MarkovHC: Markov hierarchical clustering for the topological structure of high-dimensional single-cell omics data

Zhenyi Wang, [#,1] Yanjie Zhong,[3,5] Zhaofeng Ye,[2] Lang Zeng,[6] Yang Chen,[1] Minglei Shi,[2] MinPing Qian*[,3] and Michael Q. Zhang*[,1,2,4]

1 MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, BNRist; Department of Automation, Tsinghua University, Beijing 100084, China

2 MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, BNRist; School of Medicine, Tsinghua University, Beijing 100084, China

3, School of Mathematical Sciences, Peking University, Beijing 100871, China

4, Department of Biological Sciences, Center for Systems Biology, The University of Texas, Richardson, TX 75080-3021, USA

5, Department of Mathematics and Statistics, Washington University in St. Louis, St. Louis, MO 63130, USA

6, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

Emails:
Zhenyi Wang[#]: wangzy17@mails.tsinghua.edu.cn
Yanjie Zhong: yanjie.zhong@wustl.edu
Zhaofeng Ye: yezf16@mails.tsinghua.edu.cn
Lang Zeng: langzeng@umich.edu
Yang Chen: yc@mail.tsinghua.edu.cn
Minglei Shi: shiml79@mail.tsinghua.edu.cn
MinPing Qian[*]: qianmp@math.pku.edu.cn
Michael Q. Zhang[*]: michaelzhang@tsinghua.edu.cn

Highlights:

1. MarkovHC explores the topology hierarchy in high-dimensional data.
2. MarkovHC can find clusters (basins) and cores (attractors) of clusters in different scales.
3. The trajectory of cell transition (transition paths) and critical points in the process of cell transition (critical points) among clusters can be tracked.
4. MarkovHC can be applied on diverse types of single-cell omics data.

Summary

Distinguishing cell types and cell states is one of the fundamental questions in single-cell studies. Meanwhile, exploring the lineage relations among cells and finding the path and critical points in the cell fate transition are also of great importance.

Existing unsupervised clustering methods and lineage trajectory reconstruction methods often face several challenges such as clustering data of arbitrary shapes, tracking precise trajectories and identifying critical points. Certain adaptive landscape approach[1-3], which constructs a pseudo-energy landscape of the dynamical system, may be used to explore such problems. However, algorithms based on rigorous metastability theory for constructing the landscape of individual cells are still lacking. Thus, we propose Markov hierarchical clustering algorithm (MarkovHC), which reconstructs multi-scale pseudo-energy landscape by exploiting underlying metastability structure in an exponentially perturbed Markov chain[4]. A Markov process describes the random walk of a hypothetically traveling cell in the corresponding pseudo-energy landscape over possible gene expression states. Technically, MarkovHC integrates tasks of cell classification, trajectory reconstruction, and critical point identification in a single theoretical framework consistent with topological data analysis (TDA)[5].

In addition to the algorithm development and simulation tests, we also applied MarkovHC to diverse types of real biological data: single-cell RNA-Seq data, cytometry data, and single-cell ATAC-Seq data. Remarkably, when applying to single-cell RNA-Seq data of human ESC derived progenitor cells[6], MarkovHC could not only successfully identify known cell types, but also discover new cell types and stages. In addition, when using MarkovHC to analyze single-cell RNA-Seq data of human preimplantation embryos in early development[7], the hierarchical structure of the lineage trajectories were faithfully reconstituted. Furthermore, the critical points representing important stage transitions had also been identified by MarkovHC from early gastric cancer data[8].

In summary, these results demonstrate that MarkovHC is a powerful tool based on rigorous metastability theory to explore hierarchical structures of biological data, identify cell population (basin) and critical point (stage transition), and track lineage trajectory (differentiation path).

Introduction

Cellular activities, such as division, differentiation, arrest, senescence or death in development, disease progression or drug response, can be modeled as stochastic dynamical states in the high-dimensional space of expression levels of genes, proteins, or other molecular quantities[9]. Modeling the dynamics of these processes greatly improves our understanding of the cell fate decision, carcinogenesis, and early diagnoses of diseases. S. Wright[10] first proposed the adaptive landscape concept in 1932 to describe population dynamics. Better known Waddington's epigenetic landscape[11,12] was proposed in the 1950s and has shown increasing importance in molecular and cell biology[13,14], which has widely been applied to analyze cancer[15], study evolutionary theory[16], quantify cell fate decisions[17] and model the hierarchical organization of chromatin[18] in recent years.

