Technical details

scLT-kit: a versatile toolkit for automated processing and analysis of single-cell lineage tracing data

1. scLT-statistics module

1.1 Barcoding fraction

In the single-cell lineage tracing experiments, only a subset of cells is labeled with lineage barcodes, while the remainder are not. Further, among the cells with lineage barcoding, only a part of these barcodes is inherited across time points, while the others are lost. Thus, we count the numbers of the cells in these three scenarios and calculate their respective fractions, which are then visualized using pie plots. Fig. S1 presents an example output from the scLT-kit based on the Larry-diff dataset[1]. In the following sections, we will use this dataset as a case study to introduce the scLT-kit pipeline.

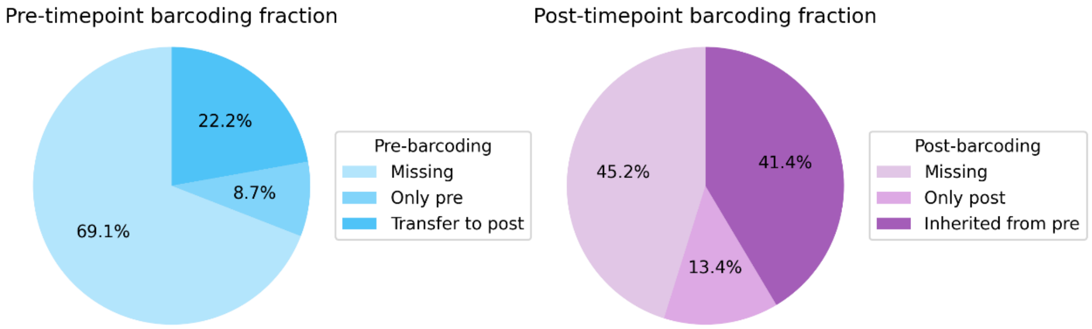

Fig. S1 | The barcoding fraction at pre- and post-timepoint of Larry-diff dataset.

1.2 Clone size

The cells sharing the same lineage barcode can be considered as a clone. For each clone, we count the number of cells in it. In the scLT-kit, the distribution of the clone sizes at both pre- and post-timepoint is presented (Fig. S2).

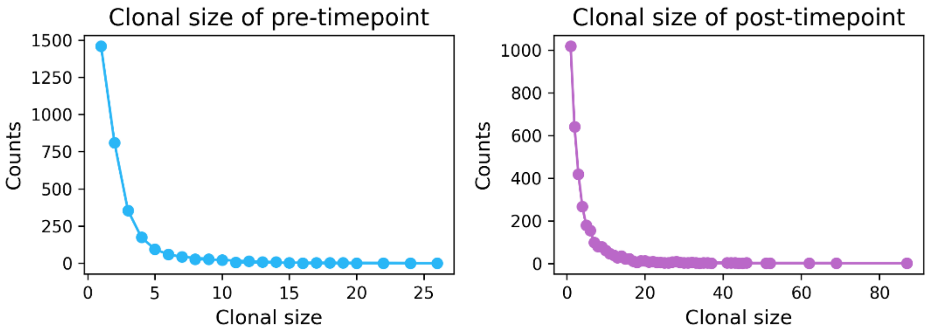

Fig. S2 | The clone sizes at pre- and post-timepoint of Larry-diff dataset.

2. scLT-analysis module

2.1 Clonal heterogeneity

To evaluate the clonal heterogeneity, we compare the transcriptomic similarities among cells within a single clone to those across different clones. These transcriptomic similarities are quantified using the Pearson correlation coefficients between the gene expression values of pairs of cells. In the scLT-kit, we present the distributions of these similarities within each time point and across two time points by histograms (Fig. S3). Further, we compared the distributions of tissue development dataset, such as Larry-diff (Fig. S3A), with those of tumor drug response dataset, such as TraCe-seq dataset[2] (Fig. S3B), and found development datasets exhibited higher transcriptome similarities within clones compared to those across clones, while this phenomenon is a little weak in tumor drug response datasets.

