dc-Mon, dendritic-like monocytes. **d**, Heatmap showing the cell-type composition in each observed clone. Only the 52 LARRY clones having more than one cell are shown. **e**, Six selected clones on the transcriptomic embedding. **f**, Correlation-bias-corrected methylation similarity matrix, without removing cell-type signals. The color bar on the right indicates the LARRY clones, methyl clones and cell type. Due to limited colors, only 21 LARRY or methyl clones are shown. Cell-type color is the same as in **c**. Clones highlighted in **e** are indicated on this heatmap. **g**, Inferred lineage phylogeny for cells associated with multipotent clones. **h**, Bar plot of lineage accuracies from raw or correlation-bias-corrected similarity, or after cell-type-aware transformation. **i**, Adjusted rank index between the predicted methyl clones from the correlation-bias-corrected similarity matrix (same as **f**) and the observed LARRY clones. **j**, Fraction of cells included in the multi-cell methyl clones. **k**, 21 largest LARRY or methyl clones on the methylation embedding. **l**, Clonal coupling score between different cell types computed with all multi-cell methyl clones. **m**, Pseudobulk DNA methylation profiles for all identified cell types on two selected genomic regions. Regions with cell-type-specific differences are indicated by arrows. **n**, Schematic of generating 'pseudo' cells by aggregating multiple single-cell profiles of the same type (indicated by shape) but from different lineages (indicated by color). **o**, Heatmap of methylation similarity between 29 pseudo cells. BC, barcode; QC, quality control. **a**, Created using BioRender.com.

cell-type-specific DNA methylation patterns near key marker genes of blood cells (Fig. 3m), yet similar methylation in other genomic regions (Extended Data Fig. 4h). To enrich for cell-type-specific signals in this DNA methylation dataset, we generated 29 'pseudo' cells, each aggregated from ~18 single cells of the same type but from roughly eight different clones, thus averaging out lineage signals (Fig. 3n). The resulting similarity matrix clustered together pseudo cells of the same type (Fig. 3o) and reflected the differentiation hierarchy between different cell types (Fig. 3l,o). We found that lineage-specific CpG sites are depleted in CpG islands and gene body, but enriched in solo-WCGW sites (Extended Data Fig. 4i), consistent with earlier reports that solo-WCGW sites are prone to hypomethylation[33,60].

We observed similar successes with an in vitro human blood differentiation assay. Here, CD34$^+$ cells were sorted from human umbilical cord blood, transfected with LARRY, cultured in a media that supports pan-myeloid differentiation and profiled with Camellia-seq on day 13 (Fig. 4a). We identified five cell types in this data from the single-cell transcriptome, and 20 LARRY clones with more than one cell, among which nine of them occupying multiple cell types (Fig. 4b–d and Extended Data Fig. 5a). MethylTree reconstructed the human blood

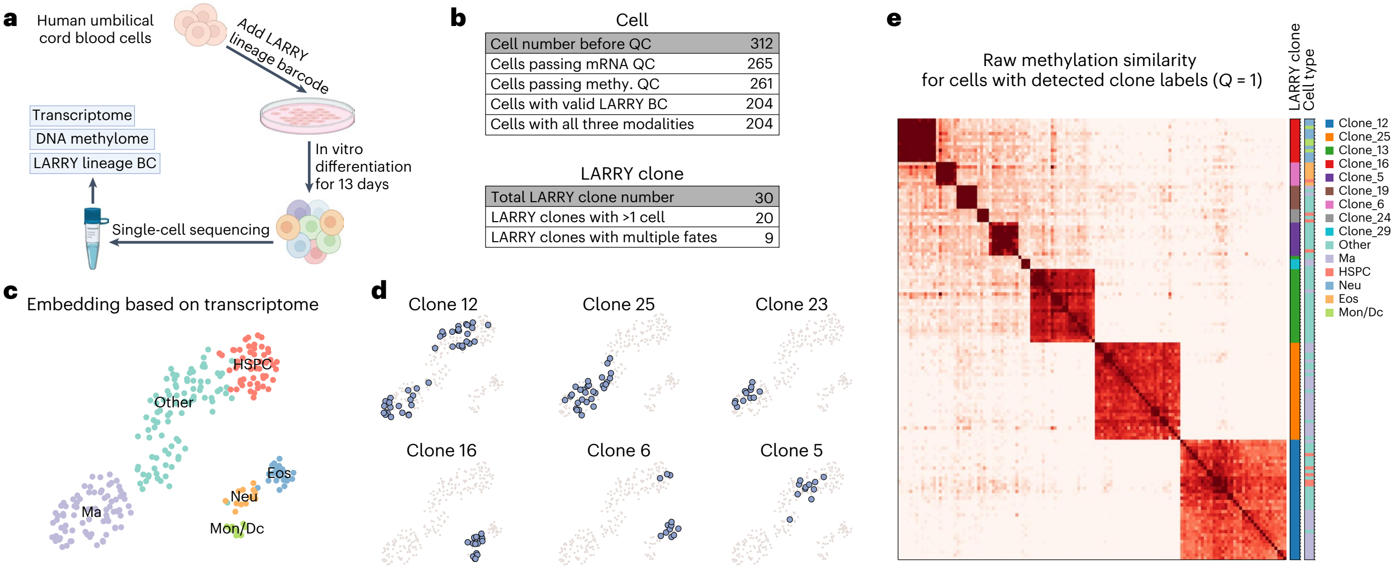


Cell

Cell number before QC	576
Cells passing mRNA QC	549
Cells passing methy. QC	539
Cells with valid LARRY BC	527
Cells with all three modalities	494

LARRY clone

Total LARRY clone number	75
LARRY clones with >1 cell	52
LARRY clones with multiple fates	21



Cell	
Cell number before QC	312
Cells passing mRNA QC	265
Cells passing methy. QC	261
Cells with valid LARRY BC	204
Cells with all three modalities	204

LARRY clone	
Total LARRY clone number	30
LARRY clones with >1 cell	20
LARRY clones with multiple fates	9

Fig. 4 | Lineage inference from in vitro differentiation of human CD34$^+$ cells. **a**, Schematic of our experimental design. **b**, Summary of data statistics. **c**, UMAP embedding and cell-type annotation with single-cell transcriptome. HSPC, hematopoietic stem and progenitor cells. Otherwise, same as Fig. 3c. **d**, Representative clones illustrated on the transcriptomic embedding. **e**, Correlation-bias-corrected methylation similarity matrix for cells with detected clone labels, without removing cell-type signals. The color bar on the right indicates the LARRY clones and cell type. The similarity matrix is generated using all the ~29 million single-CpG sites as features. **a**, Created using BioRender.com.

cell lineages with 100% accuracy, also without removing the cell type information (Fig. 4e and Extended Data Fig. 5b,c). Together, these results establish that MethylTree can be combined with single-cell multi-omic profiling to enable high-resolution, noninvasive lineage tracing in complex differentiation systems.

Early cell fate choice in human embryonic development

In mammals, cells within a developing embryo have indistinguishable morphologies and remain developmentally plastic until they become inner cell mass (ICM) or trophectoderm (Fig. 5a)[61]. This has led to the traditional view that the first fate decision toward ICM or trophectoderm is a random process (Fig. 5b)[62]. However, recent studies have revealed molecular heterogeneity in the two- to four-cell-stage embryos that influences the first cell fate decision in mice[63,64], which leads to an alternative model of early commitment (Fig. 5c). These studies require genetic manipulation to label specific genes, which is not applicable to human embryos. Below, we applied our high-resolution method based on epimutations to reveal the first fate decision in native human embryos.

In early embryo development, cells undergo drastic global demethylation and remethylation[48], which could erase shared epimutations between sister cells. To check whether our approach still works in this context, we applied MethylTree to four-cell-stage cells collected from six mouse embryos by Guo et al.[65]. MethylTree accurately identified their lineage relationships by grouping cells according to their embryonic origins (Fig. 5d,e; $Q = 1$). MethylTree performed equally well when applied to mouse cells from other stages or to human embryonic cells from previous publications (Fig. 5f,g and Extended Data Fig. 6a–d)[65,66]. At the four-cell stage, the four detected cells are correctly separated into two groups, each with exactly two cells, consistent with their division histories (Fig. 5d). Together, these results raise the possibility that epimutations on DNA methylation could be used to trace lineage histories within the same developing embryo.

Next, we investigated the first fate choice encoded in the cell lineages of a single human embryo, using the above human embryo datasets collected by Qi et al.[66]. We selected cells from day 5 or day 6 human embryos, which have just adopted the fate of either trophectoderm or ICM. In the methylation similarity of day 6 cells from embryo E49, cells are separated first into two major groups and further into four subgroups (Fig. 5h). This reflects the two- and four-cell stage in the division history, as we have seen in our simulation of single-cell expansion (Fig. 1d). The inferred lineage tree is robust and has high branch support values from bootstrapping (Fig. 5i). Inferred descendants from one four-cell-stage cell all adopted ICM fate, although another two putative four-cell-stage cells also contributed to ICM (Fig. 5h,i). In another four embryos, we also observed strong but variable early commitment toward ICM at the four-cell stage (Fig. 5j,k and Extended Data Fig. 6e–g). We also observed early commitment toward trophectoderm. Therefore, our analyses suggest a model of stochastic early commitment toward ICM or trophectoderm at the four-cell stage in early human embryo development (Fig. 5l).

Counting clones of HSCs in mice

In mice, HSCs arise through endothelial-to-hematopoietic transition (EHT) within the aorta-gonad-mesonephros (AGM) region in the embryo. HSCs start to migrate to the fetal liver at around E11.5 and expand rapidly there, before colonizing the bone marrow at around the time of birth (Fig. 6a)[67]. These EHT-derived HSC clones sustain blood production for life. The clonal diversity of HSCs is foundational to blood homeostasis and is also relevant in the aging process. Transplantation assays estimated only 70 HSC precursors at E11.5 in mice[68]. Clonal tracking based on low-capacity fluorescent labeling estimated 30 HSC clones in zebrafish[69] and hundreds of blood progenitors in mice[70]. So far, the exact number of HSC clones in mice remains unknown due to the lack of reliable and high-capacity lineage tracing approaches to directly count HSC clones in vivo. We have previously generated a dataset where we barcoded HSCs in DARLIN mice at E10, and profiled HSCs with Camellia-seq at either E15.5 or adult stage (Fig. 6a)[13]. Below, we applied MethylTree to this dataset to estimate the number of HSC clones in mice.

Applying MethylTree to HSCs from adult mouse LL731, we reconstructed cell lineages that agreed accurately with the ground-truth lineage labels from DARLIN barcodes (Fig. 6b,c and Extended Data Fig. 7a; $Q = 0.90$). We observed that each inferred methyl clone corresponds roughly to one DARLIN barcode introduced at E10 (Fig. 6b–d; ARI = 0.87), when HSCs just begin to emerge. This suggests that the

https://doi.org/10.1038/s41592-024-02567-1

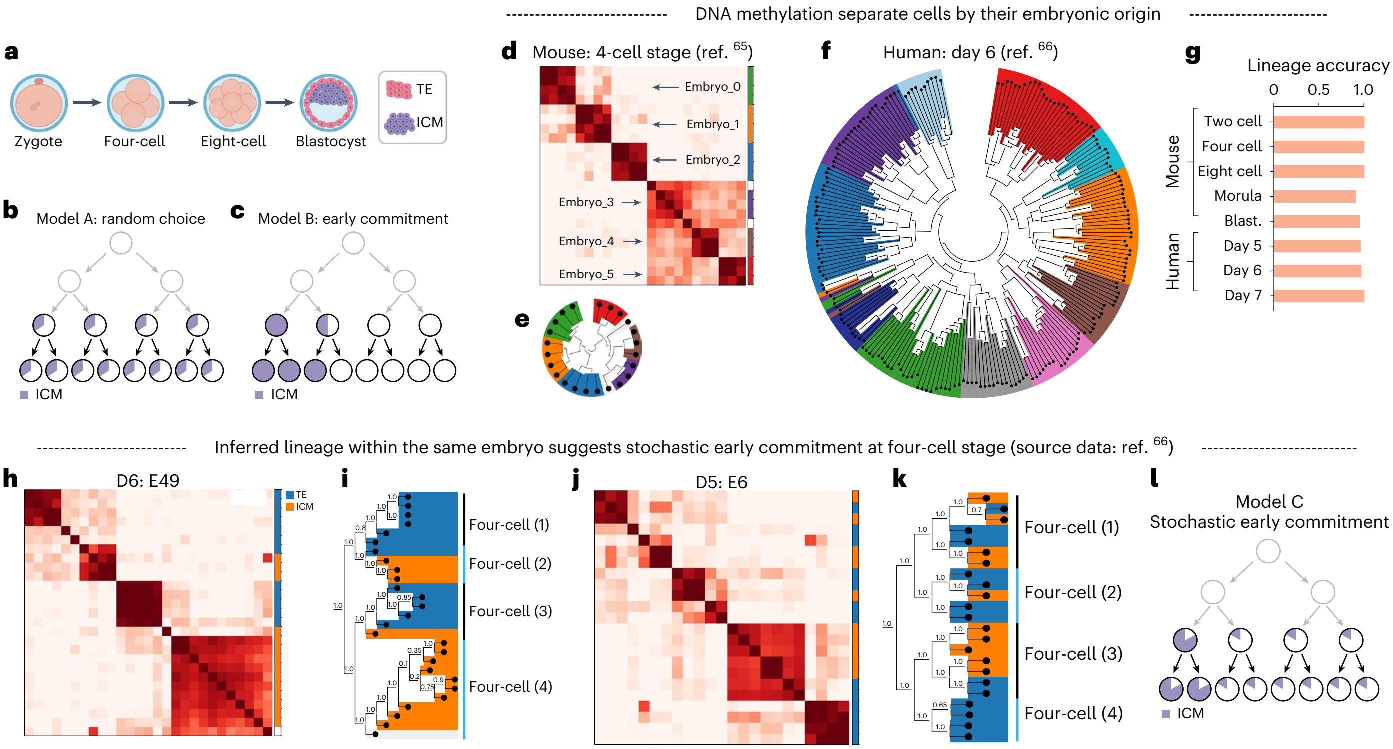


Fig. 5 | Study of the first cell fate decision in human embryonic development. **a**, Schematic of the first fate decision toward either ICM or trophectoderm (TE) in early embryonic development. **b**, A model of random fate choices. Fate choices starting from the four-cell stage are illustrated. **c**, A model of the early commitment. Otherwise, same as **b**. **d**, Heatmap of DNA methylation similarity for cells from four-cell-stage mouse embryos, using public data from Guo et al.[65]. The color bar indicates the embryonic origin of each cell, applicable to **e**, **f** and Extended Data Fig. 6. **e**, Reconstructed lineage tree corresponding to **d**. **f**, Reconstructed lineage tree for cells from day 6 human embryos using public data from Qi et al.[66]. **g**, Accuracy of MethylTree-reconstructed lineages for cells from different embryonic stages in mice and humans. Lineage accuracy is evaluated in terms of separating cells by their embryonic origins. **h**, Heatmap of DNA methylation similarity for cells from day 6 human embryo E49. The color bar indicates whether this cell adopts the TE or ICM fate on day 6. The same color scheme is used in **i**–**k** and Extended Data Fig. 6e–g. **i**, Reconstructed lineage tree corresponding to **h**. The support of each putative lineage branch is indicated accordingly. Putative descendants from the same four-cell-stage cell are indicated with a vertical bar on the right. **j**,**k**, Same as **h** (**j**) and **i** (**k**), but for day 5 human embryo E6. **l**, A proposed model of stochastic early commitment. **a**, Created using BioRender.com.

number of methylation-defined HSC clones corresponds to de novo HSCs derived from EHT.