Nowadays, high throughput single-cell technologies provide a tremendous amount of data for modeling cellular dynamics, however, the observed data suffer from high level of noises due to both intrinsic biological stochasticity and technical errors in spite of their potential to overcome heterogeneity as shown, e.g. in exploring the lineage relations among differentiated cells[19-23]and predicting cellular communications[24]. Computationally, clustering of cells and delineating density ridges/energy landscapes for trajectory reconstruction, are the two most important tasks, especially in topological data analysis (TDA)[5,25]. Existing methods remain facing number of difficult challenges such as clustering non-convex and mixture-density data, detecting rare clusters, tracking precise trajectories and identifying critical points.

Density ridge, a basic idea in topological data analysis is the essence of clustering and trajectory problems but apparently ignored by a lot of existing methods. In contract, reconstructing the underlying landscape maintaining the topological structure can shed light upon these tasks and overcome most of the challenges mentioned above. Meanwhile, cells are known to be regulated by redundant genes, therefore the gene expression data observed in a high dimensional space can be embedded in lower-dimensional manifolds[26]. Different cell types express distinctive gene sets, which can be viewed as supports in mathematics. They shape different manifolds[27] and form disconnected blocks if ignored boundary points. In addition, the definition of cell types greatly depends on classification criteria and is often ambiguous. The hierarchical structure, however, uses a series of levels correspond to cells clustering in different fineness degrees to overcome the hard threshold problem. Moreover, the hierarchical structure can integrate information concisely[28,29]and reveal the inherent hierarchy in many biological systems[30]. However, algorithms for constructing the landscape of individual cells by a cluster hierarchy based on rigorous mathematical foundation are still lacking. Thus, we developed a novel tool, MarkovHC, to address these issues in single-cell data analysis. MarkovHC is unsupervised (no assumption about the distribution of the data), which makes it intuitive, robust, and widely applicable. It recognizes cell clusters (basins) and cores of clusters (attractors) in different scales, tracks critical points (stage transition points) and transition paths (the trajectory of cell transition) between clusters and packs all information in a hierarchical structure, which

respects the topology of the data.

In this work, we compared MarkovHC with 18 popular clustering methods[23,31-50] (in part C and part D of Supplementary Materials and sheet 1 of Supplementary Table S1) and 4 representative trajectory construction methods (part C of Supplementary Materials) including SPADE[19], monocle series[20,51,52], URD[53] and Waddington-OT[14,30] theoretically. These existing works may be divided into two classes: data-based and model-based. Since MarkovHC is unsupervised, it belongs to the first class. For a given set of samples, a Markov chain in sample state space is constructed (see Supplemental Materials) and its steady state (invariant measure of the Markov dynamics) is obtained to approximating the probability density model with an adjustable coarse-graining scale ("temperature") parameter . Firstly, to validate the effectiveness of MarkovHC and show its advantages, we applied it to simulated data and several benchmark single-cell datasets. Then, we used MarkovHC to explore the hierarchical structure in mouse ES cell differentiation, the transition path and critical points in human embryo development, and critical points in gastric cancer. These results demonstrate that MarkovHC is capable of exploring hierarchical structure in high-dimensional data, finding basins and attractors in different scales, tracking critical points, and transition paths among basins systematically. In summary, MarkovHC is a powerful and extendable method based on rigorous mathematical theory for studying the landscape of single-cell data and shows great potential in analyzing extensive types of single-cell omics data.

Results

The MarkovHC algorithm

In this section, we briefly introduce the proposed algorithm. The general flowchart of MarkovHC is shown in Fig. 1.

A gene expression (or flow cytometry) data matrix ($A=\{a_{gc}: g=1,2,..,G, c=1,2,\ldots,C\}$ is taken as a generic input (Fig. 1A). In principle, treating each distinct cell as a unique point (vertex) or vector in the G-dimensional "expression" space, an initial (i.e. lowest level) Markov network can be built by defining a transition probability matrix from vertex "density" (i.e. higher local vertex weight means more cells could jump out per unit time) and edge similarity (i.e. more interaction edge strength means more "flow conductance"). The "density" may be defined as the number of cells in the same state (same expression level) or vertex degree, and the similarities among states be calculated using the cells shared between nearest neighbor (sNN) states in the (Fig. 1B). Since calculating the global Euclidean distance directly in high-dimensional space tends to introduce noises and biases, sNN is used instead, which is able to explore the natural ("geodesic") distance on the underlying "manifolds" and has been proved to get more robust results[54] in recovering cell subpopulations[55,56] and reconstructing trajectories[27,54,57,58]. In most graphical models, a similarity matrix alone is sufficient to define a Markov transition probability matrix. Since we prefer a pre-processing step (see Methods) for robustness, the degree of each node in the sNN network was used to

measure the point aggregation density, (see Methods for the density score definition). Hence, the density scores of the nodes were used together with the similarity matrix to define the basic Markov transition probability. (Fig. 1C)

