Published online 14 November 2022

https://doi.org/10.1093/nar/gkac1044

Incorporating cell hierarchy to decipher the functional diversity of single cells

Lingxi Chen[1,2] and Shuai Cheng Li[1,2,*]

[1]Department of Computer Science, City University of Hong Kong, 83 Tat Chee Ave, Kowloon Tong, Hong Kong, China and [2]City University of Hong Kong Shenzhen Research Institute, Shenzhen, 518057, Guangdong, China

Received September 02, 2022; Revised October 13, 2022; Editorial Decision October 17, 2022; Accepted October 21, 2022

ABSTRACT

Cells possess functional diversity hierarchically. However, most single-cell analyses neglect the nested structures while detecting and visualizing the functional diversity. Here, we incorporate cell hierarchy to study functional diversity at subpopulation, club (i.e., sub-subpopulation), and cell layers. Accordingly, we implement a package, SEAT, to construct cell hierarchies utilizing structure entropy by minimizing the global uncertainty in cell–cell graphs. With cell hierarchies, SEAT deciphers functional diversity in 36 datasets covering scRNA, scDNA, scATAC, and scRNA-scATAC multiome. First, SEAT finds optimal cell subpopulations with high clustering accuracy. It identifies cell types or fates from omics profiles and boosts accuracy from 0.34 to 1. Second, SEAT detects insightful functional diversity among cell clubs. The hierarchy of breast cancer cells reveals that the specific tumor cell club drives *AREG-EGFT* signaling. We identify a dense co-accessibility network of *cis*-regulatory elements specified by one cell club in GM12878. Third, the cell order from the hierarchy infers periodic pseudo-time of cells, improving accuracy from 0.79 to 0.89. Moreover, we incorporate cell hierarchy layers as prior knowledge to refine nonlinear dimension reduction, enabling us to visualize hierarchical cell layouts in low-dimensional space.

INTRODUCTION

Cells in the biological system own hierarchical functional diversity, which signifies cell types or states during development, disease, and evolution, up to the biosystem (1,2). The heterogeneity of the cell is observed with nested structures (3). In the tumor microenvironment, infiltrated lymphocytes include B cells and T cells. Furthermore, T cells can be classified into helper T cells and cytotoxic T cells (4). Specific expression of the marker genes *CD4* and *CD8* will strengthen intra-similarity within helper and cytotoxic T cells, respectively, resulting in nested cell structures. The cellular heterogeneity raised by tumor evolution presents another instance (5,6). The copy number gain, neutral, and loss classify tumor cells into aneuploid, diploid, and hypodiploid groups, respectively. Fluctuations of copy numbers in focal genome regions further categorize tumor cells into amplification or deletion subtypes. The cell cycle is a rudimentary biological process for cell replications (7). Human cells undergo a cycle G1–S–G2/M–G1 over a 24-h period, thus the cycling cells have three flat phase labels (G1, S, and G2/M). In addition, the cycling cells have an order that records the pseudo time course in the G1, S, and G2/M phases. The orders and phases reflect a hierarchical structure.

The recent maturation of single-cell sequencing technologies offers opportunities to profile large-scale single cells for their transcriptomics (8), genomics (5), epigenomics (9), etc. These technologies have blossomed revolutionary insights into cellular functional diversity under the aegis of clustering cells with similar molecular characteristics to the same groups (1,2). However, most existing clustering tools generate flat cell group (10–14). Moreover, the periodic pseudotime inference tools neglect the hierarchical structure of cycling cells (15–18). Neglection of the underlying nested structures of cells prevents full-scale detection of cellular functional diversity.

To address the issue, we incorporate *cell hierarchy* to illustrate the nested structure of cellular functional diversity. Cell hierarchy is a tree-like structure with multiple layers that capture cellular heterogeneity. From the root to the tips, the cellular heterogeneity decays. This study focuses on four main layers: global, subpopulation, club, and cell. The global layer is the root that exemplifies the whole cell population, e.g., immune cells. In contrast, the cell groups in the second and third main layers resemble *cell subpopulations* and *cell clubs*, respectively. The cell subpopulation is a broad category of cells, such as B cells and T cells (4). Cell clubs within one cell subpopulation catalog the cellular heterogeneity in a finer resolution; that is, the cells share high functional similarity within a single cell club. For example, T cell subpopulation owns helper and cytotoxic T

[*]To whom correspondence should be addressed. Tel: +852 34429412; Fax: +852 34420503; Email: shuaicli@cityu.edu.hk

cell clubs (4). The tip layer holds individual cells carrying *cell orders*, which signify the dynamic nuance of cell changes within a cell club, e.g., cellular heterogeneity varies along a periodic time course for cells undergoing a cycling process (7).

The actual cell hierarchy is difficult to determine; here, we develop SEAT, Structure Entropy hierArchy deTection, to build a pseudo cell hierarchy utilizing structure entropy to characterize the nested structures in cell–cell graphs. Structure entropy has been proposed in structural information theory to measure the dynamic global uncertainty of complex networks (19), and has benefited several biological fields (20–24). SEAT constructs cell hierarchies from a full-dimensional or dimensionally reduced single-cell molecular profile, and delivers the global-subpopulation-club-cell layers from the hierarchies. We apply SEAT to 36 datasets that cover single-cell RNA (scRNA), single-cell DNA (scDNA), single-cell assay for transposase-accessible chromatin (scATAC), and scRNA-scATAC multiome. SEAT detects the functional diversity of these single-cell omics data with cell hierarchies from three perspectives: cell subpopulation detection, cell club investigation, and periodic cell cycle pseudo-time inference.

Visualizing the functional diversity of single cells is essential since visual inspection is the most direct approach to studying the structure and pattern of cells. Nonlinear dimension reduction is a trending visualization method for high-dimensional biological data (25). Nevertheless, state-of-the-art single-cell visualization tools neglect the nested structure of cells by merely capturing at most two levels (global or local) of cell patterns (26–28). To tackle the issue, SEAT provides a component to embed the cells into a low-dimensional space by incorporating the multiple layers from the cell hierarchy as prior knowledge. Experiments demonstrate that SEAT consistently visualizes the hierarchical layout of these cells in the two-dimensional space for the above single-cell datasets.

MATERIALS AND METHODS

Problem formulation

Constructing cell–cell similarity graph. For a single-cell molecular data tabulated in a matrix, columns and rows refer to cells and their molecular features. For instance, the feature can be a gene or genome region. An entry in the matrix measures the value of the corresponding cell-feature pair, e.g., gene expression, copy number variation, or chromatin accessibility.

We reduce the dimensionality of the single-cell molecular matrix to a low-dimensional matrix $\boldsymbol{X}$ to mitigate the curse of dimensionality. We construct a dense cell–cell similarity graph $G = (V, E)$ with Gaussian kernel $e_{uv} = \exp(-\frac{||\boldsymbol{x}_u - \boldsymbol{x}_v||^2}{2\sigma^2})$ with σ as standard deviation of $\boldsymbol{X}$. Edge weight e_{uv} stands for the similarity between cells u and v in graph G.

Hierarchical coding tree. A coding tree T of a cell–cell graph $G = (V, E)$ is a hierarchical multi-nary partitioning of the cell set V, preserving the nested information in G. For clarity, we use u and v to represent the cells, and μ and ν to represent tree nodes. Each tree node $\mu \in T$ codes a cell subset $U \subset V$. Denote the cell set coded by a node $\mu \in T$ as $V(\mu)$. The root node r codes V and node μ codes U, i.e., $V(r) = V$ and $V(\mu) = U$. Denote the children of μ as $C(\mu)$. The children nodes $C(\mu)$ of the tree node $\mu \in T$ partition the cells represented by μ; that is, $V(\mu) = \bigcup_{i=1}^{|C(\mu)|} V(c_i(\mu)), V(c_i(\mu)) \cap V(c_j(\mu)) = \emptyset, 1 \leq i, j \leq |C(\mu)|, i \neq j$, where $c_i(\mu)$ signifies the i-th child node of μ and $|\cdot|$ denotes cardinality. A leaf node t codes one or multiple cells with a specific order $\pi(t) \in \mathbb{N}^{|V(t)|}$. For each cell $u \in V$ there is a unique leaf node $t \in T$ such that $\{u\} \subseteq V(t)$.

Coding tree represents the hierarchy of subpopulations, clubs, and cells. Given a pool of cells V which own k cell subpopulations, an ideal coding tree T holds k disjoint subtrees rooted at nodes $\Lambda = \{\lambda_1, ..., \lambda_k\}$ which encode k cell sets $\mathcal{P} = \{V(\lambda_1), ..., V(\lambda_k)\}$ that match the cell subpopulations. Denote the subtree $T_\lambda \Subset T$ rooted at λ as *subpopulation tree*. Suppose T_λ has ℓ_λ leaves $\{t_{\lambda,1}, ..., t_{\lambda,\ell_\lambda}\}$, they encode ℓ_λ cell sets $\{V(t_{\lambda,1}), ..., V(t_{\lambda,\ell_\lambda})\}$ that represent cell clubs inside cell subpopulation $V(\lambda)$ in a finer resolution; that is, the cells share high similarity inside one cell subpopulation. In coding tree T, the total ℓ leaves signify the ℓ cell clubs $\mathcal{C} = \{V(t_{\lambda_1,1}), ..., V(t_{\lambda_k,\ell_{\lambda_k}})\}$. Moreover, as cells in each cell club t has a specific order $\pi(t) \in \mathbb{N}^{|V(t)|}$, the ideal coding tree T also presents an overall cell order $\boldsymbol{\pi} = [\pi(t_{\lambda_1,1}), ..., \pi(t_{\lambda_k,\ell_{\lambda_k}})] \in \mathbb{N}^{|V|}$ according to the order of leaves from left to right.

Determining the hierarchy of subpopulations, clubs, and cells is now a hierarchical coding tree construction problem - partitioning the graph G hierarchically to optimize a metric. In this work, the metric is the global dynamical complexity of the graph measured by structure entropy (19–24).

Measuring coding tree with structure entropy. Recall e_{uv} is the edge weight between cells u and v in G. Term the volume of $\mu \in T$ as the sum of degrees of all cells in $V_{(\mu),\, vol(\mu) = \sum_{u \in V(\mu), v \in V} e_{uv} e_{uv}}$. Define $g(\mu)$ as the total weights of edges from cells in $V(\mu)$ to $V - V_{(\mu),\, g(\mu) = \sum_{u \in V(\mu), v \in V - V(\mu)} e_{uv} e}$. If $\mu \neq r$, its structure entropy is

$$\mathcal{S}^T(G; \mu) = -\frac{g(\mu)}{vol(G)} \log_2 \frac{vol(\mu)}{vol(p(\mu))}, \tag{1}$$

where $p(\mu)$ is the parent node of $_\mu$, $vol(G) = \sum_{u,v \in V} e_{uv} e_{uv}$ is the sum of all the edges in the graph, thus $vol(G) = vol(r)$ signifies the volume of the whole graph or the root r. The root r has structure entropy 0; that is, $\mathcal{S}^T(G; r) = 0$.

Denote $t(u)$ as the leaf node where cell u belongs to, the structure entropy of cell u in T is

$$\mathcal{S}^T(G; u) = -\frac{g(u)}{vol(G)} \log_2 \frac{vol(u)}{vol(t(u))}. \tag{2}$$

The structure entropy of graph G coded by tree T is the sum of the structure entropy of all tree nodes and all cells,

$$\mathcal{S}^T(G) = \sum_{\mu \in T} \mathcal{S}^T(G; \mu) + \sum_{u \in V} \mathcal{S}^T(G; u). \tag{3}$$

An ideal coding tree T captures the optimal hierarchy of subpopulations, clubs, and cells. Finding the optimal coding tree T for the graph G is to find the minimum structure entropy $\mathcal{S}^T(G)$ which diminishes the global variance at the random walk of G to a minimum.

Algorithm of SEAT

In previous work, we have proven that for a graph G, there exists a binary hierarchy of minimum structure entropy (23). Thus, SEAT searches the ideal coding tree T from the binary hierarchies (Figure 1A). We first construct a sparse graph G_s from dense graph G, then form cell club hierarchies with minimal structure entropy $\mathcal{S}^T(G_s)$ from sparse graph G_s with agglomerative and divisive heuristics. Then, we search the cell subpopulations by optimizing the structure entropy of the dense graph G constrained by the heuristic hierarchies. Finally, we embed the graph G into a low-dimensional space by adding the global-subpopulation-club layer constraints from cell hierarchy T.

Graph sparsification. We sparsify the dense graph G with k-nearest neighbors (kNNs), resulting in a sparse graph $G_s = (V, E_s)$ with a binary edge weight. If cell u is the k-nearest neighbor of cell v or cell v is the k-nearest neighbor of cell u in original graph G, $e_{uv} = 1$; otherwise $e_{uv} = 0$.