Due to sampling only 206 cells, most methyl clones are small in the LL731 dataset, with 82 of them having just one cell (that is, singleton), accounting for 40% of all the 206 observed cells (Fig. 6e). To estimate the actual clone number from these partial measurements, we simulated the process of sampling just 206 cells from a pool of HSC clones that have empirically observed clone size distribution. At 270 starting HSC clones, this sampling scheme produced exactly 40% cells that are singleton (Fig. 6f). We obtained consistent estimates of ~250 HSC clones across all three mice from both E15.5 and week 11 (Fig. 6g and Extended Data Fig. 7b–e). We further supported this estimate by using a bulk DARLIN dataset from our previous study[13]. There, a DARLIN mouse embryo was induced for lineage barcoding at E10, and HSCs were profiled with bulk sequencing at week 4. After correcting for low editing efficiency, we observed 312 DARLIN barcodes among HSCs (Fig. 6h), which is comparable with our estimate. The DARLIN estimate is likely inflated by the presence of background editing due to stochastic, leaky expression of the Cas9-TdT protein. Together, we demonstrated that epimutations on DNA methylation enable reliable estimation of the EHT-derived HSC clone number in mice.

Discussion

Here, we developed MethylTree, a generic lineage-tracing tool based on frequent epimutations on single-cell DNA methylation. It achieved high-resolution, noninvasive single-cell lineage tracing across multiple cell types from key developmental stages, including early embryonic stage, fetal stage and adult stage, thus covering both dynamic and static periods of global methylation (Extended Data Fig. 8a). MethylTree also works robustly with a lineage accuracy near 100% for a population with either similar or distinct cell types from mice and humans (Extended Data Fig. 8b). A genomic coverage of just 2% may be sufficient for lineage reconstruction with MethylTree, although 5% (~3 million reads per cell with scBS-seq) is better for robustness. Since the epimutation rate is ~0.001 per CpG site per division, an epimutation on a given CpG site could be stable over hundreds of cell divisions. Therefore, epimutations in principle could track cell lineages throughout the lifespan of an individual.

We observed a robust performance across different choices of genomic features in single-cell DNA methylation data. Standard practices use very large bins such as 100,000 bp (refs. 51,52). However, this may average out lineage signals and lead to a poor result ($Q = 0.2$, Extended Data Fig. 4e). Simply using all the ~30 million CpG sites, although computationally expensive, gives superior accuracy (Fig. 4e and Extended Data Figs. 1c and 8c). When averaging over a genomic window around 500 bp, certain region subsets perform better than using all 500-bp windows (Extended Data Fig. 8c). This is likely due to enrichment of epimutations within these selected regions, which could be reused in analyzing other data from similar systems, as we did in this study (Supplementary Table 1). Predicting the most informative set of genomic regions for a given system would be an interesting future direction.

In this study, we demonstrated the feasibility of epimutation-based single-cell multi-omic lineage tracing in complex differentiation

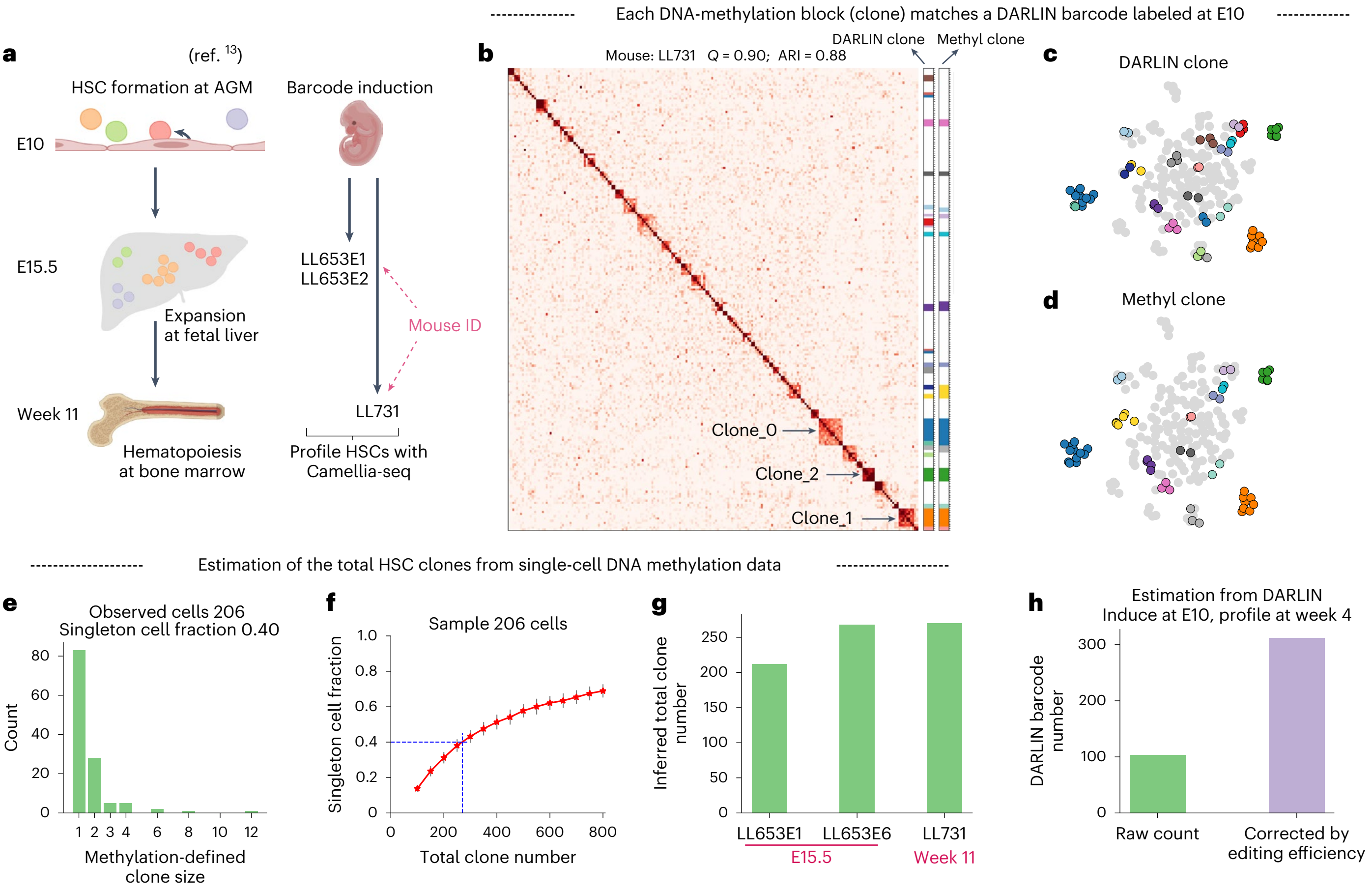


Fig. 6 | Estimation of HSC clone number in mice. **a**, HSC development in mice (left) and experimental design by Li et al. (right)[13]. **b**, Methylation similarity heatmap for HSCs from mouse LL731. The color bar indicates the observed DARLIN clone or methyl-clone ID of each cell. The three largest clones are highlighted. Due to limited colors, only the top 21 largest DARLIN clones and their overlapped methyl clones are shown. **c**, DARLIN clones from **b** on the methylation embedding. **d**, Methyl clones from **b** on the methylation embedding. Only methyl clones marked by DARLIN barcodes are shown. **e**, Histogram of the methylation-defined clone sizes in the LL731 dataset. The singleton cell fraction reports the fraction of cells with no observed sister cells from the same clone. **f**, Relationship between the hypothetic total clone number in an HSC pool and the singleton fraction after only sampling 206 HSCs. The blue dashed line indicates the total HSC clone number that corresponds to exactly 40% singleton fraction seen in the LL731 dataset. Error bars are computed from 100 realizations of random sampling at each hypothesized total clone number. **g**, Bar plot of the inferred HSC clone number across all three mice from either E15.5 and week 11. **h**, Bar plot of the DARLIN barcodes among HSCs from a mouse induced at E10 and profiled at week 4. Both the raw barcode count as well as the number corrected by editing efficiency are shown. **a**, Created using BioRender.com.

systems. We generated single-cell multi-omic datasets for in vitro differentiated hematopoietic cells for both mice and humans, each derived from a single individual. In both datasets, MethylTree achieved exact lineage inference ($Q = 1$). Furthermore, integrating methyl clones with transcriptome-based cell-state annotation recapitulated the differentiation hierarchy in hematopoiesis. Last, in these in vitro systems, we found that the genome-wide lineage signal is stronger than the cell-type signal on DNA methylation (Figs. 3f and 4e), which echoes our observation with FGCs (Fig. 2g). However, if the profiled population involve highly distinct cell types, single-cell phenotypic data are needed for removing cell-type signals in MethylTree inference. Such phenotypic data can be obtained from joint single-cell RNA sequencing, or in some cases with fluorescence-activated cell sorting (FACS).

We have applied our method to study fate choice in early embryonic development and the clonal diversity of blood. The first problem requires resolving individual cell divisions that occurs during just a few days. We found stochastic yet early fate commitment toward ICM or trophectoderm at the four-cell stage in humans, which is consistent with the report of early fate bias in mice[63,64]. In mouse HSCs, the methyl clones match the DARLIN barcodes introduced at E10, when HSCs just begin to emerge through EHT. One of the explanations is that only a small fraction of the endothelial cells undergo EHT and become HSCs[71,72], which leads to an epigenetic bottleneck that increases the methylation differences among those de novo HSCs. We used these stark epigenetic differences to estimate HSC clone number in native mice, and concluded with ~250 EHT-derived HSC clones.

We systematically compared our approach with other lineage-tracing methods in humans, which is summarized in Extended Data Fig. 9. As mentioned in the introduction, existing methods suffer from limited temporal resolution and lineage accuracy. Besides, calling somatic mutations with whole-genome DNA sequencing lacks state information of individual cells, and costs hundreds of dollars per cell[25,26]. In contrast, our method has unprecedented temporal resolution that distinguishes individual cell divisions, and achieves nearly exact lineage inference across broad contexts. Our epimutation-based lineage tracing is compatible with other modalities such as transcriptome and chromatin accessibility, as demonstrated in Camellia-seq[13] and snmCAT-seq[73]. Besides, single-cell DNA methylation (along with other modalities) can be profiled in thousands of cells per week using advanced barcoding approaches[52,74] or with robot automation[73]. With these high-throughput methods, the cost can be reduced dramatically, leading to just a few dollars per cell for our recommended sequencing depth of 5% genomic coverage. Our method also avoids the problem of inefficient barcode labeling and capture, which is typical in engineered lineage-tracing model systems[2,3,6,13,75]. Taken together, our approach provides a high-resolution, noninvasive, multi-omic and more affordable method for investigating relationships and molecular mechanisms of diverse biological processes in humans and beyond.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-024-02567-1.