Given the transition probability matrix on each level, MarkovHC can find attractors and basins on each level and construct the hierarchical structure systematically and rigorously. In this work, we term basins as clusters and attractors as cores of these clusters. Attractors are steady states in multi-stable systems and basins consist of attractors and their subsidiary points. These two notions have been widely used in previous works[15,59] modeling the nonlinear dynamics of organisms and biological phenomena as dynamics systems.

Based on the metastability theory developed by one of us (Chen et.al. 1996)[4], the hierarchical structure is built in an iterative manner. Firstly, MarkovHC finds the attractors and basins on the first level (details are in part A of Supplementary Materials). Secondly, for each point in attractors, Dijkstra's shortest path algorithm[60] is deployed to detect the shortest transition path from it to its nearest neighboring basin using the pseudo-energy matrix (Fig. 1D). The length of the shortest path represents the pseudo-energy cost in the transition from one basin (cluster) to its nearest neighboring basin (cluster). On the second level, state-space consists of basins on the first level. To form the Markov transition matrix on the second level, the transition probabilities among basins are calculated. Then, MarkovHC uses the same approach as that on the first level to find attractors and basins on the second level. Repeating this process and we can obtain attractors and basins on each level as well as the relations among basins on neighboring levels. The hierarchical structure of the levels (Fig. 1E) is built using these attractors, basins, and topological relations.

On each transition energy level, we can identify the transition path and critical points. The transition path is defined as the most probable transiting path between two basins and the critical points are defined as the closest points between two basins on the transition path. In previous works[61], the saddle points in stochastic nonlinear dynamics were regarded as critical points. However, in practice, it's impossible to locate perfect saddle points with discrete data. Hence, we regard boundary points on the transition path as critical points to approximate saddle points.

Clusters, cores, transition paths and critical points among clusters can be found by MarkovHC.

In this section, we use simple examples to intuitively show how MarkovHC can find basins, attractors, critical points, transition paths among basins, and constructing the intrinsic hierarchical structure.

Although a large number of clustering methods have been developed, the definition of a cluster is still ambiguous. Inspired by topological data analysis (TDA) [25], MarkovHC algorithm is based on the topological connectivity among points, which gives us a clear and reasonable definition of a cluster. Wasserman[25] proposed a concept called level sets for targeted density function. Here, we generalize this idea to discrete

data clustering problems. For any given $\delta > 0$, points with similarities smaller than δ are partitioned into one cluster. The larger δ is used, the larger and fewer clusters are obtained. As δ increases and reach some critical value, some small clusters will merge into a larger one and form hierarchical structures. The above definition of a cluster is clear but lacked biological application because the interaction of two states in a biological system can often be asymmetric (no detailed balance or time-reversal symmetry). Therefore, we modified the symmetric δ connectivity to a directed graph. For any fixed δ, we can obtain a directed graph, from which we can tell basins, attractors, critical points, and transition paths. We used intuitive plots (Fig. 2A) to illustrate how MarkovHC builds up a topologically structure-based hierarchy. Note that when we mention the length of a directed line segment below, we refer to some "asymmetric distance" of biological significance, for example, the amount of transition energy consumption. With a given δ, points that can be mutually connected by directed line segments consist of attractors (white border, Fig. 2A), which are the representative points in the highest density regions of basins. Other points that can link to a given attractor are assigned to the corresponding basin. As δ increasing, according to the order of basins' immergence, the hierarchical structure of these points can be constructed.

The Markov chain modeling cell transition progress includes the possibility of cell returning to the previous state, by which transition paths can be modeled precisely and reasonably. The asymmetric δ connectivity gives MarkovHC the superiority to find the transition path among the basins and the critical points on the transition paths. On a specific level, with a given start basin, the transition path is the shortest path from its attractors to another basin. Within the transition path of two cell types, only a minority of cells are critical in the progress[62]. If cells of one type transit to the critical points, these cells have a big chance to transit to another type. Meanwhile, detecting these critical points is of great importance in applications, for example, diagnosing the disease at early stage[8]. As illustrated in Fig. 2A, the transition path (red box with arrows) and critical points (black edge points) may be found.