Building cell club hierarchy. With the sparse graph G_s, we form cell club hierarchies with minimal structure entropy $\mathcal{S}^T(G_s)$ with agglomerative and divisive heuristics (Figure 1B).

Agglomerative hierarchy building. The agglomerative hierarchy building consists of three steps: initialization, forming clubs, and building club hierarchy.

We initialize the tree of height one, the root node r has $|V|$ immediate children, where each child node t is a leaf node that covers a single cell of u, $V(t) = \{u\}$. The initialized tree is multi-nary.

We merge the leaf nodes repeatedly to form cell clubs. A leaf has one of the two possible statuses at each iteration, individual or merged. Initially, all the leaves are labeled as individual. Two tree nodes μ and ν are referred to connected if there are inter-node edges between $V(\mu)$ and $V(\nu)$ in sparse graph G_s. We merge an individual leaf μ with its connected sister ν by extracting μ and ν from T and creating a new node μ' which codes all cells in $V(\mu)$ and $V(\nu)$. The new node μ' is a child of root and a leaf labeled as merged. The pair (μ, ν) is chosen by the largest merging structure entropy change $\Delta^m_{se}(\mu, \nu)$ (Supplementary Methods). This merging operation repeats until (i) there is no more individual leaf connected to other sister leaves; or (ii) there is no pair (μ, ν) yields a non-negative structure entropy difference. Then, all leaves are labeled individual, triggering subsequent iterations of the merging procedure until no non-negative structure entropy shift is possible. The above will lead to a multi-nary coding tree T of a height of one and ℓ leaves. We assume each leaf presents a cell club, and the cell order is the merging order.

To form the binary hierarchy of clubs, we iteratively combine sister node pair (μ, ν) of the root by inserting a new node ω as a child of the root and parent of μ and ν. The selection of (μ, ν) is guided by connectivity and the largest combining structure entropy change $\Delta^c_{se}(\omega, \mu, \nu)$ (Supplementary Methods). The combining operation repeats until the hierarchy is a binary coding tree.

Divisive hierarchy building. The second approach is to build the club hierarchy divisively. We initialize the tree with the root node r that codes all cells. The initialized tree has a zero height, with one node as both root and leaf. To form the hierarchy, we repeatedly split the leaf node $t \in T$ into two children guided by maximizing the bipartition structure entropy change $\Delta^s_{se}(t)$. The solution of leaf split is the Fielder vector of the normalized graph Laplacian if the sparse graph G_s is regular (Supplementary Methods). Thus, we heuristically obtain the bipartition according to the values in Fielder vector (29), the cells with smaller Fielder vectors are placed on the left. The split stops if leaf node contains only two cells or $\Delta^s_{se} < \delta$, we set cutoff $\delta = 0.05$. We assume that each leaf presents a cell club, and the value of Fielder vector reflects the cell order. Finally, we end up with a binary hierarchy T with ℓ clubs.

Finding cell subpopulations. Recall that an ideal coding tree T holds k disjoint subpopulation trees rooted at nodes $\Lambda = \{\lambda_1, ..., \lambda_k\}$ which encode k cell sets $\mathcal{P} = \{V(\lambda_1), ..., V(\lambda_k)\}$ that match the cell subpopulations. To find the k subpopulations, we *contract* the heuristic club hierarchy T into a multi-nary tree $\mathcal{T}$ with a height of one (Figure 1C). The contracted tree $\mathcal{T}$ has a root node r holding k leaf children. Each leaf node $t_\lambda \in \mathcal{T}$ maps to a subpopulation tree $T_\lambda \Subset T$ rooted at λ, thus t_λ codes the cells from T_λ, $p(t_\lambda) = r$, $V(t_\lambda) = V(\lambda)$.

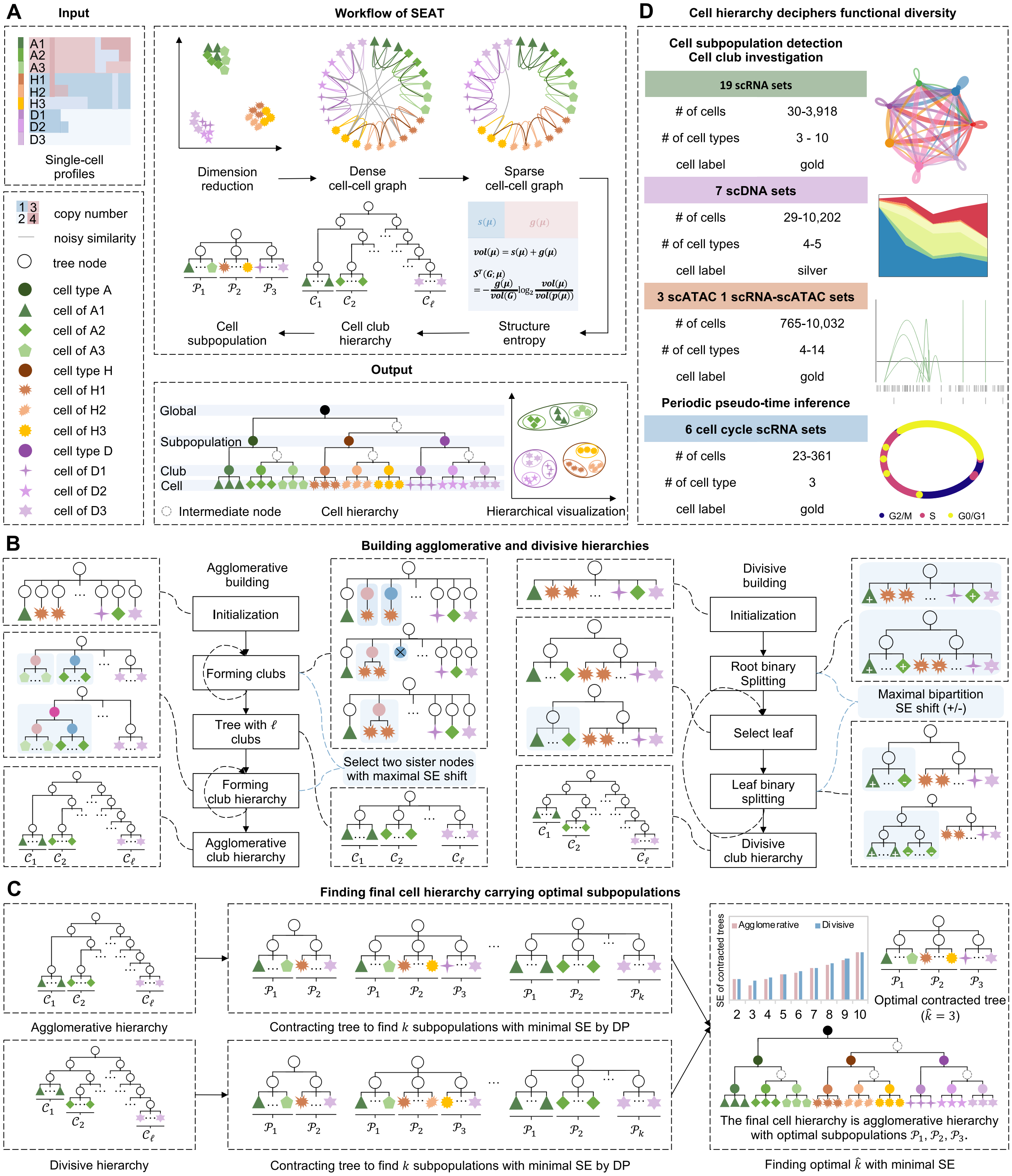
Given the heuristic club hierarchy T, contracting is optimized by minimizing the structure entropy $\mathcal{S}^{\mathcal{T}}(G)$ from dense graph G. The structure entropy associated with contracted tree $\mathcal{T}$ with k leaves focuses on measuring the global variance at the random walk of a dense graph G among k subpopulations, other than the variance in a finer cell-club resolution,

$$\mathcal{S}^{\mathcal{T}}(G) = \sum_{\lambda \in \Lambda} \left[\mathcal{S}^{\mathcal{T}}(G; t_\lambda) + \sum_{u \in V(t_\lambda)} \mathcal{S}^{\mathcal{T}}(G; u) \right]. \qquad (4)$$

To minimize $\mathcal{S}^{\mathcal{T}}(G)$, we adopt a recursive objective $\mathcal{J}(G; \omega, k)$ alongside the club agglomerative or divisive hierarchy T. Assume tree node ω in T has left and right children μ and ν, respectively. Finding k optimal subpopulation trees inside subtree $T_\omega \Subset T$ rooted at ω with minimum $\mathcal{J}(G; \omega, k)$ is equivalent to finding k' and $k - k'$ subpopulation trees inside subtrees $T_\mu \Subset T$ and $T_\nu \Subset T$ rooted at μ and ν such that sum of structure entropy in the contracted tree $\mathcal{T}$ is minimal,

$$\mathcal{J}(G; \omega, k) = \begin{cases} \mathcal{S}^{\mathcal{T}}(G; \omega) + \sum_{u \in V(\omega)} \mathcal{S}^{\mathcal{T}}(G; u), k = 1, \\ \min_{1 \leq k' < k} \{\mathcal{J}(G; \mu, k') + \mathcal{J}(G; \nu, k - k')\}, \end{cases} \qquad (5)$$

where $k = 1$ means ω is the root node of one subpopulation tree, which maps to one leaf node of the contracted tree $\mathcal{T}$.



Cell subpopulation detection
Cell club investigation

19 scRNA sets	
# of cells	30-3,918
# of cell types	3 - 10
cell label	gold
7 scDNA sets	
# of cells	29-10,202
# of cell types	4-5
cell label	silver
3 scATAC 1 scRNA-scATAC sets	
# of cells	765-10,032
# of cell types	4-14
cell label	gold

Periodic pseudo-time inference

6 cell cycle scRNA sets	
# of cells	23-361
# of cell type	3
cell label	gold

Figure 1. The schematic overview of SEAT. (**A**) The workflow of SEAT. (**B**) The algorithm of agglomerative and divisive hierarchy building. (**C**) The algorithm of finding the final cell hierarchy carrying optimal subpopulations. (**D**) The summary of experimental settings.

We solve the contracting objective using dynamic programming. We record the minimal structure entropy $\mathcal{J}(G;\omega,k)$ for finding k optimal subpopulations in a bottom-up way; that is, calculating from leaves to root. We trace back recursively to obtain the optimal cut-off k' for each node starting from the root. If $\hat{k}'_{\mu} = 1$ for one left child or $\hat{k}'_{\nu} = k-1$ for a certain right child at that state, one subpopulation $V(\mu)$ or $V(\nu)$ is found (Supplementary Methods). In this way, we obtain the contracted tree $\mathcal{T}$ with k leaves representing k cell subpopulations.

Finding final cell hierarchy carrying optimal subpopulations. For $1 \leq k \leq K$ where K is constant number, the optimal $\hat{k}$ associated with the minimal structure entropy is the optimal cut-off k' for root node, $\hat{k} = \arg\min_{1<k\leq K}\{\mathcal{J}(G;r,k)\}$.

The agglomerative and divisive hierarchies might have different hierarchical structures. The optimal subpopulations are subpopulations with less structure entropy (Figure 1C and Supplementary Methods). We choose the cell hierarchy carrying optimal subpopulations as the final cell hierarchy.

Obtaining and visualizing cell order. We find the cell hierarchy T by minimizing the structure entropy of the sparse cell–cell graph. Given the cell hierarchy T, we obtain the cell order $\boldsymbol{\pi} \in \mathbb{R}^{|V|}$ with an in-order traversal and visualize the cell order periodically with an oval shape (Supplementary Methods).

Hierarchical visualization. To convert the cell–cell similarity graph G into d-dimensional latent space $\boldsymbol{Y} \in \mathbb{R}^{n\times d}$ for visualization, state-of-the-art tool UMAP (26) adopts a cross-entropy (CE) objective,

$$CE(G) = \sum_{u,v\in G} p_{uv}\log\frac{p_{uv}}{q_{uv}} + (1-p_{uv})\log\frac{1-p_{uv}}{1-q_{uv}}. \quad (6)$$

Here, p_{uv} and q_{uv} signifies the similarity of cells u and v in original graph G and the latent space, respectively. q_{uv} is smoothly approximated by $q_{uv} = (1 + a(||\boldsymbol{y}_u - \boldsymbol{y}_v||_2^2)^b)^{-1}$, where a and b are constrained by a hyper-parameter *min-dist*, the effective minimum distance between cells in latent space.

In this study, we adjust the above embedding strategy by incorporating the final cell hierarchy. Recall that the cell partition $\mathcal{P}$ and $\mathcal{C}$ corresponds to the k and ℓ cell subpopulations and clubs, respectively. Assume cell partition $\mathcal{I} = \{V\}$ contains the one global cell population. Based on the cell partition $\mathcal{H} \in \{\mathcal{I},\mathcal{P},\mathcal{C}\}$, we set the inter-connections between different cell groups to zero, resulting in a graph $G_{\mathcal{H}}$ that focuses on the cell–cell similarity inside one cell group.