References

1. Kretzschmar, K. & Watt, F. M. Lineage tracing. *Cell* **148**, 33–45 (2012).
2. Sun, J. et al. Clonal dynamics of native haematopoiesis. *Nature* **514**, 322–327 (2014).
3. Pei, W. et al. Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature* **548**, 456–460 (2017).
4. McKenna, A. et al. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* **353**, aaf7907 (2016).
5. Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J. & van Oudenaarden, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108–112 (2018).
6. Bowling, S. et al. An engineered CRISPR-Cas9 mouse line for simultaneous readout of lineage histories and gene expression profiles in single cells. *Cell* **181**, 1410–1422.e27 (2020).
7. Chan, M. M. et al. Molecular recording of mammalian embryogenesis. *Nature* **570**, 77–82 (2019).
8. Kalhor, R. et al. Developmental barcoding of whole mouse via homing CRISPR. *Science* **361**, eaat9804 (2018).
9. Xie, L. et al. Comprehensive spatiotemporal mapping of single-cell lineages in developing mouse brain by CRISPR-based barcoding. *Nat. Methods* https://doi.org/10.1038/s41592-023-01947-3 (2023).
10. Liu, K. et al. Mapping single-cell-resolution cell phylogeny reveals cell population dynamics during organ development. *Nat. Methods* https://doi.org/10.1038/s41592-021-01325-x (2021).
11. Choi, J. et al. A time-resolved, multi-symbol molecular recorder via sequential genome editing. *Nature* **608**, 98–107 (2022).
12. Wagner, D. E. & Klein, A. M. Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-020-0223-2 (2020).
13. Li, L. et al. A mouse model with high clonal barcode diversity for joint lineage, transcriptomic, and epigenomic profiling in single cells. *Cell* https://doi.org/10.1016/j.cell.2023.09.019 (2023).
14. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, eaaw3381 (2020).
15. Rodriguez-Fraticelli, A. E. et al. Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. *Nature* https://doi.org/10.1038/s41586-020-2503-6 (2020).
16. Rodriguez-Fraticelli, A. E. et al. Clonal analysis of lineage fate in native haematopoiesis. *Nature* **553**, 212–216 (2018).
17. Patel, S. H. et al. Lifelong multilineage contribution by embryonic-born blood progenitors. *Nature* **606**, 747–753 (2022).
18. Jindal, K. et al. Single-cell lineage capture across genomic modalities with CellTag-multi reveals fate-specific gene regulatory changes. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01931-4 (2023).
19. Biddy, B. A. et al. Single-cell mapping of lineage and identity in direct reprogramming. *Nature* **564**, 219–224 (2018).
20. Yang, D. et al. Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. *Cell* **185**, 1905–1923.e25 (2022).
21. Quinn, J. J. et al. Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science* **371**, eabc1944 (2021).
22. Goyal, Y. et al. Diverse clonal fates emerge upon drug treatment of homogeneous cancer cells. *Nature* https://doi.org/10.1038/s41586-023-06342-8 (2023).
23. Woodworth, M. B., Girskis, K. M. & Walsh, C. A. Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nat. Rev. Genet.* **18**, 230–244 (2017).
24. Abyzov, A. & Vaccarino, F. M. Cell lineage tracing and cellular diversity in humans. *Annu. Rev. Genomics Hum. Genet.* **21**, 101–116 (2020).
25. Mitchell, E. et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022).
26. Fabre, M. A. et al. The longitudinal dynamics and natural history of clonal haematopoiesis. *Nature* **606**, 335–342 (2022).
27. Ludwig, L. S. et al. Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* **176**, 1325–1339.e22 (2019).
28. Campbell, P. et al. Mitochondrial mutation, drift and selection during human development and ageing. Preprint at *Research Square* https://doi.org/10.21203/rs.3.rs-3083262/v1 (2023).
29. Wang, X. et al. Clonal expansion dictates the efficacy of mitochondrial lineage tracing in single cells. Preprint at *bioRxiv* https://doi.org/10.1101/2024.05.15.594338 (2024).
30. Weng, C. et al. Deciphering cell states and genealogies of human hematopoiesis. *Nature* https://doi.org/10.1038/s41586-024-07066-z (2024).
31. Lynch, M. Evolution of the mutation rate. *Trends Genet.* **26**, 345–352 (2010).
32. Landan, G. et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat. Genet.* **44**, 1207–1214 (2012).
33. Wang, Q. et al. Imprecise DNMT1 activity coupled with neighbor-guided correction enables robust yet flexible epigenetic inheritance. *Nat. Genet.* **52**, 828–839 (2020).
34. Xie, H. et al. Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Res.* **39**, 4099–4108 (2011).
35. Ushijima, T. et al. Fidelity of the methylation pattern and its variation in the genome. *Genome Res.* **13**, 868–874 (2003).
36. Brocks, D. et al. Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Rep.* **8**, 798–806 (2014).
37. Mazor, T. et al. DNA methylation and somatic mutations converge on the cell cycle and define similar evolutionary histories in brain tumors. *Cancer Cell* **28**, 307–317 (2015).
38. Bian, S. et al. Single-cell multiomics sequencing and analyses of human colorectal cancer. *Science* **362**, 1060–1063 (2018).
39. Bian, S. et al. Integrative single-cell multiomics analyses dissect molecular signatures of intratumoral heterogeneities and differentiation states of human gastric cancer. *Natl Sci. Rev.* **10**, nwad094 (2023).
40. Wang, Y. et al. Single-cell dissection of the multiomic landscape of high-grade serous ovarian cancer. *Cancer Res.* **82**, 3903–3916 (2022).
41. Shipony, Z. et al. Dynamic and static maintenance of epigenetic memory in pluripotent and somatic cells. *Nature* **513**, 115–119 (2014).
42. Yatabe, Y., Tavaré, S. & Shibata, D. Investigating stem cells in human colon by using methylation patterns. *Proc. Natl Acad. Sci. USA* **98**, 10839–10844 (2001).
43. Gabbutt, C. et al. Fluctuating methylation clocks for cell lineage tracing at high temporal resolution in human tissues. *Nat. Biotechnol.* **40**, 720–730 (2022).

44. Gaiti, F. et al. Epigenetic evolution and lineage histories of chronic lymphocytic leukaemia. *Nature* **569**, 576–580 (2019).
45. Chaligne, R. et al. Epigenetic encoding, heritability and plasticity of glioma transcriptional cell states. *Nat. Genet.* **53**, 1469–1479 (2021).
46. Liu, Y. et al. Single-cell methylation sequencing data reveal succinct metastatic migration histories and tumor progression models. *Genome Res.* **33**, 1089–1100 (2023).
47. Grady, W. M. & Carethers, J. M. Genomic and epigenetic instability in colorectal cancer pathogenesis. *Gastroenterology* **135**, 1079–1099 (2008).
48. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* **20**, 590–607 (2019).
49. Smallwood, S. A. et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat. Methods* **11**, 817–820 (2014).
50. Guo, H. et al. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome Res.* **23**, 2126–2135 (2013).
51. Liu, H. et al. Single-cell DNA methylome and 3D multi-omic atlas of the adult mouse brain. *Nature* **624**, 366–377 (2023).
52. Bai, D. et al. Simultaneous single-cell analysis of 5mC and 5hmC with SIMPLE-seq. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-024-02148-9 (2024).
53. Adolph, S. C. & Hardin, J. S. Estimating phenotypic correlations: correcting for bias due to intraindividual variability. *Funct. Ecol.* **21**, 178–184 (2007).
54. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **15**, 72–101 (1904).
55. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
56. Lefort, V., Desper, R. & Gascuel, O. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* **32**, 2798–2800 (2015).
57. Li, L. et al. Single-cell RNA-seq analysis maps development of human germline cells and gonadal niche interactions. *Cell Stem Cell* **20**, 858–873.e4 (2017).
58. Li, L. et al. Dissecting the epigenomic dynamics of human fetal germ cell development at single-cell resolution. *Cell Res.* https://doi.org/10.1038/s41422-020-00401-9 (2020).
59. Wang, S.-W., Herriges, M. J., Hurley, K., Kotton, D. N. & Klein, A. M. CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-022-01209-1 (2022).
60. Zhou, W. et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.* **50**, 591–602 (2018).
61. Rossant, J. & Tam, P. P. L. Blastocyst lineage formation, early embryonic asymmetries and axis patterning in the mouse. *Development* **136**, 701–713 (2009).
62. Dietrich, J.-E. & Hiiragi, T. Stochastic patterning in the mouse pre-implantation embryo. *Development* **134**, 4219–4231 (2007).
63. Goolam, M. et al. Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. *Cell* **165**, 61–74 (2016).
64. Wang, J. et al. Asymmetric expression of LincGET biases cell fate in two-cell mouse embryos. *Cell* **175**, 1887–1901.e18 (2018).
65. Guo, F. et al. Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res.* **27**, 967–988 (2017).
66. Qi, S. et al. Single-cell multiomics analyses of spindle-transferred human embryos suggest a mostly normal embryonic development. *PLoS Biol.* **20**, e3001741 (2022).
67. Mikkola, H. K. A. & Orkin, S. H. The journey of developing hematopoietic stem cells. *Development* **133**, 3733–3744 (2006).
68. Rybtsov, S., Ivanovs, A., Zhao, S. & Medvinsky, A. Concealed expansion of immature precursors underpins acute burst of adult HSC activity in foetal liver. *Development* **143**, 1284–1289 (2016).
69. Henninger, J. et al. Clonal fate mapping quantifies the number of haematopoietic stem cells that arise during development. *Nat. Cell Biol.* **19**, 17–27 (2017).
70. Ganuza, M. et al. Lifelong haematopoiesis is established by hundreds of precursors throughout mammalian ontogeny. *Nat. Cell Biol.* **19**, 1153–1163 (2017).
71. Kissa, K. & Herbomel, P. Blood stem cells emerge from aortic endothelium by a novel type of cell transition. *Nature* **464**, 112–115 (2010).
72. Boisset, J.-C. et al. In vivo imaging of haematopoietic cells emerging from the mouse aortic endothelium. *Nature* **464**, 116–120 (2010).
73. Luo, C. et al. Single nucleus multi-omics identifies human cortical cell regulatory genome diversity. *Cell Genom.* **2**, 100107 (2022).
74. Cao, Y. et al. Single-cell bisulfite-free 5mC and 5hmC sequencing with high sensitivity and scalability. *Proc. Natl Acad. Sci. USA* **120**, e2310367120 (2023).
75. Pei, W. et al. Resolving fates and single-cell transcriptomes of hematopoietic stem cell clones by PolyloxExpress barcoding. *Cell Stem Cell* **27**, 383–395.e8 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Serial clonal expansion of selected cell lines

Frozen HEK 293T cells (ATCC, CRL-3216) were thawed and passaged for once to recover a healthy state. Then, each HEK 293T cell was sorted by FACS in a single well of a flat-bottom 96-well plate. Cells were cultured on growth media (DMEM + 10% FBS) (ThermoFisher). After approximately 10 days in culture, cells from each well were dissociated by 5 min of treatment with 0.25% Trypsin at 37 °C (ThermoFisher, cat. no. 25200072) and sorted into two groups: one immediately sequenced and the other plated in a new flat-bottom 96-well plate (Fig. 1g).

H9 embryonic stem cells (WICELL, CVCL_9773) were cultured similarly for clonal tracking and expansion. H9 cells were dissociated by ACCUTASE (STEMCELL, cat. no. 07920) at 37 °C for 10 min and cultured on Matrigel Matrix-plated wells (Corning, cat. no. 354277) with the following medium: mTeSR1 Basal Medium, 20% mTeSR1 5× Supplement and 10% CloneR2 (STEMCELL).

Isolation of mouse hematopoietic progenitors

After euthanasia of a C57BL/6 mouse (8 weeks, female), bone marrow from the femur, tibia, pelvis and sternum was isolated by crushing with a pestle and mortar to obtain all cells. Collected bone marrow cells were filtered through a 40-µm strainer and washed in cold EasySep buffer (STEMCELL, cat. no. 20144). Red blood cells and mature lineage cells were depleted magnetically using the EasySep Mouse Hematopoietic Progenitor Cell Isolation Kit (STEMCELL, cat. no. 19856). The resulting Lin$^-$ fraction was stained for Kit (CD117-PE, clone 2B8, Biolegend, dilution 1:100), Sca1 (Ly6a-FITC, clone D7, Biolegend, dilution 1:100) and Lin$^-$Kit$^+$Sca1$^-$ (LK) cells were isolated by FACS on Sony MA900 with a 130 µM nozzle. All animal procedures were approved by the Institutional Animal Care and Use Committee of Westlake University (AP#23-093-WSW-2). All mice were fed the normal diet at the Westlake University Laboratory Animal Resources Center. The living environment of the animal laboratory was suitable, with 20–25 °C temperature, 30–70% humidity and a 12-h light–dark cycle.

In vitro lineage tracing of mouse hematopoietic progenitors

Sorted Lin$^-$Kit$^+$Sca1$^-$ cells were barcoded using spin infection (800*g* for 90 min) in LARRY lentivirus concentrate[14] with Polybrene (Sigma) and then plated in round-bottom 96-well plates. Cells were cultured in media designed to support pan-myeloid differentiation, consisting of StemSpan SFEM media (STEMCELL, 09650), IL-3 (20 ng ml^{-1}), FLT3-L (50 ng ml^{-1}), IL-11 (50 ng ml^{-1}), IL-5 (10 ng ml^{-1}), TPO (50 ng ml^{-1}) (Peprotech) and mSCF (50 ng ml^{-1}) and IL-6 (10 ng ml^{-1}) (R&D Systems). The total number of cells plated in each well varied from 1,000 to 1,500, and there were nine such wells in parallel. After 6 days in culture, cells in each well were dissociated by 5 min of treatment with 0.25% Trypsin (ThermoFisher) and then green fluorescent protein positive (GFP$^+$) cells were sorted using FACS and sequenced immediately.

Isolation of Human CD34$^+$ hematopoietic progenitors

This study complies with all relevant ethical regulations, and was approved by the Ethics Committee of Westlake University (20240222WSW0011) and conducted in accordance to the Declaration of Helsinki protocol. Written informed consent was provided by all participants. Human cord blood samples were obtained from a single donor (30 years of age, female) from Beijing Umbilical Cord Blood Bank, without compensation. Cord blood mononuclear cells were isolated by centrifugation in SepMate-50 (STEMCELL, 86450) after adding Lymphoprep (STEMCELL, 07811) to the cord blood. Red blood cells were depleted using RBC Lysis Solution (BasalMedia, S371JV) and mature lineage cells were depleted magnetically using the EasySep Human CD34 Positive Selection Kit II (STEMCELL, 17856).

In vitro lineage tracing of human hematopoietic progenitors

Sorted CD34$^+$ cells were barcoded using spin infection (800*g* for 90 min) in LARRY lentivirus concentrate with Polybrene (Sigma) and then plated in round-bottom 96-well plates. Cells were cultured in media designed to support pan-myeloid differentiation, consisting of StemSpan SFEM media (STEMCELL, 09650), and StemSpan CC100 (STEMCELL, 02690). The total number of cells plated in each well varied from 1,500 to 2,000, and there were nine such wells in parallel. After 13 days in culture, cells in each well were dissociated by 5 min of treatment with 0.25% Trypsin (Thermofisher), and then GFP$^+$ cells were sorted using FACS and sequenced immediately.

Overview of MethylTree analysis

We first generated single-cell DNA methylation profiles from a given biological sample using scBS-seq. Modified protocols can be used to also obtain the cell type information. After preprocessing of the single-cell DNA methylation data, we first selected informative genomic regions for feature extraction. Then, for each cell, we computed the average methylation fraction within each selected genomic region, resulting in a cell-by-region methylation rate matrix. Based on this rate matrix, we calculated the Pearson correlation coefficient for each pair of cells to obtain the pairwise cell–cell similarity matrix. Subsequently, we corrected the correlation bias in this similarity matrix resulting from measurement noise, and removed any effect related to cell-type differences, if this is necessary. This similarity matrix is the core object of our methylation-based lineage analysis. Finally, based on the corrected similarity matrix containing only lineage information, we performed downstream analyses such as lineage tree reconstruction and dimension reduction. The workflow of MethylTree analysis is illustrated in Figs. 1b and 2m.

scBS-seq

We used scBS-seq[49] to profile DNA methylome in single cells with the following modifications. In brief, individual cells were directly sorted into 96-well plates containing 5 µl of cell lysis buffer (1× M-Digestion Buffer (Zymo), 2 mg ml^{-1} Proteinase K (Qiagen)). Samples were incubated for 60 min at 50 °C and stored at −80 °C until required for library preparation. EZ-96 DNA Methylation-Direct MagPrep Kit (ZYMO, D5045) was then used to carry out bisulfite conversion of DNA. Library amplification and purification were performed as described previously[76]. Primers used in library amplification are listed in Supplementary Table 2.