For demonstration, we applied MarkovHC to a set of simulated points, consisting of two spheres with a bridge connecting them in a 3-dimensional Euclidean space (Fig. 2B). MarkovHC successfully identified the two basins (green, blue) and attractors (red) located at the core with high-density respectively. Meanwhile, the transition path (orange) between these two basins were tracked. Note that the critical points (purple) around the gap without topological continuity on the transition path were detected.

To accelerate computation, we used some simple clustering methods such as k-means to cluster original points into hundreds of small clusters at the first step when MarkovHC was used to analyze data sets with more than one thousand single-cells. We regarded these small clusters as new aggregated points for further processing. As long as the primary topology of the data is kept, it should be legitimate and helpful to do pre-clustering[63]. An example with the similar topological structure in Fig. 2B but with more points (2100) were analyzed with MarkovHC similarly and results were shown in Fig. S1A-D. The results demonstrate that the base clustering will not destroy the hierarchical structure built by MarkovHC as long as the sizes of small clusters are sufficiently small to avoid clustering heterogeneous points in one small cluster.

MarkovHC is able to cluster non-convex data, hetero-density data, and differentiation stages in a continuum.

Single-cell data in high-dimensional gene space is complicated. Therefore, clustering algorithms that assume data are convex may not always be suitable for single-cell data. For instance, k-means and complete linkage hierarchical clustering get biased results in single-cell data clustering[64]. In addition, the heterogeneity and size variety of cell populations also bring obstacles to traditional density clustering. For instance, DBSCAN[65] is not applicable to identify rare cell types. In many cases, such as tissue development and cell differentiation, the single-cell data are in a continuum, which makes it hard to detect reasonable cluster boundaries. One solution would be using the topology of the data, which only depends on connectivity[5]. Several works have proposed clustering methods based on topology[66,67] and achieved improved performance. MarkovHC also takes advantage of topological connectivity, which embeds the topology of the original data into basins.

We tested MarkovHC on simulated data which consists of four groups of points in different shapes in a two-dimensional space (Fig. 2F). Within each group, points at the edge of the basins are colored grey (Fig. 2F) to indicate low aggregation, while central points are colored black to indicate high aggregation. As Fig. 2G shows, density scores (D scores) reflected point aggregation degree. A hierarchical structure from level 23 to 26 is sketched in Fig. 2I. Three basins could be found on level 26 (Fig. S2C). On level 25, the green basin split into two basins (brown and green, Fig. S2B) with critical points (purple) on the bridge connecting them. On level 24, the brown basin on level 25 was further divided into two basins (brown and dark brown, Fig. S2A), while the latter split into two smaller basins (dark and light brown, Fig. 2H).

Next, we performed MarkovHC on a simulated single-cell data with three stages in a continuum generated by splatter[68]. The data include 1000 cells in three stages in a 5000 genes space. These 1000 cells were projected onto the principal component space and colored according to their stages (Fig. 2L, Fig. S3B). Meanwhile, these are "low-density valleys" on the density landscape, which correspond to energy barriers separating basins on the pseudo-energy landscape (Fig. S3A). On level 17 these basins and corresponding attractors were identified (Fig. 2M). In comparison with other algorithms, adjust rand index (ARI) (Fig. 2N) and normalized mutual information (NMI) (Fig. S3C) were calculated, which showed that MarkovHC performed better than other methods on this dataset. Specifically, we checked the transition path (orange) and the critical points (purple) from the brown basin to the blue basin (Fig. 2O). The differentially expressed genes along the transition path (Fig. 2P) and the dramatic changes of gene expression (Fig. 2Q) around the critical points demonstrate faithfully the kinetics of gene expression along the transition path.

In the above simulations, we verified three merits of MarkovHC. Firstly, the sparse basin (grey in Fig. 2H and Fig. S2) was successfully detected, demonstrating the ability of MarkovHC to detect small basins (rare types) sensitively. Secondly, non-convex basins with arbitrary shapes (blue in Fig. 2H and Fig. S2, helixes in Fig. 2K) could be

handled by MarkovHC. Last, MarkovHC could find clusters in the continuum (Fig. 2M). These results demonstrate the potentials for MarkovHC to identify rare cell types, handle non-convex in high-dimensional single-cell data, and track development trajectory.