We minimize the disparity of cell–cell similarity between the embedding space and $G_{\mathcal{H}}$ with the objective

$$\mathcal{J}(G) = \sum_{\mathcal{H}\in\{\mathcal{I},\mathcal{P},\mathcal{C}\}} CE(G_{\mathcal{H}}) \times \theta_{\mathcal{H}}, \quad (7)$$

where hyper-parameters $\theta_{\mathcal{H}}$ are the training weights of different cell partition resolutions obtained from cell hierarchy.

We initialize the low-dimensional embedding $\boldsymbol{Y}$ with graph Laplacian (30) of $G_{\mathcal{P}}$, make *min-dist* equals 0.1, set $\theta_{\mathcal{I}} = 1$, $\theta_{\mathcal{P}} = 1$, $\theta_{\mathcal{C}} = 1$, and minimize $\mathcal{J}(G)$ to convergence with Adam gradient descent.

Outlier detection. Cellular abnormalities may distort the entire cell hierarchy, thus affecting the efficacy of cell subpopulation and club detection, cell cycle pseudo-time inference, and hierarchical visualization. Thus, we have implemented the average kNN outlier detection. We calculate the mean distance $\boldsymbol{d} \in \mathbb{R}^n$ given the single-cell molecular representation of n cells. d_i is the mean distance of i-th cell to its k-nearest neighbors. By default, we consider the cell with an average kNN distance d exceeding a distance cutoff 0.5 as the outlier. We also provide a distance percentile cutoff strategy, we regard the cell with an average kNN distance d surpassing a distance percentile cutoff (e.g., 95th percentile) as an outlier. The detected outliers will be assigned to label –1 and excluded from the cell hierarchy building.

Time complexity of SEAT. Under the graph G with n cells, the time complexity of SEAT is $O(n\log n)$ (Supplementary Methods).

Experiment Setting

scRNA data. We collect nineteen scRNA datasets with gold standard cell type labels (31–43), the description of the datasets and the download links are in Supplementary Table S1 and Supplementary Method. For these scRNA datasets, the dimension reduction transformer is UMAP (26). We adopt Seurat 'FindAllMarkers' function (44) for differential expression analysis. The $\log_2$ fold change, $\log_2$(FC), of the average expression between two groups is measured. The fold change significance P-value is evaluated by the Wilcoxon Rank Sum test, and the adjusted P-value is calculated with Bonferroni correction. The filtering criteria are $\log_2(\mathrm{FC}) \geq 0.25$, P-value < 0.05, and adjusted P-value < 0.05. Cell–cell communication analysis is conducted with CellChat (45) with default database and parameters. Any ligand-receptor interaction with less than ten supporting cells is filtered.

We also collect six scRNA datasets with gold standard cell cycle labels (Supplementary Table S2). Dataset H1-hESC has 247 human embryonic stem cells (hESCs) in G0/G1, S, or G2/M phases identified by fluorescent ubiquitination-based cell cycle indicators (46). The count expression profile and cell cycle labels are obtained with accession code GSE64016. Datasets mESC-Quartz and mESC-SMARTer have 23 and 288 mouse embryonic stem cells (mESCs) sequenced by Quartz-seq and SMARTer, respectively (47,48). Their G0/G1, S, and G2/M phases are labeled by Hoechst staining. The count expression profiles and cell cycle labels are obtained with accession codes GSE42268 and E-MTAB-2805. Datasets 3Line-qPCR_H9, 3Line-qPCR_MB, and 3Line-qPCR_PC3 own 227 H9 cells, 342 MB cells, and 361 PC3 cells, respectively. The cell cycle stages G0/G1, S, and G2/M are marked by Hoechst staining (32). The raw log2 count expression profiles and cell labels are from the paper's dataset S2. The imputation and dimension reduction are conducted by SMURF (49) and UMAP (26). We adopt Seurat (44) for differential expression analysis as described above. Cell-cell communication analysis is conducted with CellChat (45) with default database and parameters. Any ligand-receptor interaction with less than ten supporting cells is filtered. Gene Ontology (GO) is performed with ShinyGO 0.76 (50).

scDNA data. We collect seven scDNA datasets (Supplementary Table S1). Navin_T10 contains 100 cells from a genetically heterogeneous (polygenetic) triple-negative breast cancer primary lesion T10, including five cell subpopulations: diploid (D), hypodiploid (H), aneuploid 1 (A1), aneuploid 2 (A2), and pseudo-diploid (P) (51). Navin_T16 holds 52 cells from genetically homogeneous (monogenetic) breast cancer primary lesion T16P and 48 cells from its liver metastasis T16M, including four cell subpopulations: diploid (D), primary aneuploid (PA), metastasis aneuploid (MA), and pseudo-diploid (P) (51). The Ginkgo copy number variation (CNV) profiles of Navin_T10 and Navin_T16 are downloaded from http://qb.cshl.edu/ginkgo (52). The silver standard array comparative genomic hybridization (aCGH) data of Navin_T10 and Navin_T16 are downloaded with GEO accession code GSE16607 (53).

Dataset 10x_breast_S0 is a large-scale 10x scDNA-seq set without known cell population labels, where 10,202 cells from five adjacent tumor dissections (A, B, C, D, and E) of triple-negative breast cancer are sequenced. The Bam files are downloaded from 10x official site https://www.10xgenomics.com/resources/datasets. We inferred the total CNV profile utilizing Chisel (54).

Ni_CTC owns 29 circulating tumor cells (CTCs) across seven lung cancer patients (55). McConnel_neuron profiles 110 cells from human frontal cortex neurons, with an extensive level of mosaic CNV gains and losses (56). Lu_sperm has 99 sperm cells with chrX-bearing, chrY-bearing, and aneuploid groups (57). Wang_sperm contains single-cell sequencing data on 31 sperm cells with CNV gains and losses (58). The Ginkgo CNV profiles of these datasets are downloaded from http://qb.cshl.edu/ginkgo (52).

scATAC and scRNA-scATAC multiome data. We collect three public scATAC-seq data as benchmarking sets with gold standard cell type labels (Supplementary Table S1). scatac_6cl is a mixture of six cell lines (BJ, GM12878, H1-ESC, HL60, K562, and TF1) with 1,224 cells (59). Hematopoiesis owns 2,210 single-cell chromatin accessibility profiles from eight human hematopoiesis cell subpopulations (CLP, CMP, GMP, HSC, LMPP, MEP, MPP, and pDC) (60). T-cell composes of four T-cell subtypes (Jurkat_T_cell, Naive_T_cell, Memory_T_cell, and Th17_T_cell) with a total of 765 cells (61).

We collect a multiome of scRNA and scATAC dataset PBMC (human peripheral blood mononuclear cells) with 10,032 cells across fourteen cell types.

We downloaded the scOpen (62) processed accessibility profiles and cell labels from https://github.com/CostaLab/scopen-reproducibility. UMAP (26) embedded data are used to construct the kNN graphs. We adopt Cicero (63) to explore the dynamically accessible element status in different scatac_6cl GM12878 cell clubs.

Evaluating cell subpopulation detection. To detect cell subpopulations, some clustering methods require the number of clusters prespecified, while others can determine the number of clusters automatically. The SEAT package supports both. Our package requires no prespecified number of clusters by default, that is, SEAT(sub). If the number of clusters required is k, we denote the method as SEAT(k). When the context is clear, we refer to them as predefined-k and auto-k modes, respectively.

In the predefined-k mode, we access the clustering accuracy of SEAT agglomerative hierarchy and divisive hierarchy with predefined cluster number k given by the actual number of ground truth cell types, namely Agglo(k) and Divisive(k). We regard the clustering result with a lower structure entropy from agglomerative and divisive hierarchies as SEAT(k). Baselines are hierarchical clustering (HC) with four linkage strategies (ward, complete, average, and single) (12), K-means (11), and spectral clustering (10). We run them with default parameters. As the leading tool for single-cell clustering, Louvain (13) and Leiden (14) automatically detect how many communities are inside the cell–cell similarity graph. They obtain different numbers of communities at various resolutions. To benchmark Leiden and Louvain in the predefined-k setting, namely Leiden(k) and Louvain(k), we heuristically adjusted the resolution 20 times to see if the number of communities was the same as the predefined cluster number k.

As the predefined k is undetermined in most real-world scenarios, we evaluate the auto-k clustering efficacy of SEAT cell hierarchy, agglomerative hierarchy, and divisive hierarchy, namely SEAT(sub), Agglo(sub), and Divisive(sub). The baselines are Leiden and Louvain with default parameters. We also assess the clustering obtained from agglomerative and divisive hierarchy clubs, namely Agglo(club) and Divisive(club).

Adjusted Rand index (ARI) (64) and adjusted mutual information (AMI) (65) are adopted as clustering accuracy. They measure the concordance between clustering results and ground truth cell types. Perfect clustering has a value of 1, while random clustering has a value less than or near 0.

Evaluating cell cycle pseudo-time inference. SEAT cell hierarchy, agglomerative hierarchy, and divisive hierarchy generate cell orders representing the cell cycle pseudo-time for scRNA data, namely, SEAT(order), Agglo(order), and Divisive(order). We access the pseudo-time inference accuracy of SEAT given by the actual order of ground truth cell cycle phases. Benchmark methods are hierarchical clustering (HC) with four linkage strategies (ward, complete, average, and single) (12), since an in-order traversal of HC hierarchies also generates cell orders. Furthermore, we benchmark our method with four state-of-the-art tools predicting the cell cycle pseudo-time, CYCLOPS (15), Cyclum (16), reCAT (17), and CCPE (18). We run them with default parameters. CCPE fails the tasks when we follow its GitHub instruction, so we exclude CCPE for final comparison.

The change index (CI) is used to quantitatively assess the accuracy of cell pseudo-time order against known cell cycle phase labels (17). An ideal cell order changes label $k - 1$ times, where $k = 3$ is the ground truth cell cycle phase number. The change index is defined as $1 - \frac{c-(k-1)}{n-k}$, where c counts the frequency of label alters between two adjacent cells, and n is the number of cells. A value of 0 suggests the cell order is utterly wrong with $c = n - 1$, while 1 indicates a complete match between cell order and ground truth cell cycle phase with $c = k - 1$.

Evaluating hierarchical visualization. We evaluate the efficacy of SEAT hierarchical visualization, SEAT(viz), with state-of-the-art visualization tools UMAP (26), TSNE (27), and PHATE (28). The dense cell–cell similarity graph G is used as input. UMAP, TSNE, and PHATE are run with default parameters.

Evaluating cell outlier detection. We simulate the gene expression profiles of 500 cells with five subpopulations using Splatter (66). We randomly produce 20 cell outliers with gene expression disparting from all five subpopulations. We evaluate SEAT cell subpopulation detection (i) with and without the average kNN outlier detection; (ii) with different combinations of parameters (nearest neighbor number, distance cutoff, and distance percentile cutoff). The outliers are considered as a distinct group, thus the ARI and AMI are used to measure the clustering accuracy.

RESULTS

Overview of SEAT

SEAT builds a cell hierarchy annotated with global-subpopulation-club-cell layers computationally from single-cell data (Figure 1). First, SEAT constructs a pair of dense and sparse cell–cell similarity graphs from a full-dimensional or dimensionally reduced single-cell molecular profile (Figure 1A). Second, we detect cell clubs, determine the order of cells within each cell club, and build the pseudo club hierarchies by minimizing the structure entropy of the sparse graph with agglomerative (Agglo) and divisive (Divisive) heuristics (Figure 1B, Materials and Methods). We term the cell clubs and orders derived from agglomerative and divisive hierarchies as Agglo(club), Agglo(order), Divisive(club), and Divisive(order). Next, we use dynamic programming to find optimal subpopulations from agglomerative and divisive hierarchies, namely, Agglo(sub) and Divisive(sub). We choose the hierarchy carrying the lower subpopulation structure entropy as the final cell hierarchy (Figure 1C, Materials and Methods). Hence, SEAT outputs the final cell hierarchy carrying with subpopulations, clubs, and orders, namely, SEAT(sub), SEAT(club), and SEAT(order) (Figure 1A). Furthermore, by incorporating hierarchical cell partition layers, SEAT provides a component, SEAT(viz), to embed cells into a low-dimensional space while preserving their nested structures for improved visualization and interpretation (Figure 1A).