Modified Camellia-seq

We modified the original Camellia-seq[13] protocol to jointly profile the transcriptome, DNA methylome and LARRY lineage barcode in single cells. In brief, individual cells were directly sorted into 96-well plates containing the mild lysis buffer. Dynabeads in Myone Carboxylic Acid (Invitrogen, 65011) were used to capture nuclei, and the supernatant containing released RNA was transferred to a separate 96-well plate. The Dynabeads containing genomic DNA were used to carry out scBS-seq to profile DNA methylome. The RNA part was reverse transcribed and amplified for 15 cycles. The resulting complementary DNA was split evenly, from which one half was processed with a modified single-cell tagged reverse transcription sequencing (scSTRT-seq)[57,77] to obtain the single-cell transcriptome and the other half was amplified at the target locus to generate the LARRY lineage barcodes. Primers used in library amplification are listed in Supplementary Table 2.

Preprocessing of scBS-seq data

Processing of DNA methylation data generated from scBS-seq was the same as in our previous work[13], except for some minor changes described below. In brief, we used Bismark (v.0.24.0) for read alignment, deduplication and extraction of DNA methylation information. We expected cells with high quality to have low CpG methylation level around the transcription start site (TSS). To quantify this, we calculated the Pearson correlation C_{TSS} between m and $|x|$ for each cell, where m is the average methylation level across all CpG sites at a distance x to TSS. Cells with $C_{\mathrm{TSS}} > 0.7$ and having $\geq 5 \times 10^5$ distinct CpG sites were considered to pass quality control and used in downstream analysis.

Selection of informative genomic regions

To generate the cell-by-feature methylation rate matrix, we need to decide which features or genomic regions to use (Fig. 1b). A simple approach is to just use all the individual CpG sites, thus generating a cell-by-CpG matrix for downstream analysis. This approach is highly accurate (Fig. 4e and Extended Data Figs. 1c, 5b and 8c), but computationally also very expensive since there are more than 20 million CpG sites.

Here, we discuss an alternative approach that divides the genome into nonoverlapping bins or regions. In general, we found that using more genomic regions often leads to better results (Extended Data Fig. 1g). However, for a given system (for example, cells of a given type or from a tissue), certain choices of genomic regions generate more accurate cell lineages than others (Fig. 1n and Extended Data Fig. 1f), because some regions are enriched with epimutations in a cell-type-specific way. Below, we describe our approach for selecting an optimal set of genomic regions for lineage reconstruction, which can be reused to process new data from the same system, as we did in analyzing the multiple early embryo datasets as well as the HSC datasets (Figs. 5 and 6).

First, we aggregated sparse DNA methylation profiles across all cells to obtain a high-coverage pseudobulk profile for this particular system. We divided the genome into nonoverlapping 500-bp bins, computed the average DNA methylation rate m at each bin, and selected the bins whose methylation rate satisfies $m_0 \leq m \leq m_1$ (Extended Data Fig. 1d), where m_0 and m_1 are two tunable parameters. This set of selected regions, each with 500 bp, is referred to as 'not merged'. We then merged all the neighboring regions (Extended Data Fig. 1e). This 'merged' region set has a varying length distribution. In both the merged and not merged region sets, we excluded regions observed in <10% of cells.

To select the optimal set of regions for a particular system, we systematically varied m_0 and m_1, reported the corresponding lineage reconstruction accuracy from downstream analysis, and selected the best choice. We observed that the merged set of genomic regions tended to show better performance (Fig. 1n). Furthermore, the merged set required less computation because of the reduced region number, which implies a smaller cell-by-region rate matrix. Therefore, we have used the merged set of regions throughout this article, unless otherwise stated. The selection parameters for each dataset analyzed in this study are provided in Supplementary Table 1.

Generation of the similarity matrix

Given the selected genomic regions, we generated a cell-by-region methylation rate matrix A_{ix} by computing the average methylation fraction of cell i within genomic region x. To compute the similarity S_{ij} between cells i and j, we identified a subset of regions Ω_{ij}, where both cell i and j had detected values (ranging between 0 and 1, but not NaN), and computed the Pearson correlation between these two cells only in this shared subset Ω_{ij}:

$$S_{ij} = \mathrm{Corr}_{x\in\Omega_{ij}}\left(A_{ix}, A_{jx}\right).$$

Our approach does not require any imputation on the observed sparse DNA methylation data, which is important for preserving the cell specific epimutations. We found that this approach accurately extracts lineage relationships from sparse DNA methylation data, and is robust to the heterogeneity of CpG coverage between cells (Extended Data Fig. 1h). Besides, we found that Pearson correlation performs better than Euclidean or cosine similarity in our experience (Extended Data Fig. 1i).

Similarity matrix correction

As mentioned in the main text, the Pearson correlation between two observed variables will be lower than the actual value because measurement noises make these two variables less connected. To see this, consider $\mathbf{x}_o = \mathbf{x} + \boldsymbol{\eta}_x$ and $\mathbf{y}_o = \mathbf{y} + \boldsymbol{\eta}_y$, where $\mathbf{x}_o$ and $\mathbf{y}_o$ are the observed signals, $\mathbf{x}$ and $\mathbf{y}$ are the original signals and $\boldsymbol{\eta}_x$ and $\boldsymbol{\eta}_y$ are the corresponding measurement noises. The observed Pearson correlation $C(\mathbf{x}_o, \mathbf{y}_o)$ satisfies the following relationship:

$$C(\mathbf{x}_o, \mathbf{y}_o) = \frac{C(\mathbf{x}, \mathbf{y})}{Z_x Z_y}, Z_i = \frac{\sigma_{i_o}}{\sigma_i}, i \in \{x, y\}.$$

Here, $C(\mathbf{x}, \mathbf{y})$ is the actual Pearson correlation between $\mathbf{x}$ and $\mathbf{y}$. σ_i and σ_{i_o} are the standard deviations of the original and observed signal, respectively. Therefore, Z_i is the relative measurement noise of the observed signal i. A larger relative measurement noise leads to a smaller observed Pearson correlation $C(\mathbf{x}_o, \mathbf{y}_o)$. This is known in the literature of statistics as attenuation[54].

Since our similarity matrix S is based on Pearson correlation, this attenuation phenomenon leads to a biased estimation of the true similarity between two cells. We exploit only relative similarity among these cells to construct lineages. Therefore, if noise factor Z_i has the same value across all cells, the relative similarity would remain the same and the reconstructed lineage tree is correct. However, if certain cells have a very different noise factor Z_i, which could arise from heterogeneity in library preparation and sequencing, this will skew the relative similarity and may lead to erroneous estimation of cell lineages. Below, we developed an iterative algorithm that only takes the raw correlation matrix as input to correct the relative attenuation bias. This method should work in other contexts that involve a correlation matrix.

The key is to find the actual noise factor Z_i. We first initialized the noise factor Z_i for cell i from the similarity matrix S, by setting it as the inverse square root of the maximum off-diagonal value of the ith row in this matrix:

$$Z_i = \frac{1}{\sqrt{\max_{j\neq i} S_{ij}}}.$$

This is then normalized by the average value of Z to mitigate confounding factors such as sequencing depth:

$$Z_i \leftarrow \frac{Z_i}{\mathrm{mean}(Z)}.$$

Then, we generated a corrected similarity matrix S^* as

$$S^*_{ij} = Z_i S_{ij} Z_j.$$

We evaluated this corrected matrix S^* through a cost function,

$$f_c\left(S^*\right) = \frac{\sigma_{\mathrm{off-diagonal}}\left(S^*\right)}{\mu_{\mathrm{off-diagonal}}\left(S^*\right)}$$

which computed the ratio between the standard deviation σ and the mean μ of the off-diagonal values of S^*. We then use gradient descent to search for the optimal noise factor Z that minimize the cost function of the corrected S^* matrix. We iterate the above steps using this new S^* as the input until convergence. Specifically,

```
Function Similarity_Correction ($S$)
Do # iterate for the convergence of the corrected $S$
  Initialization: $Z_i = \frac{1}{\sqrt{\max_{j\neq i} S_{ij}}}, i = 1, \ldots, n$
  Normalization: $Z_i \leftarrow Z_i/\mathrm{mean}(Z)$
  While $f_c(S^*) - f_c(S) < 0$:
      # Use gradient descent to find the optimal $Z$ for a given $S$
      Update $Z$: $Z_i = Z_i - \epsilon \frac{\partial f_c}{\partial S^*} \frac{\partial S^*}{\partial Z_i}, i = 1, \ldots, n$
      Normalization: $Z_i \leftarrow Z_i/\mathrm{mean}(Z)$
      Correct $S$ matrix: $S^*_{ij} = Z_i S_{ij} Z_j$.
   **If** $||S - S^*||_1 < \delta$: **Break** # check L1 norm for convergence
   Update $S$: $S_{ij} \leftarrow S^*_{ij}$
**Return** $S$
```

Here, ϵ controls the step size of gradient descent and δ controls convergence. Both should be small in implementation. We used $\epsilon = 0.01$ and $\delta = 0.01$. In most datasets of our current study, we found that the correlation-bias correction improves lineage reconstruction (Fig. 1i,j and Extended Data Figs. 1i and 8b).

Removal of cell-type signals

Cell-type-specific DNA methylation signal may dominate the methylation similarity matrix S. Instead of trying to identify 'neutral' genomic regions that on average do not change their DNA methylation status during differentiation, we sought to computationally remove these uninteresting signals on the raw similarity matrix computed with genomic regions identified above. This approach requires cell-type information of each cell.

We hypothesized that the raw similarity matrix S can be decomposed to be a cell-type-specific similarity matrix T and a lineage-specific similarity matrix L:

$$S = T + L.$$

Then, we computed the cell-type-specific similarity T_{ij} between cells i and j. We denoted t_i and t_j as the corresponding cell type of cells i and j, respectively. We first identified the set of cells (other than j) that share the same cell type as i:

$$\Omega_i^0 = \{k | t_k = t_i, k \neq j\}.$$

In principle, the similarity set $\{S_{kj} | k \in \Omega_i^0\}$ could be used directly to compute T_{ij} through averaging when there are many lineages and cells in a dataset. However, when there are only a few lineages or cells in a dataset, we found it helpful to further exclude cell pairs (k, j) that come from the same clone, as such cell pairs would have higher similarity and therefore could inflate the estimate of T_{ij}. Since clonal information is considered unknown, we used the following approach to identify such putative clonal pairs and improve the estimation of T_{ij}. First, compute the mean similarity μ and the standard deviation σ of the similarity set $\{S_{kj} | k \in \Omega_i^0\}$. Then, generate a more restricted set of cells Ω_i that share the same cell type as i:

$$\Omega_i = \{k | t_k = t_i, k \neq j, S_{kj} \leq \mu + \sigma\}.$$

Similarly generate Ω_j as the restricted set of cells that share the same cell type as j. Finally, we computed T_{ij} as the average similarity among these cell pairs involving cell type t_i and t_j in the following way:

$$T_{ij} = \frac{\sum_{k \in \Omega_i} S_{kj} + \sum_{k \in \Omega_j} S_{ik}}{|\Omega_i| + |\Omega_j|}.$$

Here, $|\Omega_i|$ and $|\Omega_j|$ give the total number of cells in set Ω_i and Ω_j, respectively. After T was computed, the lineage-specific similarity matrix L was simply given by $L = S - T$. Applying this approach, we successfully removed cell-type-specific differences and revealed actual lineage information from both simulated and public single-cell DNA methylation data (Fig. 2e,l and Extended Data Fig. 3a,c,g,h).

Rescaling of the similarity matrix

Before downstream analysis, we rescaled the similarity matrix S from the above computation, so that its minimum value is 0 and the maximum is 1 in the off-diagonal entries:

$$S_{ij} = \frac{S_{ij} - \min\limits_{\mathrm{i \neq j}}(S_{ij})}{\max\limits_{\mathrm{i \neq j}}(S_{ij}) - \min\limits_{\mathrm{i \neq j}}(S_{ij})}.$$

We fixed the diagonal entries to be 1 afterward: $S_{ii} = 1$.

Lineage tree reconstruction

To build a lineage tree, we need to convert the similarity matrix S into a distance matrix D to use existing phylogenetic tree reconstruction algorithms such as UPGMA. For this, we generated the distance matrix with $D_{ij} = 1 - S_{ij}$. Then, we set the diagonal term $D_{ii} = 0$. We applied UPGMA to this distance matrix to infer the lineage tree. We used the inferred lineage relationships between cells to order the heatmap of the corresponding similarity matrix. Therefore, the heatmap illustrated both the similarity matrix and the inferred lineage relationships.

To estimate the reliability of the inferred lineage tree (referred to as Γ_o), we randomly sampled 80% of the selected regions in the corresponding cell-by-region methylation rate matrix A_{ix} and rerun the remaining MethylTree steps to obtain a new tree (denoted as Γ_s). We repeated this process 100 times to obtain a long list of sampled trees $\{\Gamma_s\}$. For each subtree Γ_o^k from the original tree Γ_o, we iterated through $\{\Gamma_s\}$ and checked whether a 'similar' subtree Γ_s^k can be found in Γ_s. By 'similar', we specifically meant that both subtrees were composed of the same group of cells, regardless of their structural organization within their corresponding subtrees. We reported the fraction of such 'similar' occurrences across all the 100 sampled trees $\{\Gamma_s\}$ as the support value of the observed subtree Γ_o^k, which was displayed near the root of each subtree of Γ_o (Fig. 5i,k and Extended Data Fig. 6e–g). These support values range from 0 and 1. A higher support value implies that cells belonging to the corresponding subtree are more likely clustered together in the actual lineage tree of all the cells.