MarkovHC can consistently achieve accuracy in clustering diverse types of single-cell omics data

Accurate clustering is the foundation for correct identification of cell populations and new cell types. Moving to the real data, we compared MarkovHC with popular single-cell RNA-Seq clustering methods including Seurat, SIMLR, SC3, k-means, hierarchical clustering with average linkage (ALINK), HDBSCAN, spectral clustering and model-based clustering on five benchmark datasets[38,69] (Fig. 3A-E). The ARIs and NMIs of MarkovHC are comparable to those of these popular clustering methods (Fig. 3F). The comparison among MarkovHC, Seurat with default parameters, and hierarchical clustering with other linkages are in Fig. S4. Remarkably, the cell atlas of C. elegans embryogenesis[69] (Fig. 3E) is a golden benchmark to test development trajectory and lineage clustering algorithm, in which 6090 cells were annotated as 27 sub-cell types. MarkovHC constructed the hierarchical structure of this benchmark and reconstructed sub-cell types on level 4 yielding 0.64 and 0.73 on ARI and NMI respectively (Fig. 3F), which outperformed the other methods.

In addition, the classification and hierarchical structures of human peripheral blood mononuclear cells have been well studied in many researches[70-72]. We applied MarkovHC on the 33k PBMCs dataset to validate the applicability of MarkovHC in analyzing datasets with a large number of cells. On level 113, basins correspond to B cells, T cells, NK cells, megakaryocyte, and monocyte (Fig. S5) could be identified. To identify subtypes of basins, we checked lower levels on the hierarchical structure. On level 111, the T cell basin was divided into CD4+ T cell basin and CD8+ T cell basin. On level 110, CD4+ T cell basin was further divided into CD4+ naïve, and CD4+ memory T cell basins. Similarly, the subtypes of B cell, CD8+ T cell, and monocyte could also be recovered on lower levels (Fig. S5). These results demonstrate the power of the hierarchical structure, with which a specific cell type can be identified using marker genes, and the sub-cell types or cell states can be checked in detail on lower levels.

Moreover, MarkovHC is based on topology among samples without an assumption of data distribution, it can be widely used to build the hierarchical structures of other omics data besides scRNA-Seq data. As long as an appropriate similarity matrix is provided, MarkovHC algorithm can be applied. In single-cell research, another widely used technology in profiling cell populations is mass cytometry. Here, we verified the applicability of MarkovHC on mass cytometry data from Anchang's paper[63], which includes about 150,000 cells and 38 parameters. In this dataset, cells were obtained from a healthy donor's bone marrow and 24 subpopulations were labeled according to the protein expression of 38 surface markers. We randomly sampled 10009 cells from

the dataset in proportion and projected these cells into 2-dimensional space by UMAP (Fig. S6). The results showed that MarkovHC got the comparable performance to Seurat (Fig. S7).

Nevertheless, MarkovHC can help to perform integration analyses, for instance, the integration of scRNA-Seq and scATAC data. In Fig. S8, there are 464 scRNA-Seq and 96 scATAC-seq samples generated for the retinoic acid-induced mESCs differentiation at day 4 in Zeng *et. al.*'s work[73]. MarkovHC identified three basins in scATAC-Seq (Fig. S8A) data and scRNA-Seq (Fig. S8B) data respectively. We used tSNE to visualize these data (Fig. S8A, B), and identified differentially chromatin accessibility in scATAC-Seq (Fig. S8C) and differentially expressed genes in scRNA-Seq (Fig. S8D). The peaks locating highly expressed genes are more accessible than those of repressed genes. For instance, PDE10A, RSPH1, and FBXL17 were highly expressed in basin 1. And the peaks, chr17-8725942-8730132, chr17-31394370-31396459, and chr17-63652707-63655013, which correspond to these three highly expressed genes, were open in basin1. In basin 2, highly expressed genes, RERE and SP8, correspond to open peaks, chr4-149929815-149933175 and chr12-120087188-120088390 respectively. And in basin 3, expression of these genes and accessibility of corresponding peaks were low. Therefore, three basins from these two datasets could be matched according to the peak pattern and the gene pattern respectively, which were also consistent with the three subpopulations inferred by Zeng *et. al.*[73]. These results demonstrate the potential of MarkovHC in the integration of multi-omics data.

In order to verify the effectiveness of MarkovHC in discovering new biological knowledge from real data, we applied it to three other single-cell RNA-Seq datasets, i. e., human embryonic stem cell differentiation to definitive endoderm[6], human preimplantation embryos[7], and gastric antral biopsies in the following sections.

MarkovHC's hierarchical structure revealed cell lineages, cell types, and cell stages simultaneously.