Cell hierarchy catalogs functional diversity at the subpopulation and club levels from scRNA data

We have applied SEAT to nineteen scRNA datasets carrying gold standard cell type labels. The first nine sets are cell line mixtures, including p3cl (31), 3Line-qPCR (32), sc_10x, sc_celseq2, sc_dropseq, sc_10x_5cl, sc_celseq2_5cl_p1, sc_celseq2_5cl_p2, and sc_celseq2_5cl_p3 (33). We have four datasets Yan (34), Deng (35), Baise (36), and Goolam (37) which sequence single cells from human or mouse embryos at different stages of development (zygote, 2-cell, early 2-cell, mid 2-cell, late 2-cell, 4-cell, 8-cell, 16-cell, 32-cell, early blast, mid blast, and late blast). The last six datasets are Koh (38), Kumar (39), Trapnell (40), Blakeley (41), Kolodziejczyk (42), and Xin (43), which profile different cell types in single-cell resolution. To access the efficacy of SEAT in cell subpopulations detection, we utilize the adjusted rand index (ARI) (64) and adjusted mutual information (AMI) (65) as clustering accuracy, and benchmark SEAT with state-of-the-art clustering tools (spectral clustering (10), K-means (11), hierarchical clustering (12), Louvain (13), and Leiden (14)) with predefined-k and auto-k modes (Materials and Methods, Supplementary Figures S1–S3). In predefined-k mode, SEAT(k) demonstrates comparable or higher clustering accuracy compared to other clustering baselines on most datasets (Figure 2A). Notably, Louvain(k) and Leiden(k) are unable to generate a clustering that exactly matches the number of ground truth labels after 20 different resolution trials for the Goolam and Kolodziejczyk (Figure 2A and Supplementary Figure S2). Under the auto-k mode, SEAT(sub) outperforms Louvain and Leiden on all nineteen sets. The clustering accuracies of SEAT(sub) are comparable to or better than the best clustering results with predefined-k clustering tools with the ground truth cluster number provided. This is attributed to the fact that SEAT(sub) finds a cluster number close to the ground truth (Figure 2B). Louvain and Leiden have the lowest clustering accuracy because they prefer more clusters. The two-dimensional data embedded by UMAP from full-dimensional single-cell expression profiles are inputs of all clustering tools; and the visualizations of them show that the ground truth labels are mixed for the majority of datasets (Supplementary Figures S4 and S5), explaining the low clustering accuracy of both predefined-k and auto-k clustering tools.

SEAT offers hierarchical structures of cells to study cellular functional diversity. We use differential gene expressions to investigate the biological interpretations of these hierarchies. In Supplementary Figures S6 and S7, differentially expressed genes ($P < 0.05$) between cell hierarchy clubs reveal distinct patterns that match ground truth cell subpopulations. Furthermore, visible marker gene patterns reveal the functional diversity among cell clubs within one cell subpopulation. We focus on the top five differentially expressed genes for each dataset (Supplementary Figures S8–S11). As the subpopulation detection accuracy of agglomerative hierarchy is 1 for p3cl dataset, we investigate the functional diversity revealed from the agglomerative hierarchy other than the divisive hierarchy. The agglomerative hierarchy revealed three cell subpopulations for p3cl, which correspond to the three ground truth cell types, basal (*KRT81*), luminal (*TFF1*), and fibroblast (*COL1A2* and *VIM*) (Figure 2C). We observe that each of the basal, luminal, and fibroblast has two major subclasses, controlled by the expression of cell cycle genes (*HIST1H4C*, *CDC20*, *CCNB1*, and *PTTG1*). Cell-cell communication analysis finds a total of 109 significant ($P < 0.05$) ligand-receptor (LR) pair interactions among seven agglomerative hierarchy clubs for breast cancer basal-like epithelial cell line in p3cl. The LR interactions belong to nine signaling path-

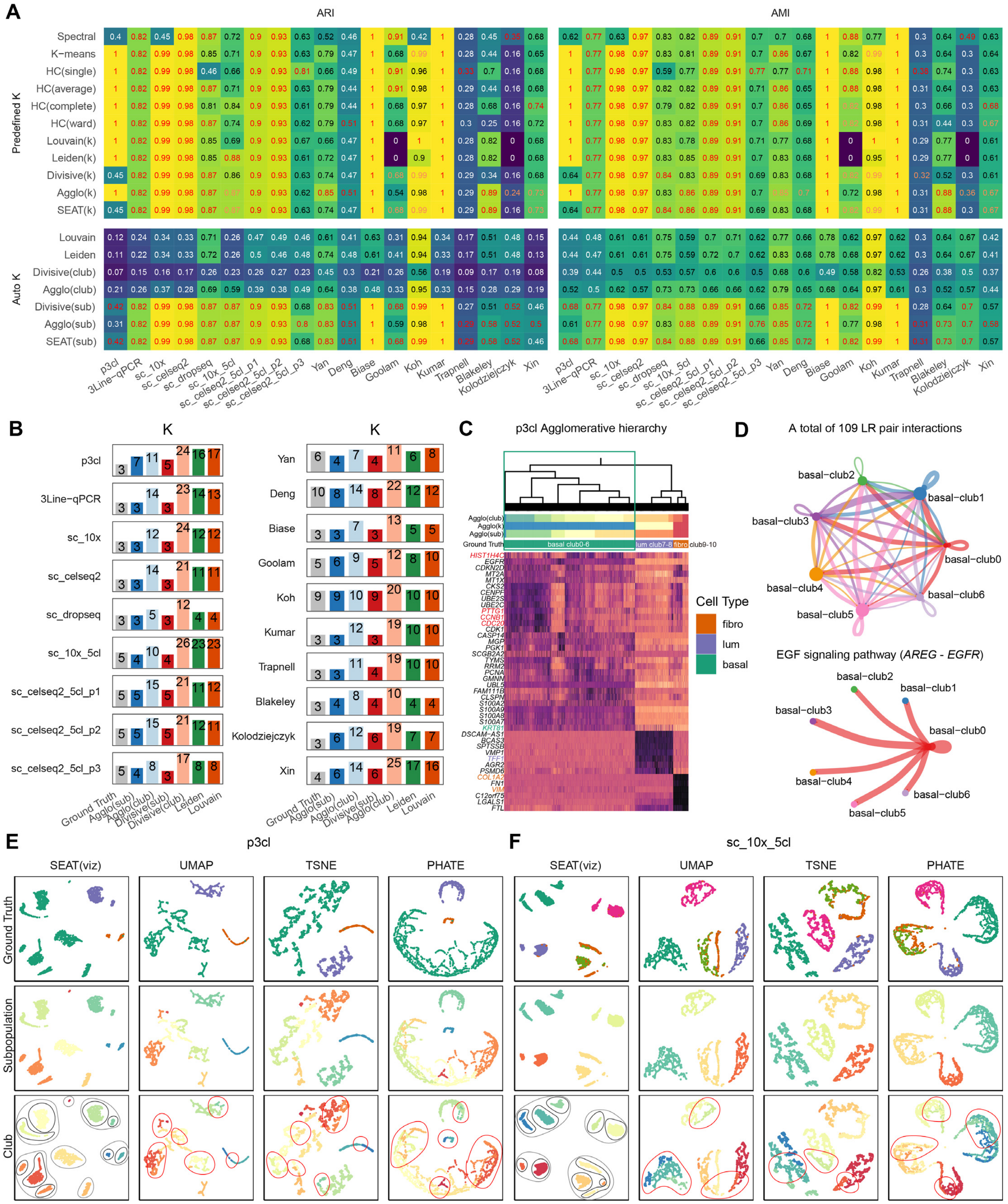


Figure 2. Applying SEAT on nineteen scRNA datasets. (**A**) The adjusted rand index (ARI) and adjusted mutual information (AMI) of predefined-k and auto-k clustering tools. The best score is colored red for each dataset in predefined and auto clustering benchmarking separately. If SEAT gets second place, we color the score orange. Spectral: spectral clustering. HC(single), HC(average), HC(complete), and HC(ward): hierarchical clustering with single, average, complete, and ward linkage. Louvain(k) and Leiden(k): Louvain and Leiden in predefined-k mode. Divisive(k) and Agglo(k): the cell subpopulations from divisive and agglomerative hierarchy in predefined-k mode. SEAT(k): the cell subpopulations from SEAT cell hierarchy in predefined-k mode. Divisive(club) and Agglo(club): the cell clubs from the divisive and agglomerative hierarchy. Divisive(sub) and Agglo(sub): the cell subpopulations from divisive and agglomerative hierarchy in auto-k mode. SEAT(sub): the optimal subpopulations from SEAT cell hierarchy in auto-k mode. (**B**) The number of subpopulations detected for auto-k clustering tools. (**C**) The top five differentially expressed genes in agglomerative hierarchy clubs for p3cl. (**D**) The cell–cell communications among seven agglomerative hierarchy clubs for breast cancer basal-like epithelial cell line in p3cl. LR: ligand-receptor. (**E**, **F**) SEAT(viz), UMAP, TSNE, and PHATE plots for p3cl and sc_10x_5cl. The cells are colored with subpopulations, clubs, and ground truth. The gray and black circles in the SEAT(viz) plot indicate the subpopulation and club boundaries, respectively. In UMAP, TSNE, and PHATE plots, the red circles mark the unclearly segregated cell clubs. SEAT(viz): the hierarchical visualization from SEAT cell hierarchy.

ways AGRN, CD99, CDH, EGF, JAM, LAMININ, MK, NECTIN, and NOTCH (Figure 2D and Supplementary Figure S12). In particular, there is a distinct breast cancer cell club (basal-club0) that drives *AREG* -*EGFR*, an oncogenic signaling (67) in breast cancer, to all basal cells, resulting in a high level of *AREG* activated *EGFR* expression (Figures 2C and 2D). The two cell clubs from the luminal subpopulation have six significant ($P < 0.05$) LR interactions involving MK, SEMA3, and CDH signaling pathways (Supplementary Figure S13). The fibroblast has three significant ($P < 0.05$) LR interactions, including two signaling pathways FN1 and ncWNT (Supplementary Figure S13). The cell club fibro-club10 releases *WNT5B* and then binds *FZD7* from fibro-club9, consistent with the observation that ncWNT is the predominant signaling pathway in skin fibroblasts (45).

Visualizations of two-dimensional data by UMAP from full-dimensional single-cell expression profiles reveal a dense layout (Supplementary Figures S4 and S5). The ground truth cell subpopulations are indistinctly separated in some high clustering accuracy datasets, and the cell clubs are densely arranged in each subpopulation clump. Here, we check whether SEAT hierarchical visualization eliminates the dense layout of clubs. We use the cell–cell graph constructed by SEAT as input and execute SEAT(viz), UMAP, TSNE, and PHATE, independently. In Figures 2E, 2F, and Supplementary Figures S14–S18, SEAT(viz), UMAP, TSNE, and PHATE separate the ground truth cell type for most datasets. It should be noted that the patterns from SEAT(viz), UMAP, TSNE, and PHATE also correspond to the subpopulation layer annotations, validating SEAT subpopulation finding efficacy. At the cell club level, SEAT(viz) shows a clear layout of cell clumps that correspond to the cell hierarchy; each cell club owns a distinct clump, and the distance between clubs belonging to the same subpopulation is within proximity. Although UMAP, TSNE, and PHATE capture the local structures of the clubs, the cell clubs marked with red circles are unclearly segregated.

Cell hierarchy deciphers periodic cell cycle pseudo-time from single-cell data

We collect six scRNA cell cycle datasets, H1-hESC (46), mESC-Quartz (47), mESC-SMARTer (48), 3Line-qPCR_H9, 3Line-qPCR_MB, and 3Line-qPCR_PC3 (32) with gold standard G0/G1, S, or G2/M stages, and then build the cell hierarchies (Supplementary Figure S19). In predefined-k and auto-k clustering benchmarking (Supplementary Figure S20), SEAT illustrates higher or comparable clustering accuracy in the six datasets. SEAT predicts the optimal number of clusters closest to ground truth three, while Leiden and Louvain generally predict more clusters than SEAT. Further investigation shows that ground truth labels are mixed or not distinctly separated in two-dimensional data derived by UMAP for all datasets (Supplementary Figure S21), explaining the poor performance of 3Line-qPCR data. Likewise, hierarchical visualization plots depict nested layouts corresponding to the cell hierarchies in visualization refinement experiments (Supplementary Figure S22).

If we order the cells in cycling progress, cells from the same phase should be lined up adjacently as they share higher similarity. Thus, the cell order obtained from an ideal hierarchy could present a periodic pseudo-time order for cell cycle data. We visualize the cell order periodically with an oval plot, the placements of the cells in the oval represent their pseudo-time in the cell cycle (Figure 3A and Supplementary Figure S23). We access the cell ordering accuracy with the change index (CI) (17), which computes how frequently the gold standard cell cycle phase labels switch along the cell order. The benchmark methods are four conventional HC strategies (12) that offer a cell order. We also recruit state-of-the-art tools dedicating to predict the cell cycle pseudo-time, CYCLOPS (15), Cyclum (16), reCAT (17), and CCPE (18). SEAT demonstrates the highest ordering accuracy for all datasets, except for 3Line-qPCR_PC3, where SEAT wins the top two (Figure 3B). We exclude CCPE as it fails the tasks. In all, this suggests that cell hierarchy obtained from SEAT facilitates the cell cycle pseudo-time order inference.