Methyl-clone identification

We used the support values computed from the previous step to identify putative clones, which we named methyl clones. To add additional information for clone identification, for each subtree we also computed a within-tree similarity score, which is defined to be the 50th percentile of the off-diagonal similarities between cells within this subtree. We expected that intra-clone similarity would be higher than that between randomly selected cells. Starting from the root of the tree, we identified putative clones that satisfied two criteria: (1) a subtree with support values above a preset support threshold; and (2) a subtree whose within-tree similarity score is larger than a preset similarity threshold. In this study, we set the similarity threshold to be the 75th percentile of the off-diagonal similarities among all cells, and the support value threshold is 0.95. We evaluated the accuracy of these methyl clones by computing the adjusted rank index using the ground-truth lineage barcode as the reference (Figs. 3i and 6b and Extended Data Fig. 7b).

Dimension reduction

We performed dimension reduction on the similarity matrix S using the function sklearn.manifold.spectral_embedding with a parameter n_components. The resulting spectral components were used to generate a k-nearest neighbors graph with scanpy.pp.neighbors function that have a tuning parameter n_neighbor. Finally, we performed uniform manifold approximation and projection for dimension reduction (UMAP) using the scanpy interface scanpy.tl.umap with a parameter min_dist to generate a two-dimensional embedding. Denoting the involved parameters as this triplet: (n_components, n_neighbor, min_dist), we used the parameter set (10,7,0.4) for Extended Data Fig. 1h, (10,10,0.5) for Extended Data Fig. 3f, (10,10,0.9) for Fig. 3k and (10,40,0.5) for Fig. 6c,d.

Accuracy of lineage reconstruction

The linear ordering of leaf nodes in the reconstructed lineage tree can be used to evaluate the accuracy of lineage reconstruction, when the ground-truth lineage or clonal label is available. We expect that cells from the same clone are more similar and appear in the same subtree. Therefore, these clonal cells should be grouped together in the linear ordering of the leaf nodes. Motivated by this expectation, we devised the following metric Q to quantify how close cells from the same clone are arranged in this lineage ordering:

$$Q = \left(\sum_k \frac{g_k - 1}{n_k - 1} \right) / M,$$

where g_k is the maximum number of cells from clone k that are grouped together in the inferred lineage ordering, n_k is the total number of cells from clone k and M is the total number of clones. Only clones with at least two cells were used in this evaluation, and we excluded cells without observed clonal labels. The '−1' in this equation ensures that accuracy score reaches a baseline of zero when none of the clones have more than one cell placed together. Therefore, Q ranges between 0 and 1, with 1 corresponding to the best scenario where all cells are grouped by their clonal origins.

To gain a better understanding of this metric, consider the accuracy of a randomized ordering. Specifically, suppose that we have M clones, each with two cells. Averaging over 100 independent simulations, we have $Q = 0.18$ for $M = 5$, $Q = 0.096$ for $M = 10$, $Q = 0.05$ for $M = 20$. Therefore, an accuracy of $Q > 0.8$ should be considered a very high accuracy. We reported the accuracy of randomized lineage ordering associated with each analyzed dataset in Supplementary Table 1 and Extended Data Fig. 8b.

Coarse-grained methylation lineages

Apart from single-cell lineage tree inference, we also inferred a phylogenetic relationship at a higher level by aggregating individual cells from similar cell types or clones (Fig. 1l). Using l to denote a group label, be it a cell type or clone annotation, we extracted the subsimilarity matrix S_{ij} with cells from group l and l', computed the median value of the off-diagonal elements and assigned it to the coarse-grained similarity $\bar{S}_{ll'}$. Then, we applied UPGMA to the coarse-grained similarity $\bar{S}_{ll'}$ to generate a lineage tree, as described above.

Analysis of single-cell transcriptomic data

Preprocessing of single-cell transcriptomic data from our modified Camellia-seq protocol is identical to the original Camellia-seq analysis[13]. For dimension reduction, we selected highly variable genes, removed cell-cycle effects, used the top 40 principal components to obtain a k-nearest neighbors graph with n_neighbors = 15 and run UMAP at min_dist = 0.3 to generate the two-dimensional embedding.

Analysis of LARRY barcode data

We modified the original LARRY bioinformatic pipeline developed by Weinreb et al.[14]. Since the sequencing data was generated through a plate-based protocol, we preprocessed data from each 96-well plate separately. After initial extraction of the cell barcode, UMI barcode and LARRY lineage barcode for each read from the fastq files, we first excluded reads that do not have the expected cell barcode or do not conform to the expected LARRY barcode structure. To correct the PCR and sequencing errors in the LARRY lineage barcode, LARRY barcodes supported by fewer than eight reads were discarded. We grouped the remaining LARRY barcodes within a Hamming distance of ten (note that a LARRY barcode has 28 variable nucleotides), and corrected them toward the most dominant LARRY barcode in this group. To further avoid the scenario where artificial cell barcodes associated with a LARRY clone were generated through PCR and sequencing errors, we excluded cell barcodes with a relative read fraction <1% (relative to the most abundant cell barcode in this clone) among all the cell barcodes sharing the same LARRY clone ID. After preprocessing, we used the latest version of CoSpar[59] (v.0.3.3) to generate clonal heatmap (Fig. 3d), visualize individual clones on the transcriptomic embedding (Fig. 3e) and compute the clonal coupling heatmap as well as the differentiation hierarchy (Extended Data Fig. 4g).

Clone identification from DNA methylation

We expect that HSCs arising from the same EHT-derived clone would share much higher methylation similarity than two random cells, and therefore would form a subtree with strong support values. We identified putative methyl clones as described above. Next, we inferred the actual number of HSC clones from the observed putative clones at the selected support threshold. Due to observing only 100–200 HSCs, it is highly likely that some HSC clones were not observed in our data. We computed the singleton cell fraction ϕ for each dataset, defined as the fraction of observed cells with no sister cells jointly detected from the same clone (Fig. 6e). To infer the actual HSC clone number, we generated M synthetic HSC clones that have the empirically observed clone size distribution, sampled N cells with replacement from this synthetic HSC pool and calculated the singleton cell fraction $\phi(M, N)$. Since LL731 is the largest dataset and therefore has the most reliable clone size distribution across our datasets, we used its distribution as the empirical clone size distribution in our simulation. For each dataset k from {LL731, LL653E1, LL653E6}, we set sampled cell number N to be the observed cell number N_k in this dataset, and inferred the actual clone number as M_k such that $\phi(M_k, N_k) = \phi_k$, where ϕ_k is the observed singleton cell fraction in this dataset (Fig. 6f,g and Extended Data Fig. 7d,e). We repeated the simulation 100 times to estimate the standard deviation of the simulated singleton cell fraction ϕ. We obtained highly consistent estimates of ~250 HSC clones across three datasets from two stages (that is, fetal liver and adult stage) (Fig. 6g).

HSC clone number estimation from bulk DARLIN data

Bulk DARLIN data were preprocessed according to our previous study[13]. We counted the total number (denoted U) of unique DARLIN barcodes observed in a mouse, excluding the unedited barcode. DARLIN sequences in some of the cells were not edited during induction, which could have led to under-estimation of the total HSC clones. To correct for this, we computed the fraction (denoted β) of DARLIN UMIs that were edited and estimated the expected DARLIN clone number to be U/β (Fig. 6h). As mentioned in the main text, background editing due to leaky expression of Cas9-TdT could inflate this HSC clone estimate from DARLIN mice.

Simulation of single-cell expansion

To evaluate the capacity of DNA methylation epimutations to reconstruct human cell lineage histories at a given genomic coverage, we simulated DNA methylation epimutations on a large array of 29 million CpG sites (Fig. 1d–f). For each CpG site, we used 0 to represent the unmethylated state and 1 to represent the methylated state. An epimutation occurs when $0 \rightarrow 1$ or $1 \rightarrow 0$, at a frequency of 0.001 per site per cell division, unless otherwise stated. Starting from a single cell, where 50% of the CpG sites were randomly methylated initially, we simulated erroneous replication of DNA methylation for seven cell divisions, resulting in 128 cells. We profiled these cells at 5% genomic coverage, unless otherwise stated. To simplify the estimation of lineage reconstruction accuracy, we introduced synthetic clone barcodes at the stage of 16 cells produced after the first four divisions, so that each clone has exactly eight cells. We reported the MethylTree accuracy for these 16 clones using the metric Q described above, after averaging over ten independent simulations (Fig. 1f).

To add more realistic complications to the simulation, we first considered the existence of epimutations without cell divisions. To simulate this, we randomly mutated a given fraction of CpG sites in each of the 128 cells after the clonal expansion and reported the MethylTree accuracy (Extended Data Fig. 1a). Second, we adapted the above simulation and modeled epimutation on a diploid genome with two independent copy of DNA molecules, each with a CpG site-specific epimutation rate sampled from a uniform distribution with a maximum value λ. In addition, each observed CpG status is obtained from sampling once on the same CpG site from either of the two DNA molecules. Afterward, we randomly select only a fraction of these CpG sites as the final observations (Extended Data Fig. 1b).

Simulation of stem-cell expansion and differentiation

To test MethylTree in inferring lineages from a mixture of different cell types, we simulated DNA methylation changes during stem-cell expansion and differentiation (Fig. 2a–e). There are three cell types: stem cell, diff_A, and diff_B. The latter two are differentiated cell types. At each generation, each cell divides once regardless of their cell identities. Each of the two daughter cells, if in the stem-cell state, has 20% chance to become either diff_A or diff_B after division. Each cell type has its specific DNA methylation pattern occupying the first half of the genome, with the remaining 50% of the genome being neutral or lineage-specific. We denote this cell-type-specific region as $\mathfrak{R}$. For simplicity, we set the methylation status on $\mathfrak{R}$ as [0,1,0,1,0,1,...] for a stem cell (which alternates between and 0 and 1), uniformly 0 for diff_A and uniformly 1 for diff_B. When there is a transition from one cell type to another (for example, when a stem cell differentiates to become diff_A), the methylation pattern on region $\mathfrak{R}$ is reset to adopt the pattern in the new cell type. At each division, cells accumulate epimutations across the genome, including $\mathfrak{R}$, at a rate of 0.001 per CpG site per cell division.

We simulated ten founder cells, each initialized as a stem cell. These founder cells adopt a similar DNA methylation pattern on the second half of the genome that is neutral to cell differentiation, but with 5% differences. There are 10^5 CpG sites in this simulated genome. Starting from these ten founder cells, we simulated five generations of cell division and differentiation, resulting in ten clones, each comprised of 32 cells with three cell types.

In Extended Data Fig. 3, to add more realistic complications, we varied the proportion α of the genome that has lineage-specific CpG sites, with the remaining $(1-\alpha)$ being cell-type specific (Extended Data Fig. 3b,c). We also modeled variation of the epimutation rate across the genome by sampling from a uniform distribution with the maximum rate being λ, and considered the effect of sampling different proportion of the CpG sites (Extended Data Fig. 3a).

Genomic localization of clone-specific CpG sites

The clone-specific CpG sites are identified from two largest clones in our in vitro mouse blood dataset (Extended Data Fig. 4i). Each qualified clone-specific CpG site should be observed in more than five cells in each clone, and its mean methylation rate within a clone should be <0.05 or >0.95 in one clone and within [0.4,0.6] in another clone. To compute the one-sided P value for the N observed clone-specific CpG sites, we generated the null localization data by randomly sampling N CpG sites among all observed CpG sites and repeated it for 100 times.

Reporting summary

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

Data availability

The sequencing data for human blood has been submitted to the Genome Sequence Archive database under the accession number HRA008624. Other sequencing data generated in this study have been submitted to NCBI GEO, with the accession number GSE262580. The methylation rate matrix associated with selected genomic regions for each analyzed dataset in our paper, along with each sample metadata and processed human blood dataset,is available via figshare at https://doi.org/10.6084/m9.figshare.27288630 (ref. 78). The accession number and analysis parameters for each analyzed dataset in this study are available in Supplementary Table 1.

Code availability

Scripts for data preprocessing are available at https://github.com/ShouWenWang-Lab/Preprocessing. MethylTree code is available at https://github.com/ShouWenWang-Lab/MethylTree. To reproduce our analysis, please check out our jupyter notebooks at https://github.com/ShouWenWang-Lab/MethylTree_notebooks. A web portal of MethylTree analysis is available at https://wangshouwen.lab.westlake.edu.cn/app/methylserver.

References

76. Clark, S. J. et al. Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat. Protoc.* **12**, 534–547 (2017).
77. Islam, S. et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
78. Chen, M., Fu, R., Chen, Y., Li L. & Wang, S.-W. MethylTree: high-resolution, noninvasive single-cell lineage tracing in mice and humans based on DNA methylation epimutations. *figshare* https://doi.org/10.6084/m9.figshare.27288630 (2024).

Acknowledgements

We are grateful to D. Li, H. Yu and K. Zhang from Westlake University for their stimulating discussions. We thank W. Liu and D. Ruan from Zhejiang University for sharing LARRY lentivirus. We acknowledge H. Guo from Zhejiang University as well as W. Yue and H. Yan from Beijing Institute of Radiation Medicine for their help. We also thank other laboratory members for their input. L.L. is supported by National Key Research and Development Project of China (2024YFA1306600). We acknowledge support from NSFC (grant no. 32470700), Westlake High-Performance Computing Center, and 'Pioneer' and 'Leading Goose' R&D Programs of Zhejiang province (grant nos. 2024SSYS0034 and 2024SSYS0034).

Author contributions

S.-W.W. and L.L. conceived the project. S.-W.W. acquired funding, developed MethylTree, designed experiments, carried out biological applications and wrote the manuscript with input from other authors. M.C. performed all experiments. R.F. carried out data analyses and generated figures. Y.C. developed the web portal. L.L. and S.-W.W. supervised M.C. to set up the experimental system and generate experimental datasets. S.-W.W. supervised the entire project.

Competing interests

S.-W.W. is named inventor on a patent application for MethylTree (PCT/CN2024/095497). The other authors declare no competing interests.