In this section, we show that the hierarchical structure obtained by MarkovHC could be used to explore cell lineages, cell types, and cell stages. In human cell atlas[74], cell lineage means developmental relationship of cells; cell type implies a notion of homeostatic persistence; and cell state refers to more inducible or transient properties. However, the boundaries among these concepts at the gene transcriptional level can be fuzzy, partly may be due to the limit of our knowledge and understanding of cellular dynamics. Fortunately, data-driven approaches, especially hierarchical algorithms can be helpful to refine these concepts.

Single-cell RNA-seq data[6] from experiments profiling snapshots of lineage-specific progenitor cells differentiated from H1 human ES cells in specific induction circumstances was used to illustrate this point. Cells were sorted with specific markers using fluorescence-activated cell sorting (FACS) and captured for single-cell RNA-seq using the Fluidigm C1 system (Scale bar 200μm). In total, 1018 single-cells of six cell types were analyzed in this dataset. All cells were projected into 2-dimensional space

using phateR[75], in which six clusters could be found (Fig. 4A). Low-level results could be swamped by a lot of noises, while high level results may average out true biological signals. Therefore, appropriate levels for downstream analyses need to be chosen based on biological prior knowledge and low dimensional visualization. There were six basins on level 42 (Fig. 4B), around which we went for further analyses.

On level 42, the six basins (Fig. 4B) match the six cell types in ground truth (Fig. 4A). On level 41, we noticed that the neuronal progenitor cells were divided into two basins. Meanwhile, the H1 and H9 ES basin was also divided into two individual basins on level 40. Relations among these basins can be visualized and studied using the hierarchical structure (Fig. 4C). Specifically, on the higher levels, the difference among basins might reflect different cell lineages, while on the lower levels it could correspond to cell types or stages. Meanwhile, the attractors are representative cells of basins, which consist of cells with significantly higher homogeneity than that of other cells in basins (Fig. S9).

To closely inspect the difference among these basins, the 50 most significantly up-regulated genes in each basin were found by Wilcoxon rank sum test (Fig. 4D left). In checking the biological significance of these basins, we selected 34 lineage-specific markers from Chu's paper[6] to show the expression patterns of basins (Fig. 4D right). In these two figures (Fig. 4D), the pattern of basins obtained by MarkovHC through differentially expressed genes (in sheet 2 of Supplementary Table S1) were consistent with that of the selected lineage-specific markers, which demonstrated that the basins were of biological significance. To reveal the difference between H1 and H9 ES cell basins on level 40 and two small basins on level 41 in neuronal progenitor cells, we enriched gene ontology terms using the differentially expressed genes. On level 40, the differentially expressed genes (Fig. S10) between the two basins in ES cells were enriched in terms related to cell division and cell cycle (Fig. 4E), which suggests these ES cells were in different cell cycle stages. In the brown basin, cells were in more active division stage, while cells in the light brown basin were in the suppressive division stage. On level 41, the up-regulated genes (Fig. S11) of cells in the blue basin were enriched in neuron differentiation (Fig. S12A), while terms were enriched in synapse organization and forebrain development (Fig. S12B) in the light blue basin, which indicated that two subtypes of cells could be found in neuronal progenitor cells. Obviously, the definitive endoderm cell and endothelial cell are two lineages that would be merged on level 43. From the bottom to top in the hierarchical structure (Fig. 4C), cell stages, cell types and cell lineages are presented, which gives a good example of how hierarchical structure can be used to define cell identities and relationship at different resolutions along the developmental lineage trees.

MarkovHC revealed transition paths and critical points in human preimplantation embryo development.

MarkovHC can be used to track transition paths among cell types and detect critical points. This dataset includes scRNA-Seq data[7] of 1529 single-cells collected

from 88 human preimplantation embryos at different stages. The overall layout of these 1529 cells in the space of PHATE1-3[75] can reflect the order of samples (Fig. 5A).

Cells collected at the same time point can be heterogeneous due to stochasticity in cell differentiation. We first identified homogeneous clusters (Fig. 5B) of cells with MarkovHC and then tracked the transition paths and critical points among these clusters. Meanwhile, the bar plot (Fig. S13) shows that there is only one basin in E3, and then this basin separated into two basins in E4. Basin3 is dominant in E5 early and more new basins emerge after that, which suggests that new cell populations burst at that time. As the work from Petropoulos *et. al.*[7] demonstrated, cells in these data were identified as eight cell groups, including pre-lineages, trophectoderm (TE), inner cell mass (ICM), epiblast (EPI), primitive endoderm (PE), E5mid (these cells collected at seven time points), mural and polar (Fig. 5K). For there were eight basins on level 8, we selected the levels (from level 6 to level 15) around level 8 from the hierarchical structure (Fig. 5C) for further analyses.