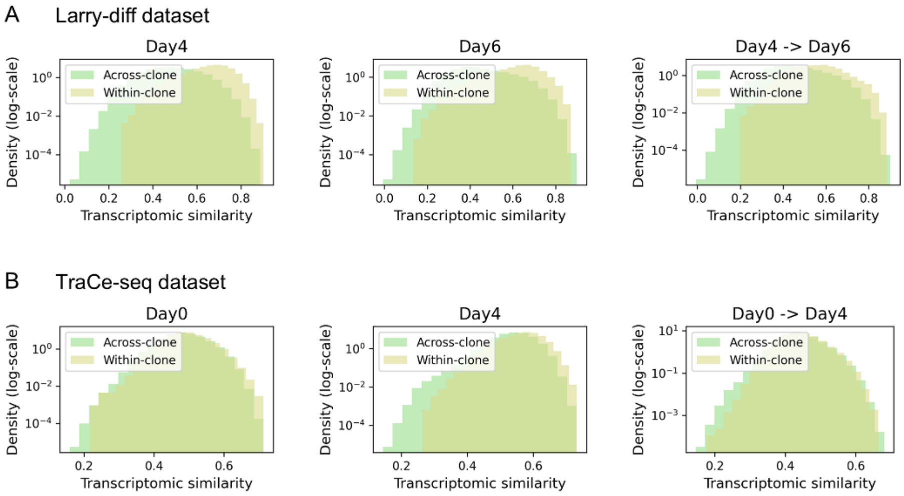

Fig. S3 | Histograms comparing the transcriptomic similarities among cells within the same clone (yellow) versus those across different clones (green) in Larry-diff dataset (**A**) and TraCe-seq dataset (**B**).

2.2 Cell dynamics

According to the lineage tracing information across two time points, we construct a binary adjacency matrix $\boldsymbol{L}^{(N\times M)}$ to represent the cell-cell lineage relationships, where N and M denote the number of cells at the pre- and post-timepoint, respectively. Each entry in $\boldsymbol{L}$ indicates whether the corresponding pair of cells shares a common lineage barcode, thereby reflecting cell-level dynamic relationships.

To derive cluster-level dynamic relationships, we first perform clustering on the cells at each time point individually using the 'Scanpy' package[3] (Fig. S4A). Then, we merge the entries in matrix $\boldsymbol{L}$ to create a "cell-to-cluster" matrix $\boldsymbol{L}_{cls}^{(N\times K)}$, utilizing the clustering results from the post-timepoint, where K denotes the number of the clusters at post-timepoint. In this way, each row in $\boldsymbol{L}_{cls}$ reflects the cluster-level fate possibilities for the corresponding cell. Further, we assign a dominant cluster-level fate to the cell by identifying which post-cluster has the highest number of cells sharing the same lineage barcode as that cell (Fig. S4B). Based on these dominant cluster-level fates, we employ Sankey plots to show the inferred cell dynamics (Fig. S4C).

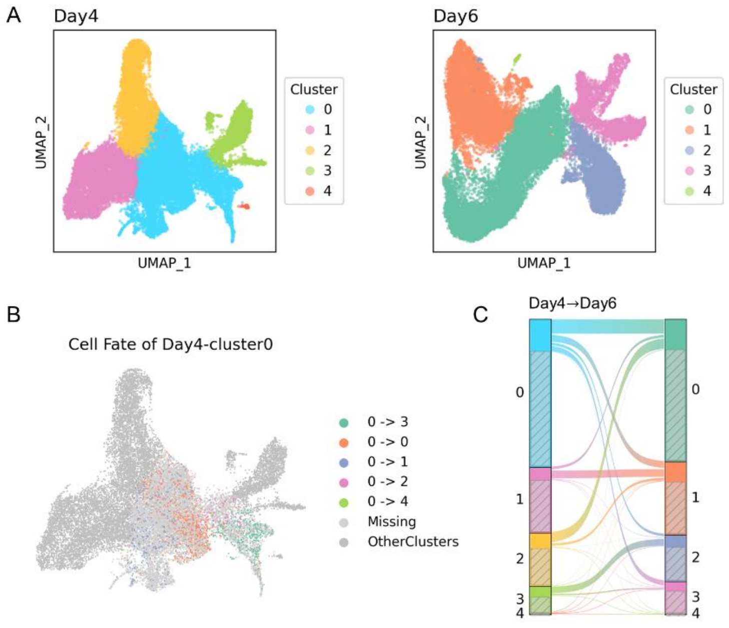

Fig. S4 | (A) UMAP (Uniform Manifold Approximation and Projection) plots showing the clustering labels of the cells in Larry-diff dataset at days 4 and 6. **(B)** The inferred cluster-level dominant fates of the cells in cluster 0 at day 4. **(C)** Sankey plot showing the cluster-level cell dynamics. The shaded areas in light grey indicate cells without lineage link relationships across time points.