Additional information

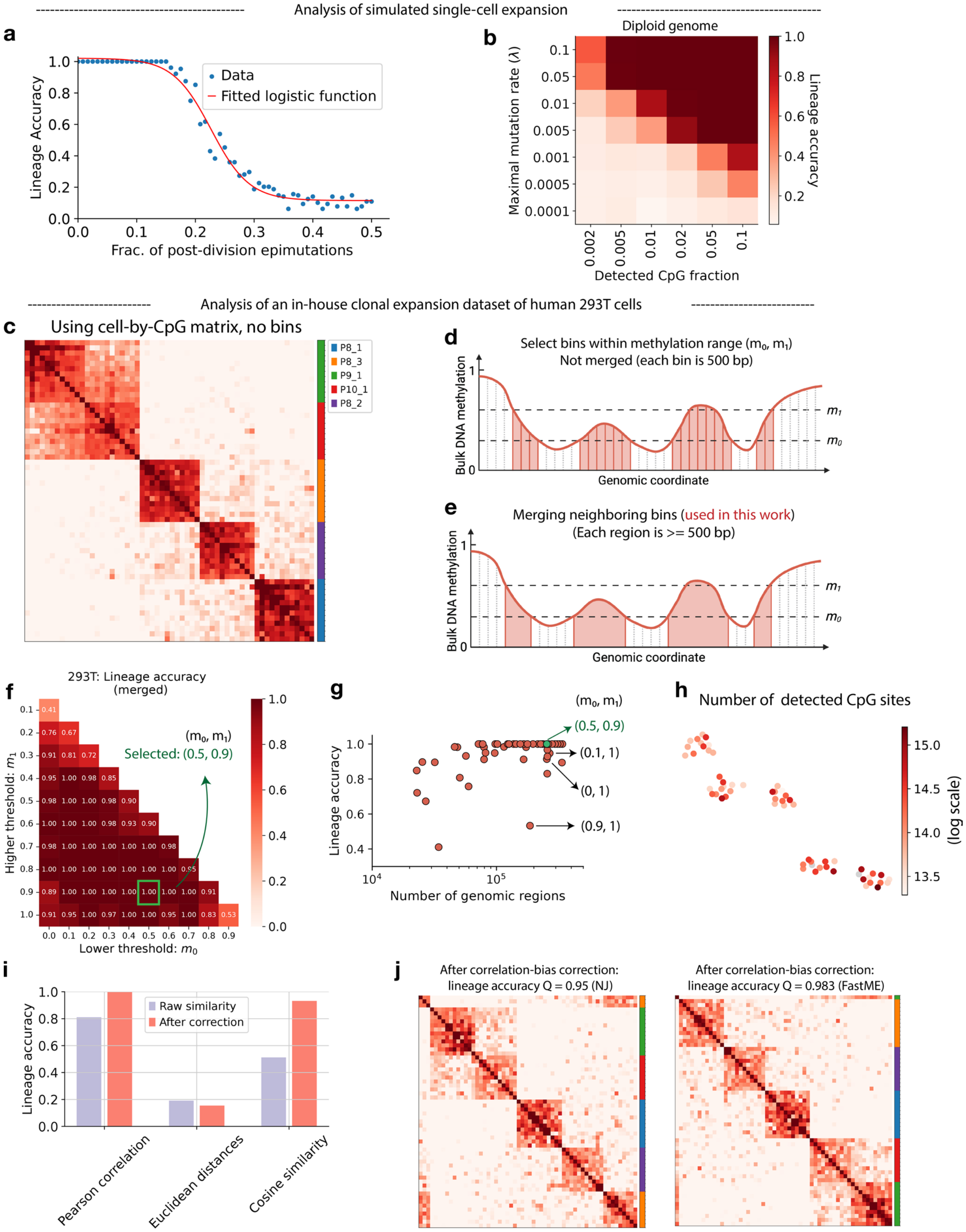
Extended data is available for this paper at https://doi.org/10.1038/s41592-024-02567-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02567-1.

Correspondence and requests for materials should be addressed to Li Li or Shou-Wen Wang.

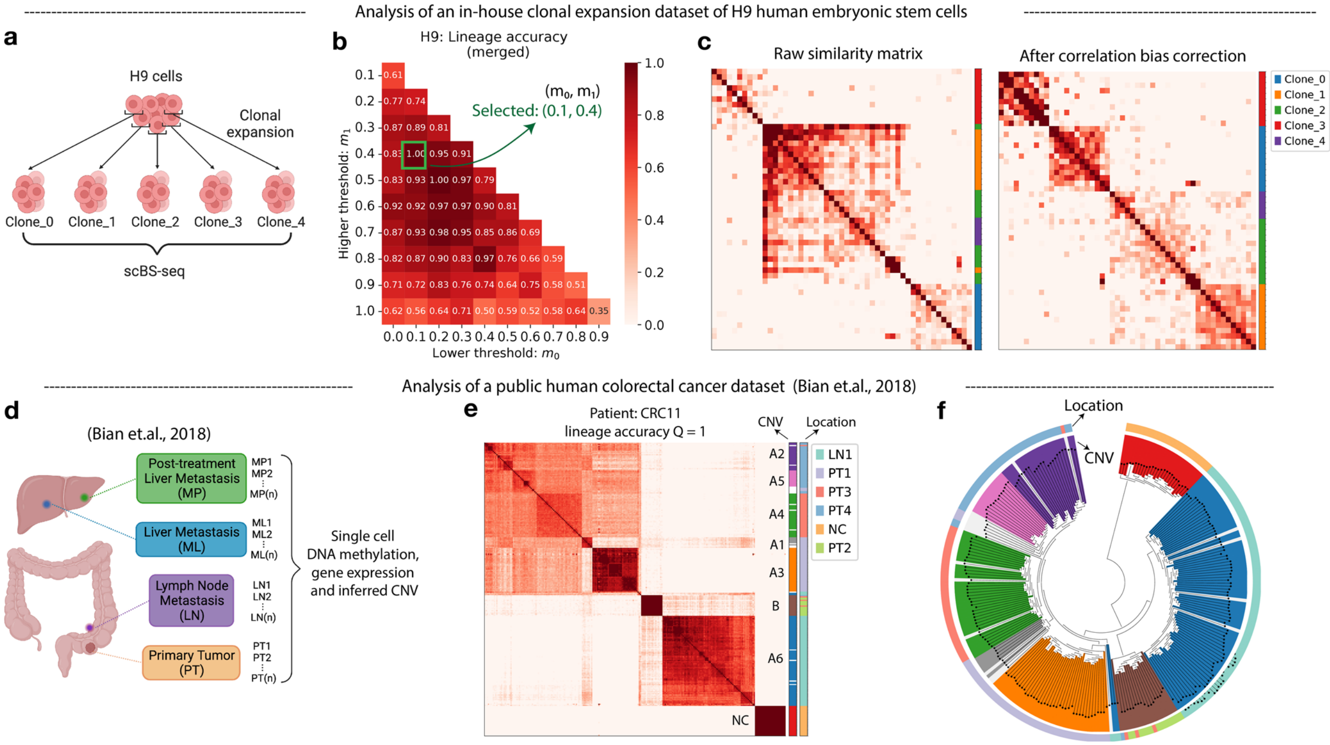
Peer review information *Nature Methods* thanks Simon Anders and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Lei Tang, in collaboration with the *Nature Methods* team.

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

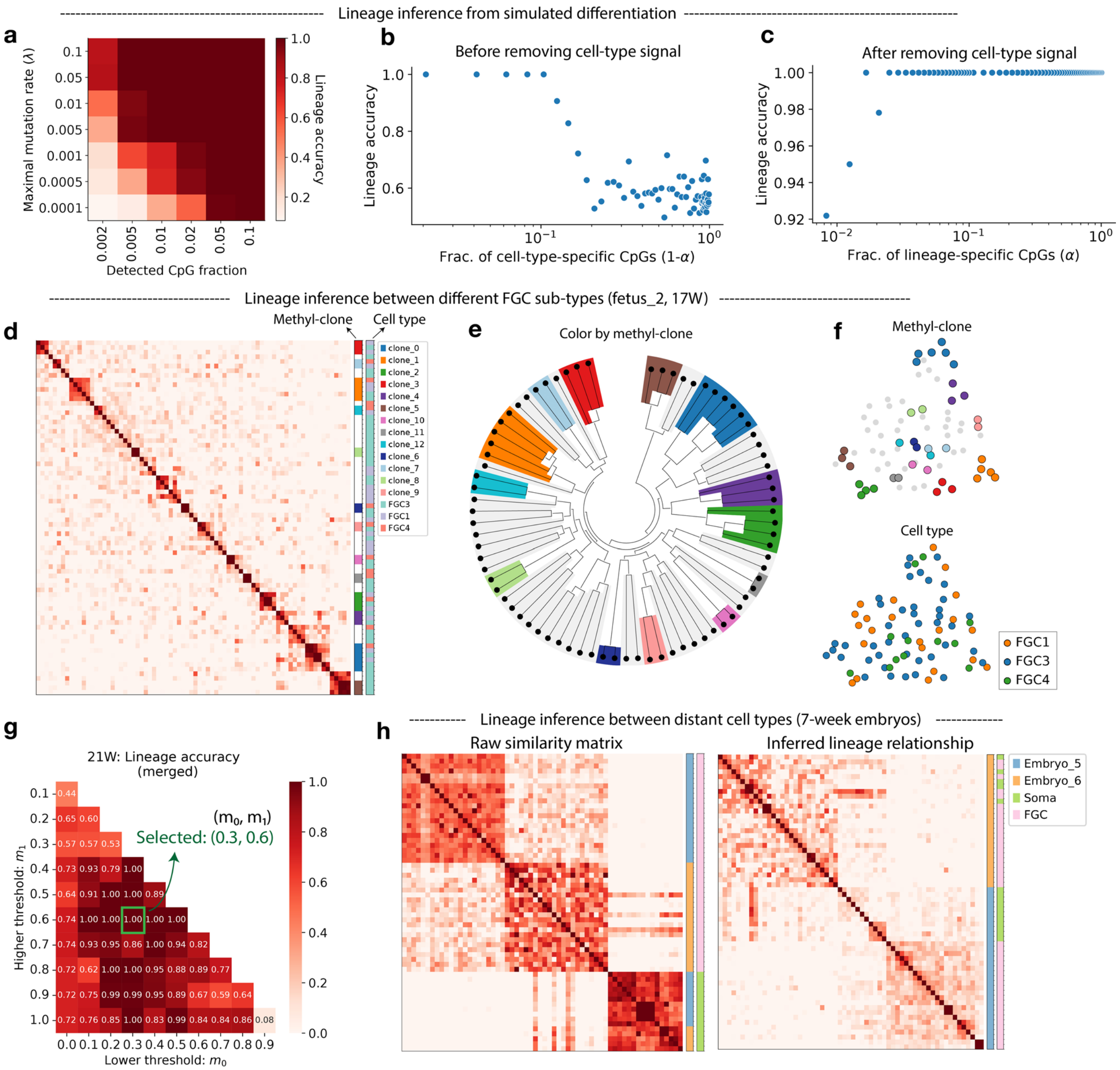


Extended Data Fig. 1 | See next page for caption.

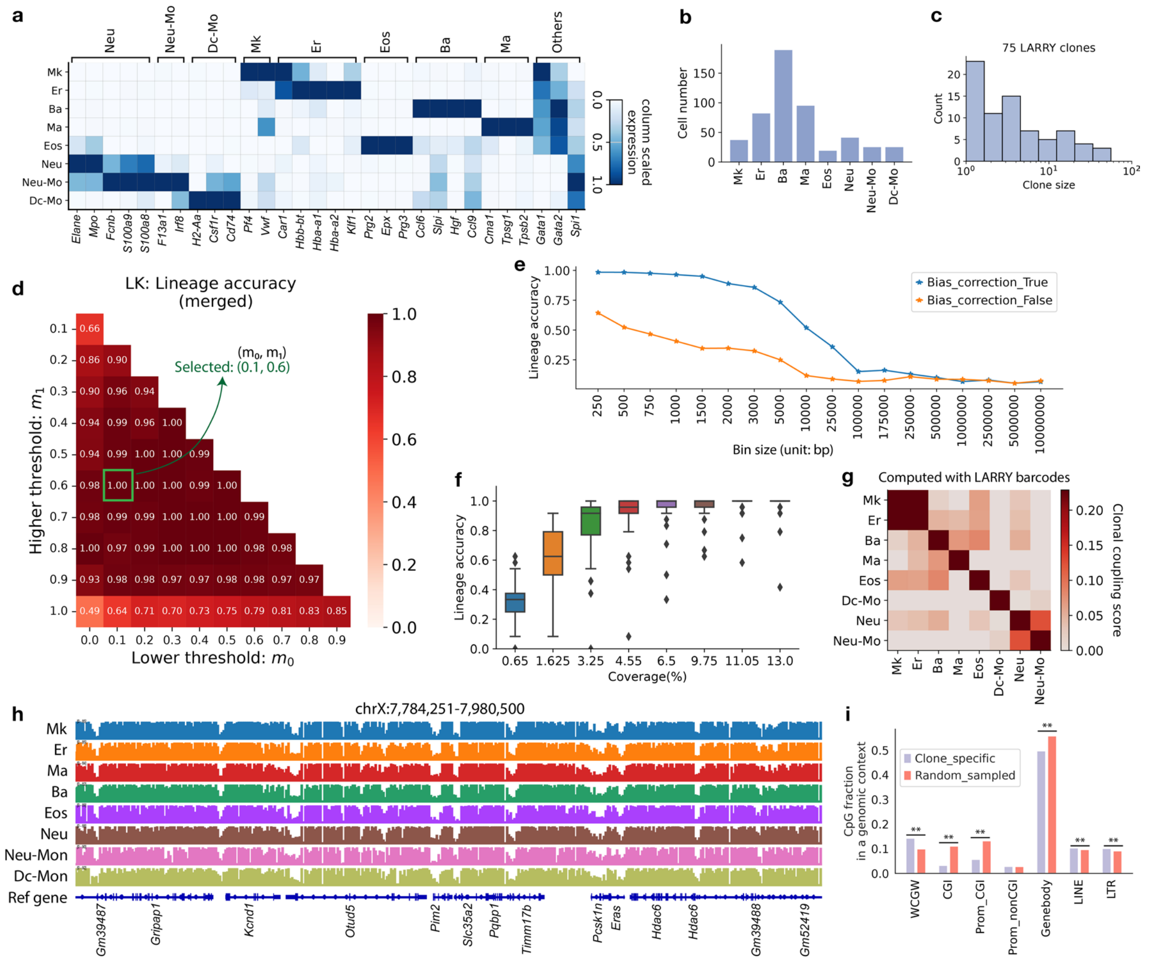
Extended Data Fig. 1 | Systematic characterization of MethylTree performance in a homogeneous population. a, **b**, Analysis on simulated single-cell expansion with more realistic features. **a**, The impact of division-free CpG mutations on lineage inference accuracy. After simulated clonal expansion as in Fig. 1d, we randomly mutated a given fraction of CpG sites in each of the 128 cells. **b**, Heatmap of lineage accuracy as a function of CpG coverage and the variation of epimutation rate controlled by the parameter λ. Compared with Fig. 1f, we modeled epimutation on a diploid genome with a CpG-site specific epimutation rate sampled from a uniform distribution with a maximum value λ. Each observed CpG status is obtained from sampling once on the same CpG site from either of the two DNA molecules. **c**–**j**, MethylTree analysis of a clonal expansion dataset of human HEK 293T cells. **c**, Heatmap of the similarity matrix computed with the cell-by-CpG matrix, without binning. **d**, Schematic of region selection. Non-overlapping 500-bp genomic bins with an intermediate methylation rate between m_0 and m_1 were selected. **e**, Merging neighboring bins after selection in **d**. This procedure was used in analyzing all datasets in this article. **f**, Heatmap of MethylTree lineage accuracies on the 293T dataset using 'merged' genomic regions selected at different thresholds according to **e**. The parameters indicated on this plot ($m_0 = 0.5$, $m_1 = 0.9$) were used to generate Fig. 1i–k. **g**, A scatter plot showing the number of genomic regions associated with each selection and the corresponding accuracy of MethylTree-inferred lineages, using the data from **f**. The selection parameters (m_0, m_1) for some data points are highlighted. **h**, Number of detected CpG sites per cell on the methylation embedding of 293T cells. **i**, Lineage accuracy using different metrics to compute the cell-cell similarity. With Euclidean distance matrix X, we converted it to a similarity with $1 - X/\max(X)$, where $\max(X)$ is the largest value in this matrix. **j**, Similarity heatmap ordered with the phylogenetic tree inferred from the neighbor-joining[56] (NJ, left) or FastME[57] (right) method.