To reveal the biology processes in each basin, the 50 most significantly up-regulated genes of each basin were identified using Wilcoxon rank sum test (in sheet 3 of Supplementary Table S1). Overall, genes were enriched on gene ontology terms related to the process of embryonic development, while specific groups of gene ontology terms could be found in each basin (Fig. 5D, in sheet 4 of Supplementary Table S1). To infer transition paths among these basins, we calculated the pseudo-energy matrix (Fig. 5E). Because a basin transit to its closest basin (yellow boxes in Fig. 5E) on the energy landscape with the highest probability, we inferred the trajectories (Fig. 5F) and built the development hierarchy (Fig. 5H) which is basically consistent with the knowledge of human embryo development hierarchy (Fig. 5K). To identify each basin, we used canonical marker genes used by Petropoulos *et. al.*[7] and LifeMap Discovery (https://discovery.lifemapsc.com/in-vivo-development/) in the embryonic development (Fig. 5G) and identified four main cell types which were 8-cell embryo (GATA6-, SOX7-, SOX17-), morula cell (GATA6+, SOX7-), ICM (inner cell mass, SOX2+ and PDGFRA+) and TE (trophectoderm, GATA2+ and GATA3+) according to the marker genes expression. To identify sub-groups in TE, we used the differentially expressed genes of mural (WIPF3+, PLCE1+, PDCD4+, SLC15A2+) and polar cells (GCM1+, NRP1+, SLCO4A1+)[7]. To further explore sub-groups in ICM, we checked the sub-basins of ICM on level 4 and found that there were three main basins (Fig. 5I). According to their marker genes expression (Fig. 5J), we identified these three basins as PE (PDGFRA, FGFR2, LAMA4, and HNF1B), EPI (SOX2, TDGF1, DPPA5, GDF3, and PRDM14) and ICM.early. To further check gene-expression “flow” trends along the transition path from 8-cell embryo to ICM, we tracked the cells in the path (Fig. 5L, yellow points indicate cells on the transition path and purple points indicate critical points from morula cells to ICM). 1631 differentially expressed genes along this path were found (Fig. 5M, in sheet 5 of Supplementary Table S1). GATA2, PTGES, PDGFA (Fig. 5N) showed increasing trends, while SLC34A2, MYC, ARGFX (Fig. 5O) displayed decreasing trends. Dramatic changes in gene expression were observed around critical points (purple points), indicating these points as pivotal stages in the developmental process. These results demonstrate the significance of transition paths

and critical points in understanding the developmental dynamics and time-series biological data.

MarkovHC detected critical points from MSCs to gastric cancer cells

The transition process from normal gastric cells to cancer cells is complex and poorly studied. Many studies have been carried out to better understand the carcinogenesis[76-78]. For example, Zhang *et. al.*[8] conducted a single-cell transcriptomic study on gastric antral biopsies and identified gastric cancer related cell populations. As mentioned above, transition paths and critical points can be helpful to reveal cluster relations. Here, we used MarkovHC to analyze 830 mesenchymal stem cells (MSCs) and 649 MSC-origin early gastric cancer cells (EGCs) (Fig. 6A) from one patient in Zhang *et. al.*'s study[8].

MarkovHC found two basins on level 22 (Fig. 6B), which was basically consistent with the ground truth (Fig. 6A). However, it was obvious that MSCs and EGCs got mixed up in the boundary region (Fig. 6A). Therefore, we checked the lower levels and found that the MSC basin (green) and EGC basin (red) split into three and two basins on level 19 respectively (Fig. 6C). Not surprisingly, basin 3 and 4 (Fig. 6C) were a mixture of EGCs and MSCs. The differentially expressed genes of these two basins to basin 2 (Fig. 6G) suggested that these EGCs might be in two different transition states. Meanwhile, basin 1 and 5 might be two subpopulations in EGCs (Fig. 6G). Based on these observation, two transition paths from basin 2 to basin 1 were plotted (yellow and brown arrows in Fig. 6D). From the pseudo-energy matrix (Fig. 6E), we could see it would take less energy to transit from MSCs (basin 2) to transition states (basin 3 and 4) than directly to EGCs (basin 1 and 5). Namely, MSCs could transit to transition states with larger probability. Interestingly, we also noticed that it would cost more energy to transit from basin 3 to EGCs (basin 1 and 5) than from basin 4 to EGCs. This might suggest that basin 3 be more stable than basin 4 on the landscape and reflect the heterogeneity of these two transit states, with the latter having a greater potential to process into cancer. In addition, the two EGC basins (basin 1 and 5) could transit to each other suggesting plasticity and instability of EGCs. We presented the inferred hierarchy of these cell populations in Fig. 6F.