2.3 Cell fate diversity

For a selected cell u at pre-timepoint, we denote the set of the five nearest neighbors of u in the two-dimensional UMAP space as $N(u)$, the cluster-level cell fate vector of u in the "cell-to-cluster" matrix as f_u, the dominant fate labels of $N(u)$ as l_u, and the length of a set as $|\cdot|$.

Then, to evaluate the diversity of cell fate in scLT datasets, we design four indicators, including cell fate randomness (CFR), neighbor fate randomness (NFR), neighbor fate consistency (NFC) and neighbor fate similarity (NFS).

Cell fate randomness (CFR) applies information entropy to evaluate the randomness of cell fates. For every cell u, we normalize the cell fate vector f_u into a discrete distribution P_u and calculate the information entropy of the elements in P_u. The CFR degree for a dataset d is defined as the average fate randomness over all cells:

$$\mathrm{CFR}(d) = \frac{1}{|d|}\sum_{u \in d}\left(-\sum_{f_{ui} \in f_u} P_u(f_{ui}) \log P_u(f_{ui})\right).$$

Neighbor fate randomness (NFR) applies information entropy to evaluate the distribution complexity of the dominant fates in neighboring cells. For each cell u, we obtain a discrete distribution Q_u by counting and normalizing the dominant cell fate labels l_u of its neighbors, and then calculate the entropy of Q_u. The NFR for a dataset d is defined as the average NFR over all cells:

$$\mathrm{NFR}(d) = \frac{1}{|d|}\sum_{u \in d}\left(-\sum_{l_{ui} \in l_u} Q_u(l_{ui}) \log Q_u(l_{ui})\right).$$

Neighbor fate consistency (NFC), measuring the consistency of dominant fate labels between neighboring cells, is calculated by the average proportion of nearest neighbors that share the same fate as a selected cell. The NFC for a dataset d is defined as the average NFC over all cells:

$$\mathrm{NFC}(d) = \frac{1}{|d|} \sum_{u \in d} \frac{|\{v \in N(u): \mathrm{fate}(v) = \mathrm{fate}(u)\}|}{|N(u)|}.$$

Neighbor fate similarity (NFS) measures the similarity of fate vectors between neighboring cells. For a selected cell u, we calculate the average Pearson product-moment correlation coefficients (using the “corrcoef” function in the Numpy package) between its fate vectors and those of its nearest neighbors. The NFS for a dataset d is defined as the average NFS over all cells:

$$\mathrm{NFS}(d) = \frac{1}{|d|} \sum_{u \in d} \frac{1}{|N(u)|} \left(\sum_{v \in N(u)} \mathrm{corr}(f_u, f_v)\right).$$

An LT-scSeq dataset exhibiting higher CFR and NFR, alongside lower NFC and NFS, indicates greater randomness in cell fates. Conversely, a dataset with lower CFR and NFR, and higher NFC and NFS, reflects a more explicit dynamic pattern. These indicators provide comprehensive insights into the regularity of fate information derived by lineage tracing technology. We calculate these four indicators on the real datasets, and compare the patterns between the tissue development datasets against tumor drug response datasets (Fig. S5).