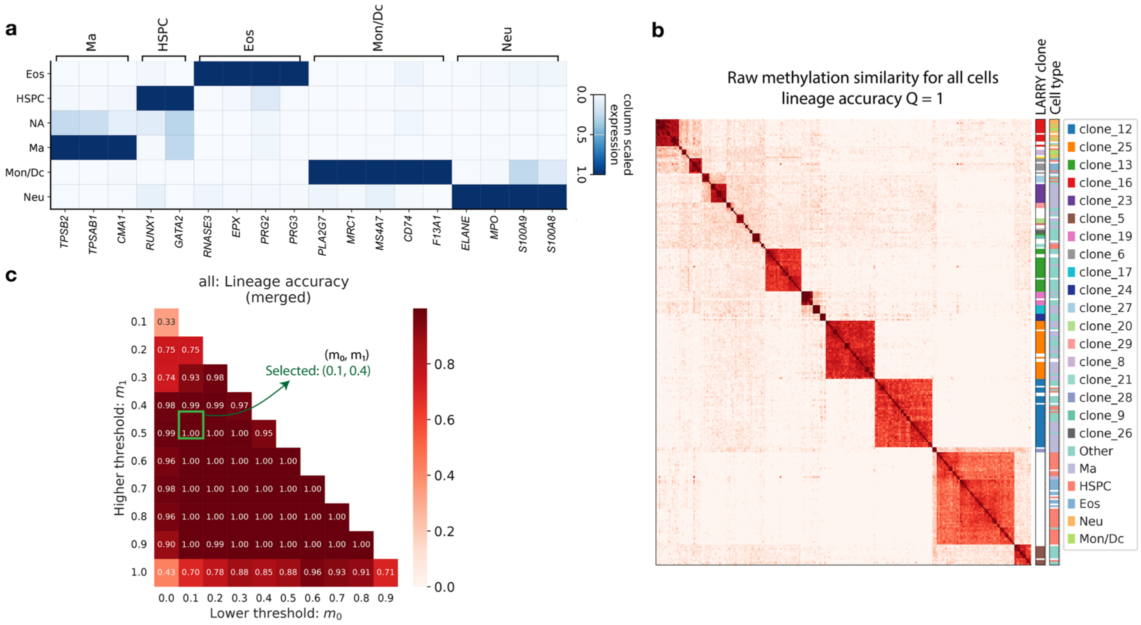
Extended Data Fig. 2 | Lineage inference from human embryonic stem cells and colorectal cancer. **a**–**c**, Lineage analysis of clonal expansion of H9 human embryonic stem cells. **a**, Schematic of our experimental design, created using BioRender.com. There were five clones generated in this experiment. **b**, Heatmap of MethylTree lineage accuracies on the H9 dataset, similar with Extended Data Fig. 1f. **c**, Heatmap of the similarity matrix of the H9 dataset before (left) and after (right) correlation-bias correction. The color bar shows the actual clonal identify of each cell. **d**–**f**, Lineage inference from human colorectal cancer. Data is obtained from patient CRC11 in Bian et al.[38]. **d**, Schematic of tissue sampling and cell profiling, created using BioRender.com. **e**, Heatmap of the cell-cell similarity matrix computed from single-cell DNA methylation. Here, A1–A6 and B were inferred cancer lineages based on copy number variations (CNV) in the original analysis by Bian et al. NC marks the normal cells. **f**, Lineage phylogenetic tree inferred from the methylation matrix. Same color as **e**.



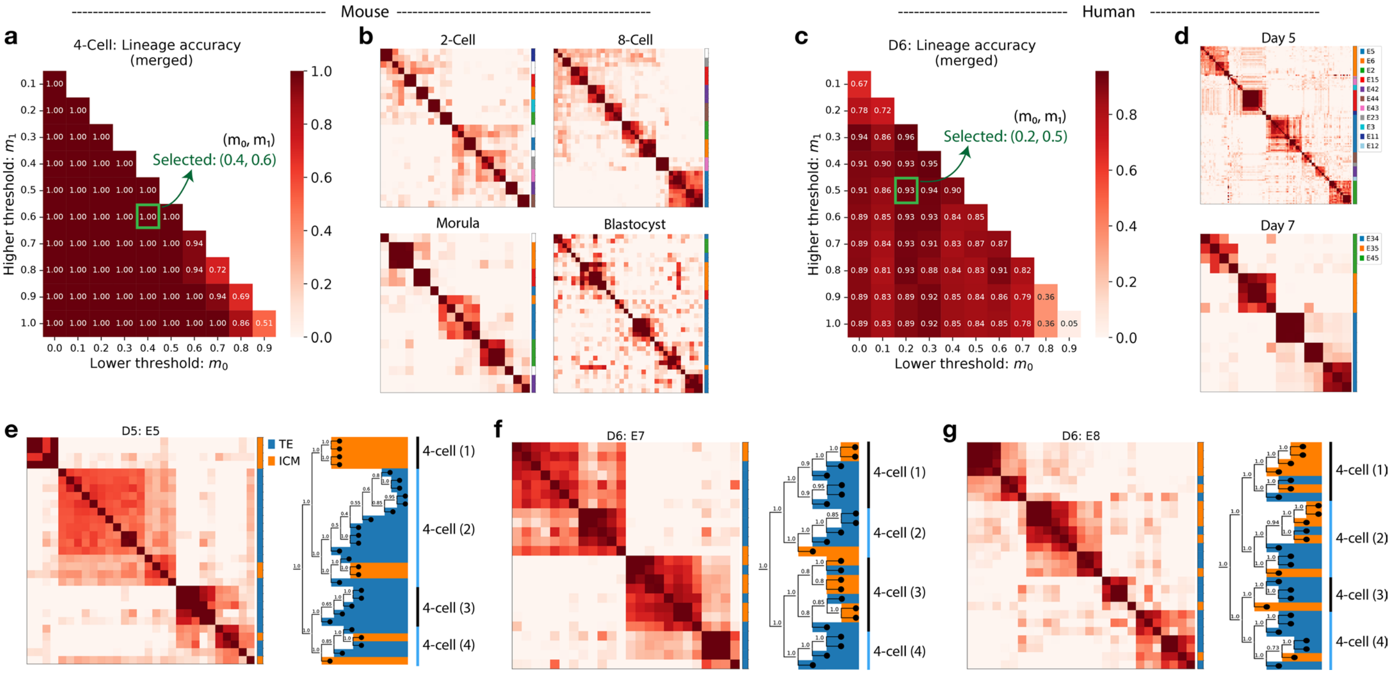
Extended Data Fig. 3 | MethylTree analysis in a heterogeneous population. **a**–**c**, Lineage inference from simulated differentiation. **a**, Heatmap of lineage accuracy as a function of CpG coverage and the variation of the epimutation rate controlled by λ. Here, we simulated differentiation on a haploid genome with a site-specific epimutation rate sampled from a uniform distribution with a maximum value λ. Here, the lineage-specific CpG fraction $\alpha = 0.5$. **b**, Inferred lineage accuracy from simulated differentiation with different fractions of cell-type-specific CpG sites $(1 - \alpha)$. The cell-type signals are not removed. **c**, Lineage accuracy after removing cell-type signals, evaluated at different fractions of lineage-specific CpG sites (α). **d**, Heatmap of methylation similarity associated with fetus_2 from 17 weeks. **e**, Inferred lineage tree from **d**, colored by inferred methyl-clones. **f**, Methylation embedding colored by methyl-clone ID (top) or FGC sub-types (bottom). **g**, Heatmap of MethylTree lineage accuracies associated with Fig. 2l using different region choices. The selected regions associated with (0.3,0.6) were re-used to analyze other datasets in Fig. 2 and Extended Data Fig. 3. **h**, Similarity heatmaps of FGCs and somatic cells from two 7-week human embryos. Left panel: the raw similarity matrix; right panel: after removing cell-type-specific signals.



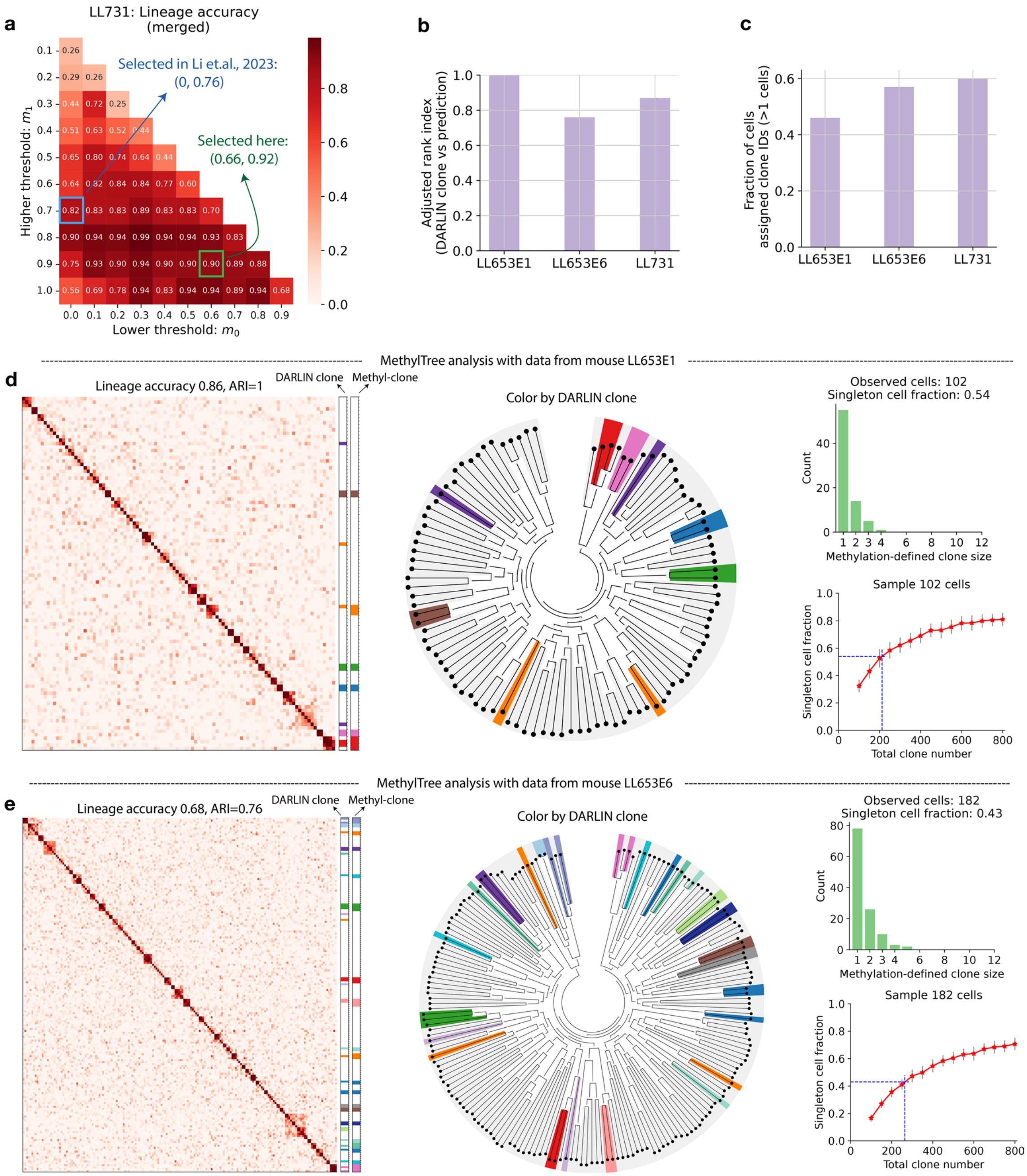
Extended Data Fig. 4 | Analysis of the single-cell multi-omic blood dataset from mouse. a, Heatmap showing the expression of cell-type-specific marker genes (columns) in each annotated cell types (rows) in Fig. 3c. Expression values were column-wise normalized by the highest value in each column. **b**, Bar plot of cell counts of each cell type identified in this dataset. **c**, Histogram of LARRY clone sizes in this dataset. **d**, Heatmap of MethylTree lineage accuracies associated with different region choices on these blood cells. We highlight the parameters used to generate Fig. 3f. **e**, Lineage accuracy computed with non-overlapping bins at different sizes, with either correlation-bias correction or not. **f**, Box plot of lineage accuracies at different genomic coverages. At each coverage, results from all genomic choices are shown. See Fig. 1n for box plot description. **g**, Heatmap of clonal coupling scores computed from the observed LARRY lineage barcodes. **h**, Pseudobulk DNA methylation profiles on genomic regions not specifically related to hematopoiesis. Otherwise, same as Fig. 3m. **i**, Fraction of clone-specific CpG sites in different genomic contexts. These were differentially methylated CpG sites between the two largest clones in this dataset. WCGW: a solo CpG site franked by either A or T; CGI: CpG islands; Prom_CGI: CGI-enriched promoter region (within 2000 bp from transcription starting site); Prom_nonCGI: CGI-depleted promoter region; Genebody: gene body region; LINE: long interspersed nuclear elements; LTR: long terminal repeats. Results from randomly sampled CpG sites are also shown. **, one-sided p-value < 0.01, obtained from directly simulating the null distribution. See Methods.