SEAT orders cells in H1-hESC, mESC-Quartz, and mESC-SMARTer alongside the oval that closely matches the G0/G1-S-G2/M cycle (Figure 3A). Differential expression analysis among ground truth phases reveals distinct cell cycle phase markers (Supplementary Figure S24). These visible cell cycle marker patterns remain consistent when rearranging with SEAT cell order (Supplementary Figure S25). The top 20 differential expression genes ($P < 0.05$) for hESC and mESC cells include well-known cell cycle markers *UBE2C*, *TOP2A*, *CDK1*, and *CCNB1* (Supplementary Figure S26). Their expressions rise progressively with SEAT recovered pseudo-time order and are peaked with significant fold changes at the M phase (Figure 3C).

In H9, MB, and PC3 cell lines, the cell orders in the S and G2/M phases are partially arranged compared to the exact time course (Figure 3A). The differential expression makers of ground truth phases show that there are subpatterns within the S and G2/M phases. Moreover, there are similar patterns shared between the S and G2/M phases (Supplementary Figure S24), suggesting the cause of poor performance in pseudo-time ordering. Interestingly, after rearranging the marker expression heatmap with SEAT cell hierarchy, we observe distinct marker gene patterns among SEAT discovered cell subpopulations (Supplementary Figure S25). For the H9 cell line, SEAT detects four cell subpopulations (Figure 3D), G0/G1 phase corresponds to sub2. Cell cycle S and G2/M phases together have three cell subpopulations, sub0, sub1, and sub3. The top 20 differential expression genes ($P < 0.05$) exhibits two groups (Figure 3D). The genes from the first group are enriched in GO cell cycle signaling pathways. The genes from the second group are enriched in GO chemokine-mediated signaling and immune response pathways with CXC and IL gene families, respectively (Supplementary Figure S27). We demonstrate the top 20 differential expression genes for MB and PC3 cell lines in Supplementary Figures S26 and S27. Finally, we verify the cellular interactions among cell subpopulations with cell–cell communication analysis. We find a total of 124, 87, and 77 significant ($P < 0.05$) LR pair interactions among cell subpopulations for H9, MB, and PC3 cell

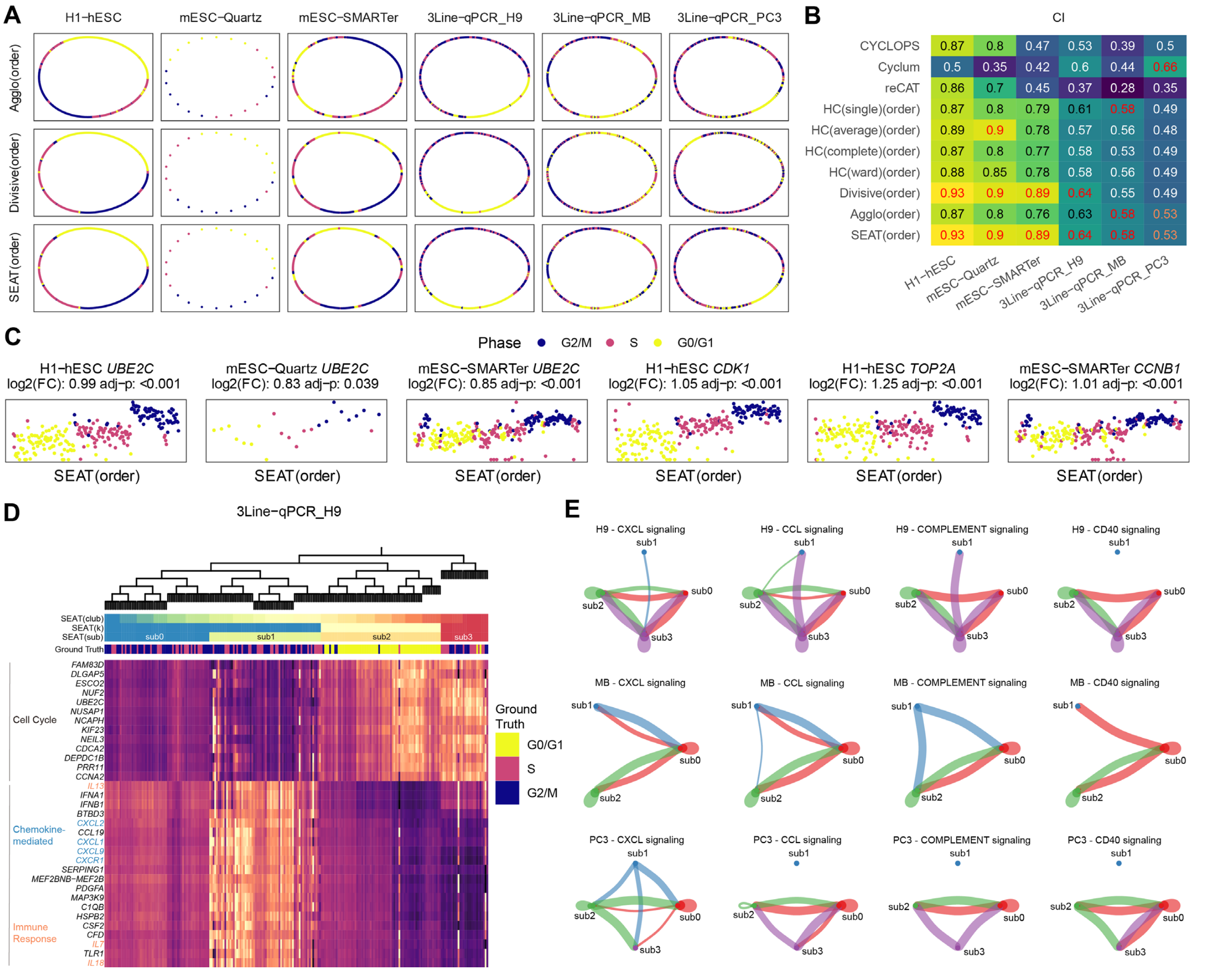


Figure 3. Applying SEAT on six scRNA cell cycle datasets. (**A**) The oval visualization of cell pseudo-time. From left to right are H1-hESC, mESC-Quartz, mESC-SMARTer, 3Line-qPCR_H9, 3Line-qPCR_MB, and 3Line-qPCR_PC3. From top to bottom are cell orders obtained from agglomerative hierarchy, divisive hierarchy, and SEAT cell hierarchy; namely, Agglo(order), Divisive(order), and SEAT(order). (**B**) The accuracy of cell pseudo-time order is measured by change index (CI) for baseline tools. The best score is colored red for each dataset. If SEAT gets second place, we color the score orange. HC(single)(order), HC(average)(order), HC(complete)(order), and HC(ward)(order): the cell orders from hierarchical clustering with single, average, complete, and ward linkage. (**C**) The normalized expression of M phase marker genes alongside the SEAT cell order. (**D**) The top 20 differentially expressed genes in G0/G1, S and G2/M ground truth phases for p3cl, arranged with SEAT cell hierarchy. SEAT(club): the cell clubs from SEAT cell hierarchy. SEAT(k): the cell subpopulations from SEAT cell hierarchy in predefined-k mode. SEAT(sub): the optimal subpopulations from SEAT cell hierarchy in auto-k mode. (**E**) The cell–cell communications among SEAT cell subpopulations for H9, MB, and PC3 cell lines.

lines, respectively. All datasets exhibit CXCL, CCL, COMPLEMENT, and CD40 signaling interactions among cell subpopulations (Figure 3E).

Cell hierarchy detects rare subclones on scDNA data

SEAT catalogs the clonal subpopulations of solid tumors and circulating tumor cells in four scDNA datasets. SEAT also identifies the CNV substructures of neuron and gamete cells in three scDNA datasets. Owning to the unique characteristics of CNV profiles, we only adopt SEAT agglomerative hierarchy to investigate the functional diversity of CNV substructures.

Navin *et al.* have profiled 100 cells from a genetically heterogeneous (polygenetic) triple-negative breast cancer primary lesion Navin_T10 (51). Fluorescence-activated cell sorting (FACS) analysis has confirmed that Navin_T10 carried four main cell subpopulations: diploid (D), hypodiploid (H), aneuploid A (A1), and aneuploid B (A2). Furthermore, Navin *et al.* have reported pseudo-diploid cells (P) with varying degrees of chromosome gains and losses from diploids. They are unrelated to the three tumor cell subgroups (H, A1, and A2) (51). Therefore, given whole-genome single-cell CNV profiles, we verify whether SEAT and the state-of-the-art clustering tools identify the four major cell groups and the rare pseudo-diploid cell group (Figure 4A). In predefined-k mode, SEAT agglomerative hierarchy successfully recognizes five cell subpopulations consistent with the patterns of CNV profiles. From top to bottom, the ranks are cancer normal cell group

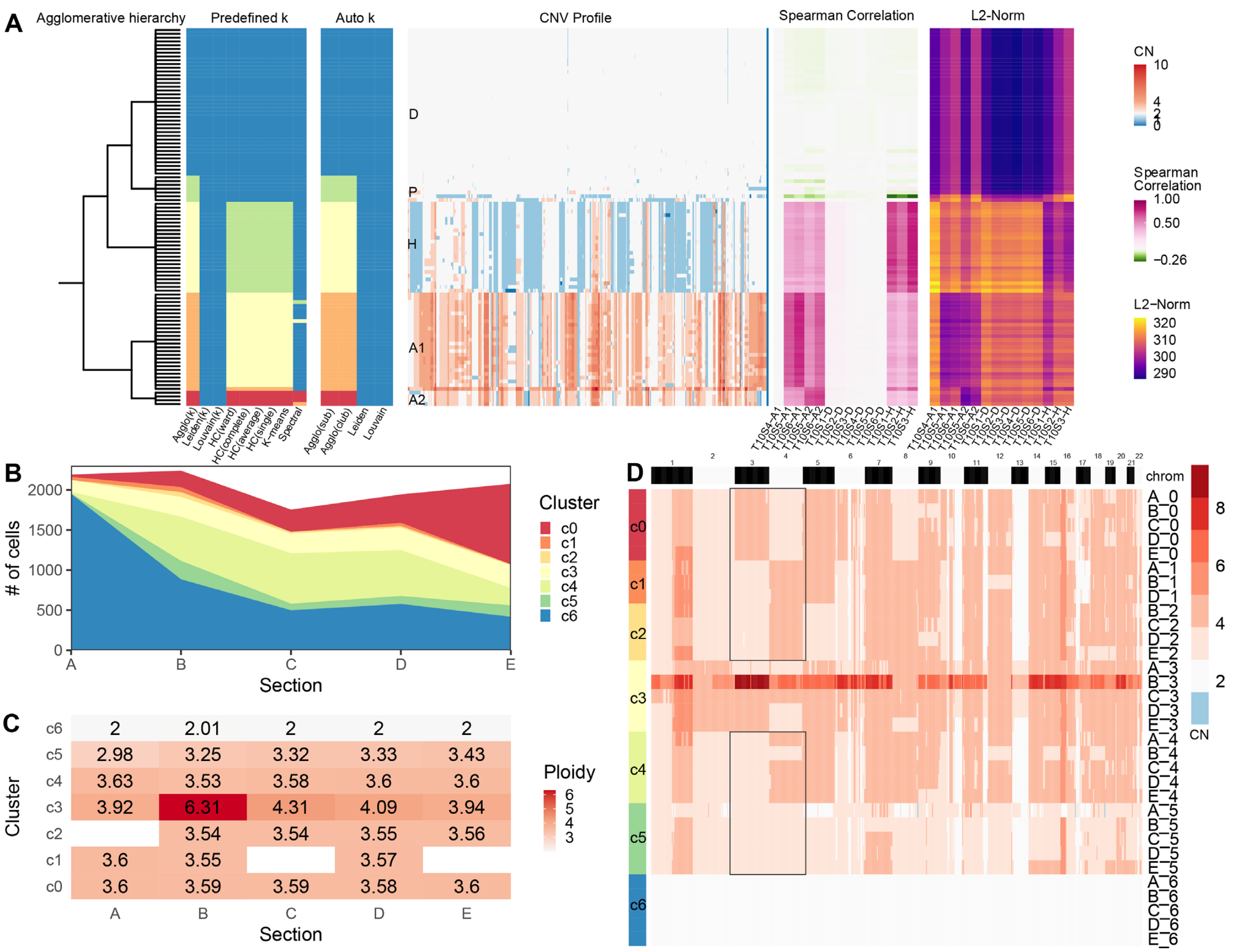


Figure 4. Applying SEAT on scDNA datasets. (**A**) The analysis result of Navin_T10. From left to right is the SEAT agglomerative hierarchy, subpopulation detecting results for predefined-k ($k = 5$) and auto-k clustering tools, the whole genome single-cell CNV heatmap of Navin_T10, the Spearman correlation, and Euclidean distance (L2-Norm) between scaled copy number profiled by scDNA and copy number density profiled by aCGH. Spectral: spectral clustering. HC(single), HC(average), HC(complete), and HC(ward): hierarchical clustering with single, average, complete, and ward linkage. Louvain(k) and Leiden(k): Louvain and Leiden in predefined-k mode. Agglo(k): the cell subpopulations from agglomerative hierarchy in predefined-k mode. Agglo(club): the cell clubs from the agglomerative hierarchy. Agglo(sub): the cell subpopulations from agglomerative hierarchy in auto-k mode. aCGH: array comparative genomic hybridization. (**B**) The stacked area plot illustrates the SEAT subpopulations across 10x_breast_S0 tumor sections. Cluster c6 (blue) signifies the diploid cells. (**C**) The mean ploidy of SEAT subpopulations across 10x_breast_S0 tumor sections. (**D**) The whole-genome single-cell CNV heatmap of SEAT subpopulations across 10x_breast_S0 tumor sections. The black boxes highlight the mutually exclusive amplification events on chr3 and ch4 across subclones. CNV: copy number variation.