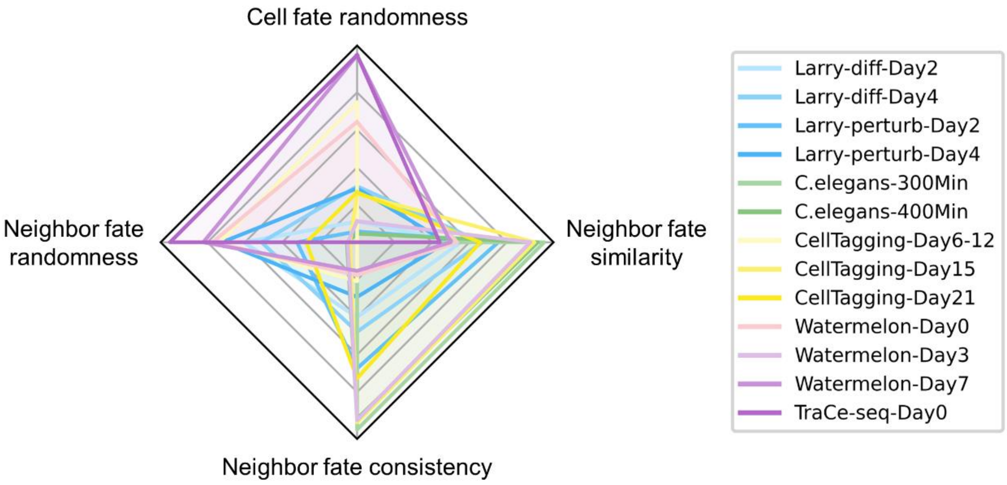

Fig. S5 | Radar plot displaying the indicators of cell fate diversity in developmental datasets and tumor drug response datasets.

2.4 Fate-associated genes

To identify fate-associated genes, we perform differential expression (DE) analysis between the subclusters transitioning to different post-clusters, using the Wilcoxon rank-sum test in the Scanpy package[3]. The DE genes are identified based on the adjusted p-value < 0.05. If the cells are annotated by cell type information, the cell type labels can be used instead of the cluster labels.

Fig. S6 shows an example of Larry-diff dataset, illustrating the transition of undifferentiated cells into different mature cell types (such as neutrophils and monocytes, Fig. S6B). We identify the DE genes associated with the undifferentiated cells that exhibit distinct fates (Fig. S6C).

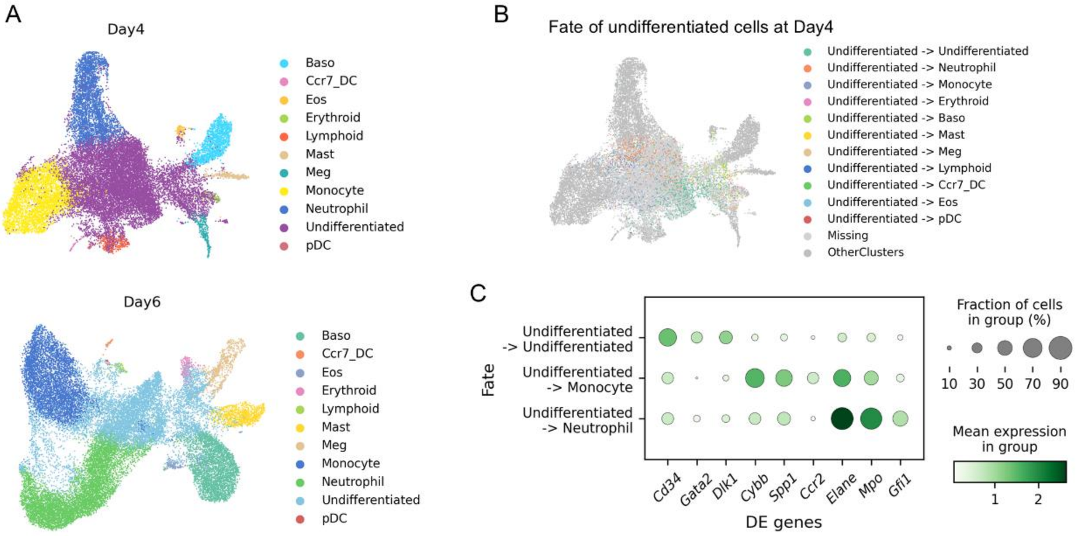

Fig. S6 | (A) UMAP plots showing the cell type labels of the cells in Larry-diff dataset at days 4 and 6. **(B)** The inferred dominant fates at cell type level for the undifferentiated cells at day 4. **(C)** Bubble plot showing the expression profiles of selected DE genes in the undifferentiated cell subclusters that maintain undifferentiated states, or transition to monocytes or neutrophils.

References

1. 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).

2. Chang, M. T. *et al.* Identifying transcriptional programs underlying cancer drug response with TraCe-seq. *Nat Biotechnol* **40**, 86–93 (2022).

3. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).