Extended Data Fig. 5 | Analysis of the single-cell multi-omic blood dataset from human. a, Heatmap showing marker gene expression of each cell type. **b**, Similarity heatmap created the same as in Fig. 4e, but for all the cells passing methylation quality control. **c**, Heatmap of MethylTree lineage accuracies associated with different region choices.

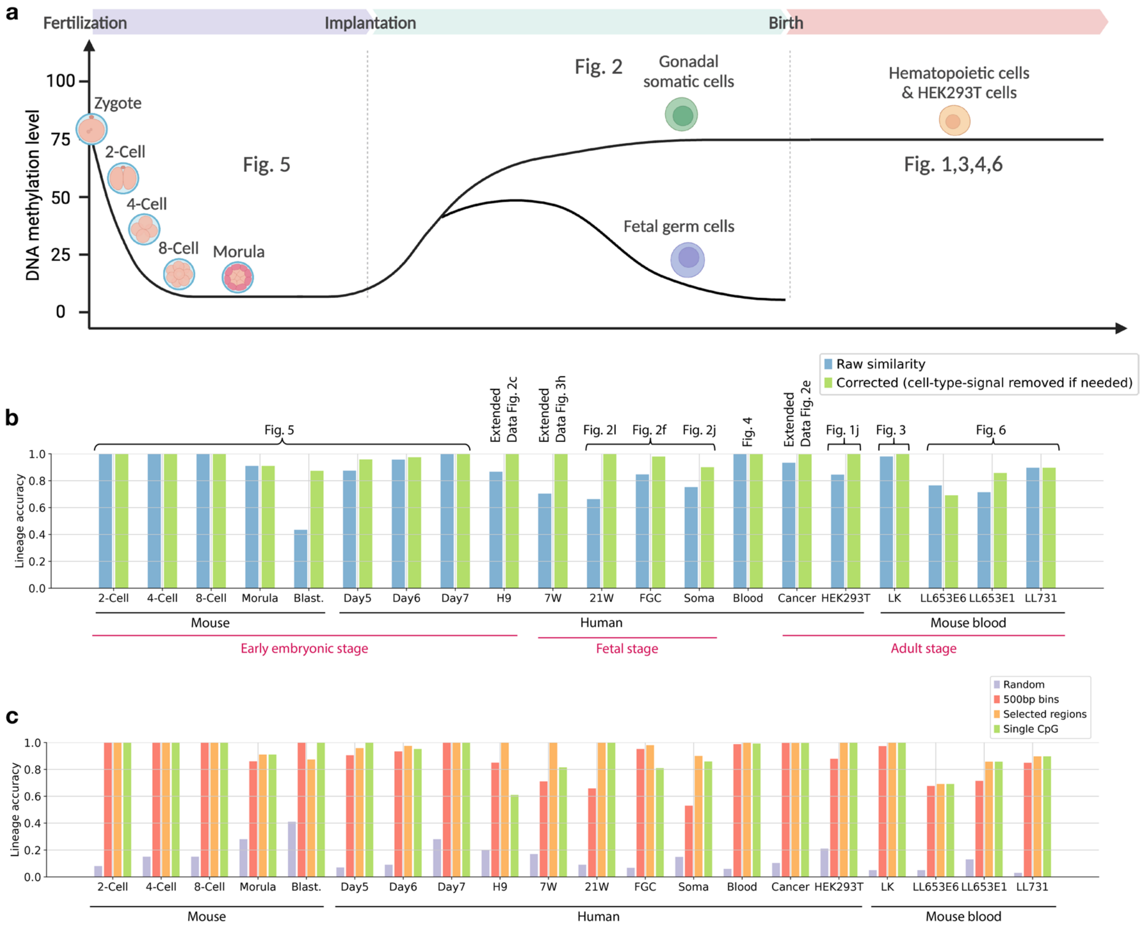


Extended Data Fig. 6 | MethylTree analysis on developing human and mouse embryos. a, Heatmap of MethylTree lineage accuracies associated with different region choices on 4-cell-stage cells from mouse embryos. The selected regions associated with (0.4,0.6) were re-used to analyze mouse datasets from other stages in Fig. 5 and this figure. **b**, Methylation similarity heatmaps of mouse cells from other developmental stages, with the color bar indicating their embryonic origins. **c**, **d**, Same as **a** and **b**, but for cells from human embryos. The selected regions associated with (0.2,0.5) were re-used to analyze human datasets from other stages in Fig. 5 and this figure. **e**–**g**, Methylation similarity heatmaps and reconstructed lineages (with support values from bootstrap sampling) for additional three human embryos, in addition to those shown in Fig. 5h–k. These include E5 from day 5 (**e**), E7 from day 6 (**f**), and E8 from day 6 (**g**).

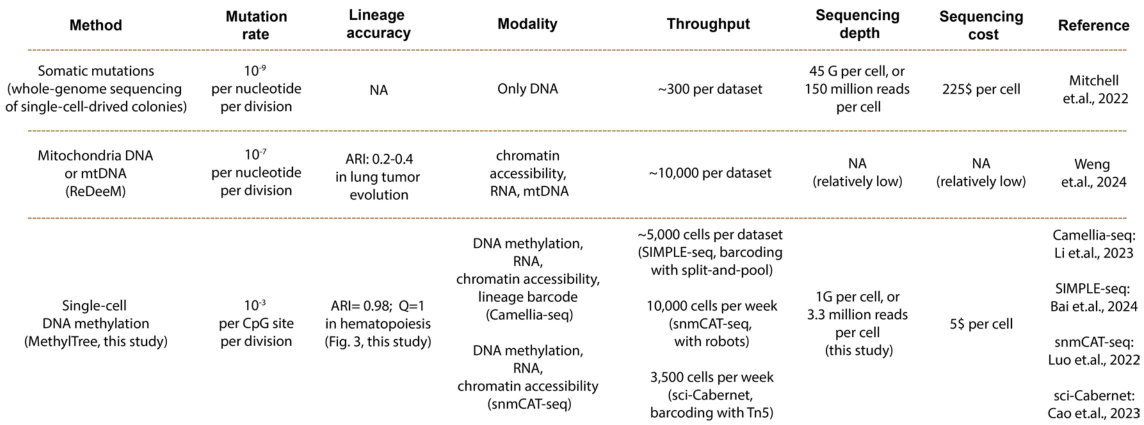


Extended Data Fig. 7 | MethylTree analysis on mouse HSCs. a, Heatmap of MethylTree lineage accuracies associated with different region choices on HSCs from mouse LL731. We highlight parameters used to generate Fig. 6b, and also the choice used in our previous study[13]. The same set of genomic regions was re-used in analyzing the remaining HSC datasets in Fig. 6 and this figure. **b**, Bar plot of adjusted rank index associated with each HSC dataset. **c**, Bar plot of the fraction of cells among the multi-cell methy-clones. **d**, HSC clone number inference on mouse LL653E1. From left to right: methylation similarity matrix, inferred lineage tree, distribution of putative clone sizes (same as Fig. 6e), and HSC clone number inference based on the observed singleton cell fraction (same as Fig. 6f). **e**, HSC clone number inference on mouse LL653E6. Otherwise, same as **d**.

https://doi.org/10.1038/s41592-024-02567-1



Extended Data Fig. 8 | Summary of all datasets analyzed in this study. a, Schematic of the global methylation dynamics over the life time of an individual, created using BioRender.com. Our study analyzed datasets across all three key stages of methylation dynamics, including two global de-methylation waves before birth and a stable period after birth. **b**, Bar plot comparing lineage accuracies from raw similarity or corrected (cell-type-signal removed if needed) similarity across all datasets analyzed in this study. **c**, Bar plot comparing lineage accuracies from using all 500-bp bins, selected genomic regions, all single-CpG sites, or from a randomized cell ordering, across all datasets analyzed in this study. The accuracies from selected-region and single-CpG methods are significantly higher than those from randomization, each with a p-value < 0.0005.



Method	Mutation rate	Lineage accuracy	Modality	Throughput	Sequencing depth	Sequencing cost	Reference
Somatic mutations (whole-genome sequencing of single-cell-drived colonies)	10^{-9} per nucleotide per division	NA	Only DNA	~300 per dataset	45 G per cell, or 150 million reads per cell	225$ per cell	Mitchell et.al., 2022
Mitochondria DNA or mtDNA (ReDeeM)	10^{-7} per nucleotide per division	ARI: 0.2-0.4 in lung tumor evolution	chromatin accessibility, RNA, mtDNA	~10,000 per dataset	NA (relatively low)	NA (relatively low)	Weng et.al., 2024
Single-cell DNA methylation (MethylTree, this study)	10^{-3} per CpG site per division	ARI= 0.98; Q=1 in hematopoiesis (Fig. 3, this study)	DNA methylation, RNA, chromatin accessibility, lineage barcode (Camellia-seq) DNA methylation, RNA, chromatin accessibility (snmCAT-seq)	~5,000 cells per dataset (SIMPLE-seq, barcoding with split-and-pool) 10,000 cells per week (snmCAT-seq, with robots) 3,500 cells per week (sci-Cabernet, barcoding with Tn5)	1G per cell, or 3.3 million reads per cell (this study)	5$ per cell	Camellia-seq: Li et.al., 2023 SIMPLE-seq: Bai et.al., 2024 snmCAT-seq: Luo et.al., 2022 sci-Cabernet: Cao et.al., 2023

Extended Data Fig. 9 | Comparison between different lineage tracing methods in humans. We assume that one Gigabyte bases cost 5$ here. This cost could decrease over time as the technology improves.

nature portfolio

Corresponding author(s): Shou-Wen Wang, Li Li

Last updated by author(s): Oct 20, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	https://github.com/ShouWenWang-Lab/Preprocessing
Data analysis	MethylTree code is available at https://github.com/ShouWenWang-Lab/MethylTree. To reproduce our analysis, please check out our jupyter notebooks at https://github.com/ShouWenWang-Lab/MethylTree_notebooks.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The sequencing data for human blood has been submitted to GSA database under the accession number HRA008624 (https://ngdc.cncb.ac.cn/gsa/). Other sequencing data generated in this study has been submitted to NCBI GEO, with the access number GSE262580. A table containing accession number and analysis parameters for each analyzed dataset in this study is available at Table S1.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender: The human cord blood donor included in the study was a 30-year-old female. No chromosomal or developmental abnormality was reported.

Population characteristics: Only a 30-year-old female was included in this study.

Recruitment: The human cord blood samples were obtained through a collaboration with Beijing Umbilical Cord Blood Bank. Written informed consent was provided by all participants..

Ethics oversight: This study complies with all relevant ethical regulations and was approved by the Ethics Committee of Westlake University (20240222WSW0011) and conducted in accordance to the Declaration of Helsinki protocol.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size: No statistical methods were used to determine the number of samples. Four samples were used to be able to recover lineage in vitro. The number of samples sequenced was chosen to capture clonal relationships among polyclonal cells.

Data exclusions: We excluded some single cell samples from downstream analysis due to poor sequencing quality. This cut-off was decided empirically (see manuscript Methods for further details).

Replication: We replicated the experiment in multiple different datasets across broad range of conditions. All attempts were successful an no other experiments have been performed.

Randomization: No randomization was performed as the experiments were designed to study lineage tracing and no treatments or interventions were involved.

Blinding: Blinding is not relevant, as we are not studying specific pathology or disease, and there are no intervention/ control arms in the study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
□	☒	Antibodies
□	☒	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
□	☒	Animals and other organisms
☒	□	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
□	☒	Flow cytometry
☒	□	MRI-based neuroimaging

Antibodies

Antibodies used: Antibodies used are as follows:

Antibodies used	PE anti-mouse CD117 (c-Kit) (Biolegend, Cat# 105807), FITC anti-mouse Ly-6A/E (Sca-1) (Biolegend, Cat# 108105)
Validation	All antibodies used in the study have been validated by the manufacturers for the application and species relevant for this manuscript. PE anti-mouse CD117 (c-Kit) - The reactivity of the ab was validated by the manufacturer and results shows on https://www.biolegend.com/en-us/products/pe-anti-mouse-cd117-c-kit-antibody-75 FITC anti-mouse Ly-6A/E (Sca-1) - The reactivity of the ab was validated by the manufacturer and results shows on https://www.biolegend.com/en-us/products/fitc-anti-mouse-ly-6a-e-sca-1-antibody-227

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	HEK293T (ATCC, CRL-3216), H9 human embryonic stem cell (WICELL, CVCL_9773)
Authentication	Cell lines derived from credible sources. No authentification done
Mycoplasma contamination	All cell lines tested negative
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in the study

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	C57BL/6 mice (8 weeks, female)
Wild animals	No wild animals were used in the study.
Reporting on sex	For focused on the cellular research, no sex were considered in study.
Field-collected samples	No field collected samples were used in the study.
Ethics oversight	All animal procedures were approved by the Institutional Animal Care and Use Committee of Westlake University (AP#23-093-WSW-2).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	The Lin- Cells were stained with antibodies Kit (CD117-PE, clone 2B8, Biolegend, dilution 1:100) and Sca-1 (Ly6a-FITC, clone D7, Biolegend, dilution 1:100) for 30 min at 4 ºC, protected from light, and washed once in 1 mL of cold DPBS. After final centrifugation (5 min at 300 g, 4 ºC) cells were resupend in 0.5 mL DPBS with 2% FBS and filtered through a 40 µM cell strainer before preceeding with FACS-sorting. Lin-Kit+Sca1- (LK) cells were isolated by FACS on Sony MA900 with a 130µM nozzle.
Instrument	Sony MA900
Software	FlowJo v.10
Cell population abundance	Kit+Sca1- (LK) cells represented 14.3% of the lineage depleted cell population.
Gating strategy	Kit+, Sca1-

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.