(D), pseudo-diploid cell subgroup (P), hypodiploid cell subgroup (H), and two tumor aneuploid groups, A1 and A2 (Figure 4A). Leiden(k) and Louvain(k) fail at this task after 20 different resolution trials. Four HC strategies and K-means fail to distinguish the four pseudo-diploid cells as in the Navin *et al.*'s HC trial (51). Spectral clustering performs poorly by mixing tumor and normal cells. Regarding auto-k clustering algorithms, agglomerative hierarchy identifies five concordant subpopulations as predefined-k mode. Leiden and Louvain fail with the same sparse cell–cell similarity graph as input. Then, we leverage CNV density signals detected by aCGH from FACS identified D, H, A1, and A2 dissections of Navin_T10 (53) as silver standard to validate the clustering result. We calculate the pairwise Spearman correlation and Euclidean distance (L2-norm) between scaled single-cell CNV profiles and aCGH CNV signals. As a proof of concept, the single-cell CNV profiles of three bottom clusters separately own higher correlation and lower distance to aCGH profiles of H, A1, and A2 sections. The cells in the uppermost subpopulation detected by SEAT have almost zero correlation and the lowest distance with aCGH D section, suggesting that they are diploid cells. Pseudo-diploid cells illustrate a low correlation with all aCGH sections, validating their unique CNV profiles. Navin *et al.* have sequenced 100 cells from a monogenic triple-negative breast cancer tumor and its seeded liver metastasis (Navin_T16) (51). SEAT clusters the 100 samples into four distinct subpopulations (Supplementary Figure S28). Two are primary and metastasis aneuploid cells, corresponding to the published population structure. Notably, SEAT catalogs diploid cells and pseudo-diploid cells while baseline tools failed.

We collect a large-scale 10x scDNA-seq dataset (10x_breast_S0) without known subclone labels, where 10,202 cells from five adjacent tumor dissections (A, B, C, D, and E) of triple-negative breast cancer are sequenced. We check whether SEAT seizes the substantial intra-tumor heterogeneity. In Figure 4B–D, SEAT automatically detects seven subpopulations, and the proportions of the cell subpopulations vary across the five lesions. The blue subpopulation c6 gathers normal cells, with the mean cellular ploidy being diploid across all sections. The number of normal cells gradually decreases from sections A to E. SEAT identifies six clonal subpopulations (c0–c5), where c3 manifests the highest average ploidy. The mutually exclusive amplification events (marked with black boxes in Figure 4D) on chr3 and chr4 of subclones c0, c1, c2, and c4, indicate an early branching evolution which is consistent with the findings of Wang *et al.* (68); that is, originated from normal cell group c6, the earliest subclone could be c5, with CN=3 on ch3 and ch4. Subclone c5 derived to subclone c0 with amplification on chr3 (CN=4). Moreover, subclone c5 derived to an intermediate subclone with amplification on chr4 (CN=4). Then, the intermediate subclone derived to subclone c1, c2, and c4 with CN gains on other chromosomes.

Furthermore, SEAT distinguishes cells with CNV gains and losses in circulating tumor cells of seven lung cancer patients (55) and in human cortical neurons (56) (Supplementary Figure S28). SEAT also detects the loss of heterogeneity event, it successfully classifies chrX-bearing, chrY-bearing, and aneuploid sperm cells (57,58) (Supplementary Figure S28).

Cell hierarchy dissects chromatin accessibility heterogeneity of single-cell data

SEAT dissects chromatin accessibility heterogeneity of single cells. We utilize three public scATAC-seq data as benchmarking sets with gold standard cell type labels. scatac_6cl is a mixture of six cell lines (BJ, GM12878, H1-ESC, HL60, K562, and TF1) (59). Hematopoiesis consists of eight types of human hematopoiesis cells (CLP, CMP, GMP, HSC, LMPP, MEP, MPP, and pDC) (60). T-cell composes of four T-cell subtypes (Jurkat_T_cell, Naive_T_cell, Memory_T_cell, and Th17_T_cell) (61). We collect a multiome of scRNA and scATAC dataset, PBMC, for peripheral blood mononuclear cells (PBMCs) with 14 cell types.

The orders of the cells in the agglomerative and divisive hierarchies are consistent with their ground truth cell types (Supplementary Figure S29). The clustering accuracies of SEAT against its baselines are in Figure 5A. In predefined-k mode, SEAT(k) demonstrates the highest clustering accuracies on scatac_6cl and T-cell sets. For auto-k clustering, SEAT(sub) beats Louvain and Leiden on all four sets. For scatac_6cl and T-cell, the optimal number of clusters obtained by SEAT matches the ground truth, thus yielding the comparable ARI against predefined-k clustering algorithms. Leiden and Louvain have lower performance due to predicting more clusters than ground truth (Supplementary Figure S29).

We check whether SEAT reveals the functional diversity of single-cell chromatin accessibility. We select cells from scatac_6cl GM12878 cell line, then conduct *cis*-regulatory DNA interaction analysis on chr22 for SEAT cell club1 and club2. Figures 5B and 5C depict the *cis*-regulatory map on chr22 of club1 and club2 cells, respectively. The co-accessibility correlations among peaks of club2 cells are significantly higher ($P < 0.05$) than club1 cells (Figure 5D). Meanwhile, we identify 29 and 179 *cis*-co-accessibility networks (CCANs) from GM12878-club1 and GM12878-club2, respectively (Figure 5E). The CCANs detected in GM12878-club1 and GM12878-club2 are heterogeneous. Figure 5F illustrates a GM128780-club1 specified CCAN at chr22:20,827,398–21,441,482. The *cis*-regulatory elements surrounding gene *SNAP29* are co-accessible only in GM128780-club1. Moreover, we found dense pairwise connections among peaks at chr22:39,778,355–40,451,820 in GM12878-club2 (Figure 5G), harboring genes *TAB1*, *MGAT3*, *MIEF1*, *CACNA1I*, *ENTHD1*, *GRAP2*, *FAM83F*, *TNRC6B*, etc.

Similar to the scRNA visualization refinement experiments, the SEAT(viz) reveals a clear pattern of cells corresponding to ground truth; and the nested layouts of subpopulations and clubs are clearly illustrated with gray and black circles (Figures 5H, 5I, and Supplementary Figure S30). However, UMAP visualizations derived from high-dimensional data mix ground truth cell subpopulations in one clump (Supplementary Figure S29). Furthermore, UMAP, TSNE, and PHATE visualizations derived from cell–cell similarity graphs fail to place cells from K562 (light green) and TF1 (yellow) within the vicinity in scatac_6cl; and they fail to place all effector CD8 T cells (magenta) together in PBMC (Figures 5H and 5I). Likewise, the cell clubs marked with red circles are unclearly segregated in UMAP, TNSE, and PHATE plots.

DISCUSSION

Detecting and visualizing cellular functional diversity are essential in single-cell analysis. Neglection of the underlying cellular nested structures prevents the capture of full-scale cellular functional diversity. To address the challenge, we incorporate cell hierarchy to investigate the functional diversity of cellular systems at the subpopulation, club, and cell layers, hierarchically. The cell subpopulations and clubs catalog the functional diversity of cells in broad and fine resolution, respectively. In the cell layer, the order of cells further records the slight dynamics among cells locally. Accordingly, we establish SEAT to construct cell hierarchies utilizing structure entropy by diminishing the global uncertainty of cell–cell graphs. In addition, SEAT offers an interface to embed cells into low-dimensional space while preserving the global-subpopulation-club hierarchical layout in cell hierarchy.

Currently, state-of-the-art clustering tools for cell subpopulation or club investigation neglect the underlying nested structures of cells. Flatten clustering tools, such as spectral clustering (10) and K-means (11), do not support the cell hierarchy. Although conventional hierarchical clustering (12), Louvain (13), and Leiden (14) derive cell hierarchy layer by layer via optimizing merging or splitting metrics, computing these metrics merely uses single-layer information. When constructing subsequent layers, they have not

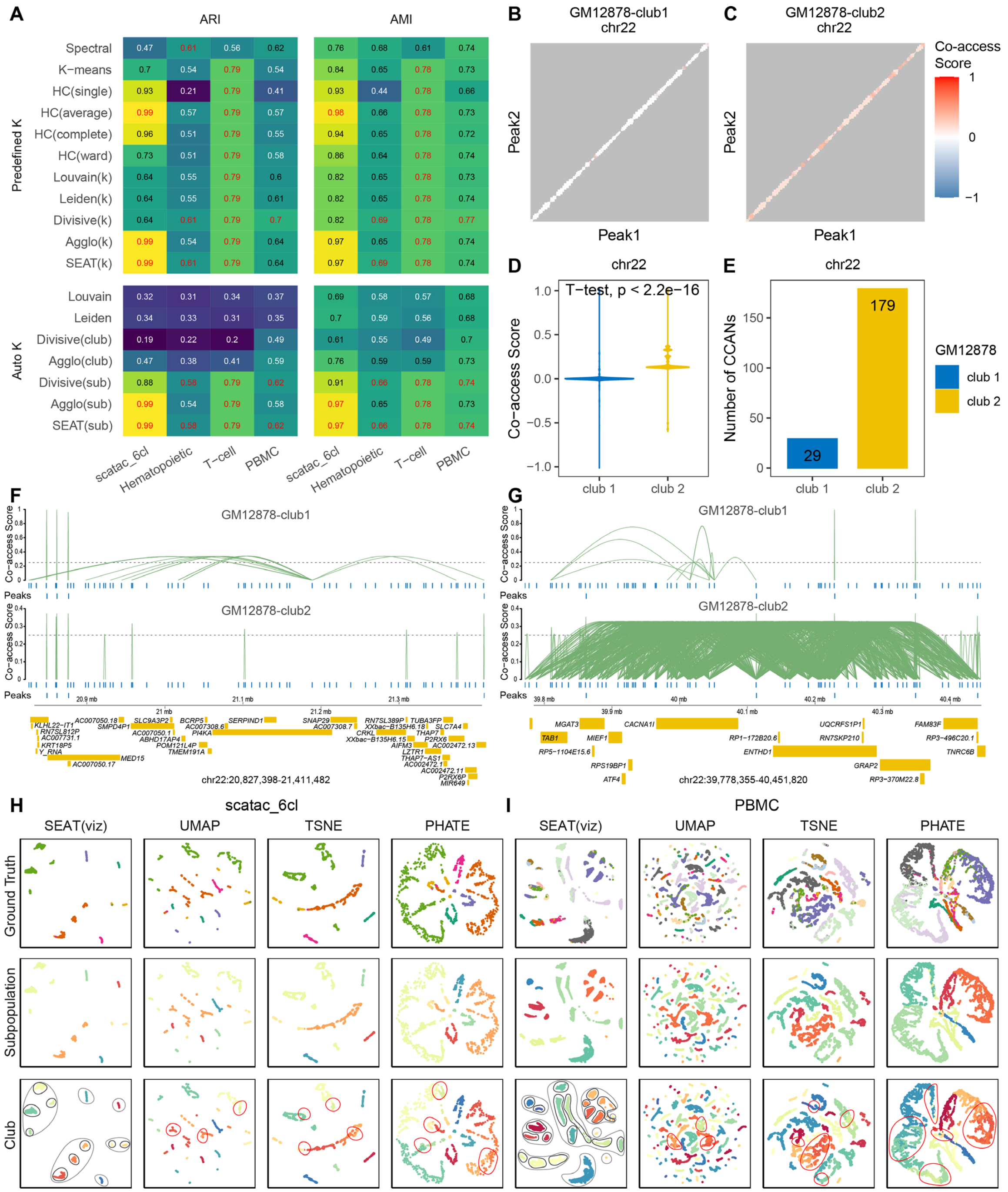


Figure 5. Applying SEAT on three scATAC datasets and one scRNA-scATAC multiome dataset. (**A**) The adjusted rand index (ARI) and adjusted mutual information (AMI) of predefined-k and auto-k clustering tools. The best score is colored red for each dataset in predefined and auto clustering benchmarking. Spectral: spectral clustering. HC(single), HC(average), HC(complete), and HC(ward): hierarchical clustering with single, average, complete, and ward linkage. Louvain(k) and Leiden(k): Louvain and Leiden in predefined-k mode. Divisive(k) and Agglo(k): the cell subpopulations from divisive and agglomerative hierarchy in predefined-k mode. SEAT(k): the cell subpopulations from SEAT cell hierarchy in predefined-k mode. Divisive(club) and Agglo(club): the cell clubs from the divisive and agglomerative hierarchy. Divisive(sub) and Agglo(sub): the cell subpopulations from divisive and agglomerative hierarchy in auto-k mode. SEAT(sub): the optimal subpopulations from SEAT cell hierarchy in auto-k mode. (**B–D**) The co-accessibility score among peak pairs at chr22 for cells at SEAT club1 and club2 from scatac_6cl GM12878 cell line. (**E**) The number of *cis*-co-accessibility networks (CCANs) among pair of peaks at chr22 for cells at SEAT club1 and club2 from scatac_6cl GM12878 cell line. (**F**) The co-accessibility connections among *cis*-regulatory elements in chr22:20,827,398–21,441,482. The height of links signifies the degree of the co-accessibility correlation between the pair of peaks. The top panel illustrates cells in scatac_6cl GM12878-club1, and the bottom shows cells in scatac_6cl GM12878-club2. (**G**) The co-accessibility connections among *cis*-regulatory elements in chr22:39,778,355–40,451,820. The height of links signifies the degree of the co-accessibility correlation between the pair of peaks. The top panel illustrates cells in scatac_6cl GM12878-club1, and the bottom shows cells in scatac_6cl GM12878-club2. (**H**, **I**) SEAT(viz), UMAP, TSNE, and PHATE plots of scatac_6cl and PBMC. The cells are colored with subpopulations, clubs, and ground truth. The gray and black circles in the SEAT(viz) plot indicate the subpopulation and club boundaries, respectively. In UMAP, TSNE, and PHATE plots, the red circles mark the unclearly segregated cell clubs. SEAT(viz): the hierarchical visualization from SEAT cell hierarchy.

incorporated the built-in cell hierarchy in the previous layers. Structure entropy is a metric that encompasses the previously constructed internal cell hierarchy. Experiments validate that SEAT delivers robust cell-type clustering results and forms insightful hierarchical structures of cells.

SEAT is good at finding the optimal subpopulation number with high accuracy. We have collected scRNA, scDNA, scATAC, and scRNA-scATAC profiles with the number of cell types ranging from 2 to 14. SEAT consistently predicts the optimal cluster number closest to the gold or silver standards, while Louvain and Leiden predict too many clusters. Especially for scRNA set Kumar, SEAT boosts the accuracy from 0.34 to 1 compared to Louvain and Leiden (Figure 2A). Auto-k clustering mode of SEAT is comparable to or better than the best clustering results of predefined-k clustering methods for most datasets.

SEAT specializes in hierarchically deciphering cellular functional diversity at subpopulation and club levels. We observe visible marker gene patterns that match cell clubs within one cell subpopulation. For the p3cl set, the basal, luminal, and fibroblast cell subpopulations have their own cell clubs, determined by differentially expressed cell cycle genes (*HIST1H4C*, *CDC20*, *CCNB1*, and *PTTG1*) (Figure 2C). Looking at the seven agglomerative clubs for the basal subpopulation, we find a distinct breast cancer cell club that drives oncogenic *AREG-EGFR* signaling in all basal cells (Figure 2D), suggesting a promoting role in tumorigenesis (67). Cell hierarchy obtained from copy number profiles of 10x_breast_S0 demonstrates a mutually exclusive subclones layout (Figure 4D), indicating an early branch evolution (68). Furthermore, we find that there is a club-specified dense co-accessible network of *cis*-regulatory elements at chr22:39,778,355–40,451,820 in GM12878-club2, harboring genes *TAB1*, *MGAT3*, *MIEF1*, *CACNA1I*, *ENTHD1*, *GRAP2*, *FAM83F*, *TNRC6B*, etc. (Figure 5G).

Inferring the periodic pseudo-time for the cell cycle data is crucial as it reveals the functional diversity of cells undergoing the cell cycle process. Several tools are dedicated to cell cycle pseudo-time inference. CYCLOPS (15) and Cyclum (16) utilize deep autoencoders to project expression profiles into cell pseudo-time in the periodic process, which act as black boxes and lack explainability. reCAT (17) employs the Gaussian mixture model to group cells into clusters, and constructs a cluster-cluster graph weighted by the Euclidean distance between the mean expression profile of each cluster, then takes the traveling salesman path of the cluster-cluster graph as the order. Finding a traveling salesman path is NP-hard, and no polynomial time algorithms are available (17). CCPE (18) learns a discriminative helix to represent the periodic process and infer the pseudo-time. However, we fail to run CCPE according to its GitHub instruction. Moreover, CYCLOPS, Cyclum, reCAT, and CCPE bypass the nested structure of cells when inferring the pseudo-time. In this study, we propose that the cell layer of a hierarchy encodes the pseudo-time of cells for cycling data. We build the hierarchy by minimizing the structure entropy of the kNN cell–cell graph. The built hierarchy carries the nested structure between individual cells and their ancestral cell partitions. Then, the order of individual cells is acquired with an in-order traversal of the hierarchy. scRNA data exemplify that SEAT cell orders outperform CYCLOPS, Cyclum, reCAT, and CCPE by accurately predicting the periodic pseudo-time of cells in the cell cycle process. In hESC and mESC cells, the expressions of M phase marker genes *UBE2C*, *TOP2A*, *CDK1,* and *CCNB1* rise progressively alongside the SEAT recovered order and are peaked at the M phase with significant fold changes (Figure 3C).

Visualizing the hierarchical functional diversity of cells in biological systems is crucial for obtaining insightful biological hypotheses. UMAP (26) intends to maintain the global cell structures by minimizing the binary cross entropy. TSNE (27) preserves the local cell structures. PHATE (28) tackles the general shape and local transition of cells. However, none of them impart the nested structures of cells into the visualization. We propose a nonlinear dimension reduction refinement based on UMAP by incorporating cell hierarchy as supervised knowledge. We acquire three cell–cell graphs that only store the intra-connections of cells within each global, subpopulation, and club partition. Then, we minimize the weighted binary cross-entropy of the three cell–cell graphs. This approach guarantees the global structure of the cells. Moreover, it ensures that cells within one cell club and cell clubs within one subpopulation are closely placed in the visualization. In contrast, cells from different clubs and subpopulations are kept at a considerable distance. One can adjust the cross-entropy weights of global-subpopulation-cell layers so that the patterns in visualization retain a desired degree of hierarchy. Experiments with scRNA and scATAC data demonstrate that SEAT hierarchical visualization consistently produces a clear layout of cell clumps corresponding to the cell hierarchy.

Cellular abnormalities may distort the entire cell hierarchy. When there are cell outliers presented, the original SEAT will assign each cell outlier to its nearest cell subpopulation. Thus, the downstream biological interpretation may be skewed. To tackle the issue, we provide an optional average kNN outlier detection step before constructing the cell hierarchy. In Supplementary Results and Supplementary Figures S31–S35, we demonstrate the distance cutoff is more stable than the distance percentile cutoff because the latter heavily depends on the ratio of outliers in the whole population. Thus, we set distance cutoff as the default outlier detection strategy.

The structure entropy evaluates the global uncertainty of random walks through a network with a nested structure (19). The minimum structure entropy interprets a stable nested structure in the network. Li *et al.* has used structure entropy to define tumor subtypes from bulk gene expression data (21) and to detect the hierarchical topologically associating domains from Hi-C data (22). These works utilize greedy merging and combining operations to build a local optimal multi-nary hierarchy and cutting hierarchy roughly by keeping the top layers. As we have proven that a binary hierarchy of minimum structure entropy exists for a graph (23), Li *et al.*'s strategy to search for a multi-nary hierarchy is not optimized. Adopted by Louvain and Leiden, modularity is a popular optimization metric to capture community structure in a single-cell network. Agglo(club) is analogous to Louvain's if we switch the merging metric to modularity. Agglo(club) achieves better or comparable clustering performance against Louvain in most benchmark sets (Figures 2A and 5A), suggesting the superiority of structure entropy over modularity in measuring the strength of hierarchically partitioning a network into subgroups. We have discussed the differences and advantages of SEAT against the existing structure entropy and modularity approaches at the algorithmic level in the Supplementary Method.

SEAT detects the cell hierarchy, assuming that the structure entropy codes nested structures of cells. There is no assurance that the resultant cell hierarchy will resemble accurate nested structures of cells. SEAT finds a pseudo cell hierarchy of cells. Nevertheless, the pseudo cell hierarchy showcases profound efficacy and biological insights in subpopulation detection, cell club investigation, and periodic pseudo-time inference for single-cell multiomics benchmarking datasets. In future work, we aim to refine the algorithm to find a more accurate and insightful pseudo cell hierarchy.

Recall that the cell hierarchy has multiple layers to present cellular heterogeneity. In this study, we merely utilize four main layers (global, subpopulation, club, and cell) to interpret and visualize the cellular functional diversity. In the future, we intend to investigate possible biological insights and visualization layouts derived from more cell hierarchy layers.

Moreover, the order of the cell clubs can be flipped in the cell hierarchy. There is only a partial order among cells bounded by the cell hierarchy. We plan to refine the algorithm to provide a proper non-partial one-dimensional order, which might infer the nuance of pseudo-time or development trajectory among cells outside the periodic cell cycle.

DATA AVAILABILITY

The 25 scRNA, seven scDNA, three scATAC and one scRNA-scATAC multiome datasets are publicly available. The details are summarized in Experiment Setting and Supplementary Methods.

CODE AVAILABILITY

The source code of SEAT is available at https://github.com/deepomicslab/SEAT.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Author contributions: L.X.C. conducted the project and wrote the manuscript. S.C.L. supervised the project and revised the manuscript.

FUNDING

CityU/UGC Research Matching Grant Scheme [9229012]. Funding for open access charge: The Science Technology and Innovation Committee of Shenzhen Municipality (Project No. JCYJ20200109143216036).

Conflict of interest statement. None declared.

REFERENCES

1. Andrews,T.S., Kiselev,V.Y., McCarthy,D. and Hemberg,M. (2021) Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data. *Nat. Protoc.*, **16**, 1–9.
2. Nayak,R. and Hasija,Y. (2021) A hitchhiker's guide to single-cell transcriptomics and data analysis pipelines. *Genomics*, **133**, 606–619.
3. Wu,Z. and Wu,H. (2020) Accounting for cell type hierarchy in evaluating single cell RNA-seq clustering. *Genome Biol.*, **21**, 123.
4. Gao,Y., Chen,L., Cai,G., Xiong,X., Wu,Y., Ma,D., Li,S.C. and Gao,Q. (2020) Heterogeneity of immune microenvironment in ovarian cancer and its clinical significance: a retrospective study. *Oncoimmunology*, **9**, 1760067.
5. Minussi,D.C., Nicholson,M.D., Ye,H., Davis,A., Wang,K., Baker,T., Tarabichi,M., Sei,E., Du,H., Rabbani,M. *et al.* (2021) Breast tumours maintain a reservoir of subclonal diversity during expansion. *Nature*, **592**, 302–308.
6. Chen,L., Qing,Y., Li,R., Li,C., Li,H., Feng,X. and Li,S.C. (2022) Somatic variant analysis suite: copy number variation clonal visualization online platform for large-scale single-cell genomics. *Brief. Bioinform.*, **23**, bbab452.
7. Kowalczyk,M.S., Tirosh,I., Heckl,D., Rao,T.N., Dixit,A., Haas,B.J., Schneider,R.K., Wagers,A.J., Ebert,B.L. and Regev,A. (2015) Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.*, **25**, 1860–1872.
8. Zheng,G.X., Terry,J.M., Belgrader,P., Ryvkin,P., Bent,Z.W., Wilson,R., Ziraldo,S.B., Wheeler,T.D., McDermott,G.P., Zhu,J. *et al.* (2017) Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.*, **8**, 14049.
9. Cusanovich,D.A., Daza,R., Adey,A., Pliner,H.A., Christiansen,L., Gunderson,K.L., Steemers,F.J., Trapnell,C. and Shendure,J. (2015) Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, **348**, 910–914.
10. Ng,A.Y., Jordan,M.I. and Weiss,Y. (2002) On spectral clustering: Analysis and an algorithm. In: *Advances in Neural Information Processing Systems*. pp. 849–856.
11. Hartigan,J.A. and Wong,M.A. (1979) Algorithm AS 136: a k-means clustering algorithm. *J. Roy. Stat. Soc. Series C (Appl. Stat.)*, **28**, 100–108.
12. Johnson,S.C. (1967) Hierarchical clustering schemes. *Psychometrika*, **32**, 241–254.
13. Blondel,V.D., Guillaume,J.-L., Lambiotte,R. and Lefebvre,E. (2008) Fast unfolding of communities in large networks. *J. Stat. Mech.: Theor. Exp.*, **2008**, P10008.
14. Traag,V.A., Waltman,L. and Van Eck,N.J. (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.*, **9**, 5233.
15. Anafi,R.C., Francey,L.J., Hogenesch,J.B. and Kim,J. (2017) CYCLOPS reveals human transcriptional rhythms in health and disease. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 5312–5317.
16. Liang,S., Wang,F., Han,J. and Chen,K. (2020) Latent periodic process inference from single-cell RNA-seq data. *Nat. Commun.*, **11**, 1441.
17. Liu,Z., Lou,H., Xie,K., Wang,H., Chen,N., Aparicio,O.M., Zhang,M.Q., Jiang,R. and Chen,T. (2017) Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. *Nat. Commun.*, **8**, 22.
18. Liu,J., Yang,M., Zhao,W. and Zhou,X. (2022) CCPE: cell cycle pseudotime estimation for single cell RNA-seq data. *Nucleic Acids Res.*, **50**, 704–716.
19. Li,A. and Pan,Y. (2016) Structural information and dynamical complexity of networks. *IEEE Trans. Inform. Theor.*, **62**, 3290–3339.
20. Li,A., Li,J. and Pan,Y. (2015) Discovering natural communities in networks. *Physica A: Stat. Mech. Appl.*, **436**, 878–896.
21. Li,A., Yin,X. and Pan,Y. (2016) Three-dimensional gene map of cancer cell types: Structural entropy minimisation principle for defining tumour subtypes. *Sci. Rep.*, **6**, 20412.
22. Li,A., Yin,X., Xu,B., Wang,D., Han,J., Wei,Y., Deng,Y., Xiong,Y. and Zhang,Z. (2018) Decoding topologically associating domains with ultra-low resolution Hi-C data by graph structural entropy. *Nat. Commun.*, **9**, 3265.
23. Zhang,Y.W., Wang,M.B. and Li,S.C. (2021) SuperTAD: robust detection of hierarchical topologically associated domains with optimized structural information. *Genome Biol.*, **22**, 45.
24. Zhang,Y.W., Chen,L. and Li,S.C. (2022) Detecting TAD-like domains from RNA-associated interactions. *Nucleic Acids Res.*, **50**, e88.

25. Chen,L., Xu,J. and Li,S.C. (2019) DeepMF: Deciphering the latent patterns in omics profiles with a deep learning method. *BMC Bioinformatics*, **20**, 648.
26. McInnes,L., Healy,J. and Melville,J. (2018) Umap: Uniform manifold approximation and projection for dimension reduction. arXiv doi: https://arxiv.org/abs/1802.03426, 18 September 2020, preprint: not peer reviewed.
27. Van der Maaten,L. and Hinton,G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.
28. Moon,K.R., van Dijk,D., Wang,Z., Gigante,S., Burkhardt,D.B., Chen,W.S., Yim,K., Elzen,A. v.d., Hirn,M.J., Coifman,R.R. *et al.* (2019) Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.*, **37**, 1482–1492.
29. YU,Y., Chen,L., Miao,X. and Li,S.C. (2021) SpecHap: a diploid phasing algorithm based on spectral graph theory. *Nucleic Acids Res.*, **49**, e144.
30. Von Luxburg,U. (2007) A tutorial on spectral clustering. *Stat. Comput.*, **17**, 395–416.
31. Dong,M., Thennavan,A., Urrutia,E., Li,Y., Perou,C.M., Zou,F. and Jiang,Y. (2021) SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief. Bioinform.*, **22**, 416–427.
32. McDavid,A., Dennis,L., Danaher,P., Finak,G., Krouse,M., Wang,A., Webster,P., Beechem,J. and Gottardo,R. (2014) Modeling bi-modality improves characterization of cell cycle on gene expression in single cells. *PLoS Comput. Biol.*, **10**, e1003696.
33. Tian,L., Dong,X., Freytag,S., Lê Cao,K.-A., Su,S., JalalAbadi,A., Amann-Zalcenstein,D., Weber,T.S., Seidi,A., Jabbari,J.S. *et al.* (2019) Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nat. Methods*, **16**, 479–487.
34. Yan,L., Yang,M., Guo,H., Yang,L., Wu,J., Li,R., Liu,P., Lian,Y., Zheng,X., Yan,J. *et al.* (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.*, **20**, 1131–1139.
35. Deng,Q., Ramsköld,D., Reinius,B. and Sandberg,R. (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, **343**, 193–196.
36. Biase,F.H., Cao,X. and Zhong,S. (2014) Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Res.*, **24**, 1787–1796.
37. Goolam,M., Scialdone,A., Graham,S.J., Macaulay,I.C., Jedrusik,A., Hupalowska,A., Voet,T., Marioni,J.C. and Zernicka-Goetz,M. (2016) Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. *Cell*, **165**, 61–74.
38. Koh,P.W., Sinha,R., Barkal,A.A., Morganti,R.M., Chen,A., Weissman,I.L., Ang,L.T., Kundaje,A. and Loh,K.M. (2016) An atlas of transcriptional, chromatin accessibility, and surface marker changes in human mesoderm development. *Scientific Data*, **3**, 160109.
39. Kumar,R.M., Cahan,P., Shalek,A.K., Satija,R., DaleyKeyser,A.J., Li,H., Zhang,J., Pardee,K., Gennert,D., Trombetta,J.J. *et al.* (2014) Deconstructing transcriptional heterogeneity in pluripotent stem cells. *Nature*, **516**, 56–61.
40. Trapnell,C., Cacchiarelli,D., Grimsby,J., Pokharel,P., Li,S., Morse,M., Lennon,N.J., Livak,K.J., Mikkelsen,T.S. and Rinn,J.L. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**, 381–386.
41. Blakeley,P., Fogarty,N.M., Del Valle,I., Wamaitha,S.E., Hu,T.X., Elder,K., Snell,P., Christie,L., Robson,P. and Niakan,K.K. (2015) Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development*, **142**, 3151–3165.
42. Kolodziejczyk,A.A., Kim,J.K., Tsang,J.C., Ilicic,T., Henriksson,J., Natarajan,K.N., Tuck,A.C., Gao,X., Bühler,M., Liu,P. *et al.* (2015) Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell*, **17**, 471–485.
43. Xin,Y., Kim,J., Okamoto,H., Ni,M., Wei,Y., Adler,C., Murphy,A.J., Yancopoulos,G.D., Lin,C. and Gromada,J. (2016) RNA sequencing of single human islet cells reveals type 2 diabetes genes. *Cell Metab.*, **24**, 608–615.
44. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck III,W.M., Zheng,S., Butler,A., Lee,M.J., Wilk,A.J., Darby,C., Zager,M. *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184** , 3573–3587.
45. Jin,S., Guerrero-Juarez,C.F., Zhang,L., Chang,I., Ramos,R., Kuan,C.-H., Myung,P., Plikus,M.V. and Nie,Q. (2021) Inference and analysis of cell–cell communication using CellChat. *Nat. Commun.*, **12**, 1088.
46. Leng,N., Chu,L.-F., Barry,C., Li,Y., Choi,J., Li,X., Jiang,P., Stewart,R.M., Thomson,J.A. and Kendziorski,C. (2015) Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments. *Nat. Methods*, **12**, 947.
47. Sasagawa,Y., Nikaido,I., Hayashi,T., Danno,H., Uno,K.D., Imai,T. and Ueda,H.R. (2013) Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol.*, **14**, R31.
48. Buettner,F., Natarajan,K.N., Casale,F.P., Proserpio,V., Scialdone,A., Theis,F.J., Teichmann,S.A., Marioni,J.C. and Stegle,O. (2015) Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, **33**, 155–160.
49. Wang,B., Pu,J., Chen,L. and Li,S. (2022) SMURF: embedding single-cell RNA-seq data with matrix factorization preserving self-consistency. bioRxiv doi: https://doi.org/10.1101/2022.04.22.489140, 22 April 2022, preprint: not peer reviewed.
50. Ge,S.X., Jung,D. and Yao,R. (2020) ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics*, **36**, 2628–2629.
51. Navin,N., Kendall,J., Troge,J., Andrews,P., Rodgers,L., McIndoo,J., Cook,K., Stepansky,A., Levy,D., Esposito,D. *et al.* (2011) Tumour evolution inferred by single-cell sequencing. *Nature*, **472**, 90.
52. Garvin,T., Aboukhalil,R., Kendall,J., Baslan,T., Atwal,G.S., Hicks,J., Wigler,M. and Schatz,M.C. (2015) Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods*, **12**, 1058.
53. Navin,N., Krasnitz,A., Rodgers,L., Cook,K., Meth,J., Kendall,J., Riggs,M., Eberling,Y., Troge,J., Grubor,V. *et al.* (2010) Inferring tumor progression from genomic heterogeneity. *Genome Res.*, **20**, 68–80.
54. Zaccaria,S. and Raphael,B.J. (2021) Characterizing allele-and haplotype-specific copy numbers in single cells with CHISEL. *Nat. Biotechnol.*, **39**, 207–214.
55. Ni,X., Zhuo,M., Su,Z., Duan,J., Gao,Y., Wang,Z., Zong,C., Bai,H., Chapman,A.R., Zhao,J. *et al.* (2013) Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 21083–21088.
56. McConnell,M.J., Lindberg,M.R., Brennand,K.J., Piper,J.C., Voet,T., Cowing-Zitron,C., Shumilina,S., Lasken,R.S., Vermeesch,J.R., Hall,I.M. *et al.* (2013) Mosaic copy number variation in human neurons. *Science*, **342**, 632–637.
57. Lu,S., Zong,C., Fan,W., Yang,M., Li,J., Chapman,A.R., Zhu,P., Hu,X., Xu,L., Yan,L. *et al.* (2012) Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science*, **338**, 1627–1630.
58. Wang,J., Fan,H.C., Behr,B. and Quake,S.R. (2012) Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell*, **150**, 402–412.
59. Buenrostro,J.D., Wu,B., Litzenburger,U.M., Ruff,D., Gonzales,M.L., Snyder,M.P., Chang,H.Y. and Greenleaf,W.J. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**, 486–490.
60. Buenrostro,J.D., Corces,M.R., Lareau,C.A., Wu,B., Schep,A.N., Aryee,M.J., Majeti,R., Chang,H.Y. and Greenleaf,W.J. (2018) Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell*, **173**, 1535–1548.
61. Satpathy,A.T., Saligrama,N., Buenrostro,J.D., Wei,Y., Wu,B., Rubin,A.J., Granja,J.M., Lareau,C.A., Li,R., Qi,Y. *et al.* (2018) Transcript-indexed ATAC-seq for precision immune profiling. *Nat. Med.*, **24**, 580–590.
62. Li,Z., Kuppe,C., Ziegler,S., Cheng,M., Kabgani,N., Menzel,S., Zenke,M., Kramann,R. and Costa,I.G. (2021) Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen. *Nat. Commun.*, **12**, 6386.
63. Pliner,H.A., Packer,J.S., McFaline-Figueroa,J.L., Cusanovich,D.A., Daza,R.M., Aghamirzaie,D., Srivatsan,S., Qiu,X., Jackson,D., Minkina,A. *et al.* (2018) Cicero predicts cis-regulatory DNA

interactions from single-cell chromatin accessibility data. *Mol. Cell*, **71**, 858–871.

64. Rand,W.M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, **66**, 846–850.
65. Cover,T.M. and Thomas,J.A. (2012) In: *Elements of Information Theory*. John Wiley & Sons.
66. Zappia,L., Phipson,B. and Oshlack,A. (2017) Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.*, **18**, 174.
67. Kappler,C.S., Guest,S.T., Irish,J.C., Garrett-Mayer,E., Kratche,Z., Wilson,R.C. and Ethier,S.P. (2015) Oncogenic signaling in amphiregulin and EGFR-expressing PTEN-null human breast cancer. *Mol. Oncol.*, **9**, 527–543.
68. Wang,R., Lin,D.-Y. and Jiang,Y. (2020) SCOPE: a normalization and copy-number estimation method for single-cell DNA sequencing. *Cell Syst.*, **10**, 445–452.