To further investigate the gene expression patterns in these basins, Wilcoxon rank sum test was deployed to find the top 50 differentially expressed genes (DEGs) (Fig. 6G, in sheet 6 of Supplementary Table S1), with which the GO terms were enriched in the relevant functional categories (in sheet 7 of Supplementary Table S1). In the heatmap, we observed relative homogeneity in MSCs (basin 2) and obvious heterogeneity in transition states (basin 3) and EGCs (basin 1 and 5). Specifically, highly expressed genes in basin 2 were enriched in response to topologically incorrect protein and unfolded protein related terms. This suggested that although basin 2 cells were MSC-like cells, they might have been stimulated by environment factors and could be resisting to incorrect protein folding. DEGs in basin 3 were enriched in protein targeting and protein localization related terms, while DEGs in basin 4 were enriched in intranuclear biological processes such as DNA strand elongation and DNA

conformation change. The differences indicated that these two transition states responded to environmental stimulus differently. Interestingly, the up-regulated genes in basin 1 were enriched in epidermis development and primary alcohol catabolic process related terms, while those in basin 5 were response to metal ion and detoxification related terms. These results suggested that the formation of basin 5 might be related to metal ion stimulus, while basin 1 might be associated with alcohol stimulus and epidermis damage. The latter was partially supported by the fact that this patient (p8 in Zhang *et. al.*'s paper[8]) had a chronic alcohol consumption history.

To check the transition trajectories from MSCs to EGCs (path 1 and path 2 in Fig. 6D), we examined the gene expressions of some marker genes along paths (Fig 6.H-J). MSC markers (OLFM4, EPHB2, and SOX9) and EGC markers (KLK10, SLC11A2, SULT2B1, KLK7, ECM1 and LMTK3) were gradually decreasing and increasing respectively (Fig. 6H). DEGs along path 1 and 2 were also identified (Fig. 6I-J), and used for gene function enrichment analysis(in sheet 8 of Supplementary Table S1).

Basically, two clusters of DEGs could be found in Fig. 6I and 6J. In path 1, the decreased genes were enriched in response to chemokine, while the increased genes were enriched in cellular response to metal ion suggesting that the potential disease progress might be driven by metal pollution[79]. Specifically, we observed OLFM4[80], a marker for stem cells in human intestine, deceased from MSCs to EGCs, while CEACAM6[81], which plays important roles in invasion and metastasis in Gastric Cancer, was increased (Fig. 6K). In path 2, the decreased genes were enriched in response to mechanical stimulus, DNA damage response and signal transduction by p53 class mediator resulting in cell cycle arrest, while the increased genes were enriched in neutrophil mediated immunity suggesting this path could be the classic p53 related progression of gastric cancer[82,83]. One key gene, SOX4, deceased along path 2 (Fig. 6L), which was consistent with the results that MiR-596 down-regulated SOX4 expression and was a potential novel biomarker for gastric cancer[84]. Another marker, NEAT1[85], is a long non-coding RNA promoting viability and migration of gastric cancer cells through up-regulation of microRNA-17, which increased along path 2 (Fig. 6L). Furthermore, expression values of these genes dramatically changed around the critical points (purple points in Fig. 6K, L), which suggested critical points were unstable cells locating at the boundaries of the basins. The two paths indicated that there might be two separating trajectories from MSCs to EGCs for this patient, corresponding to two types of potential disease progression routes. These results again demonstrate the great capability and flexibility of MarkovHC in clarifying relationship of the cell subpopulations and the usefulness of transition paths and critical points in revealing the underlying mechanisms.

Discussion

In this paper, we developed MarkovHC based on the metastability of exponentially perturbed Markov chain to explore the cell landscape which corresponds to the intrinsic topological structure of the data. And we also developed an R package, "MarkovHC" (https://github.com/ZhenyiWangTHU/MarkovHC). This tool is easy to use and only

requires a feature×cell data matrix as the input. We showed that MarkovHC can cluster cells into populations at different resolutions, which could correspond to cell lineages, cell types, and cell states. Furthermore, we showed that the transition paths and the critical transition points among cell populations tracked by MarkovHC can reveal developmental processes, such as human embryonic development or disease procession.

Besides clustering and identifying differentiation paths, there could be other interesting type of analyses with the help of MarkovHC. First, one could calculate pseudo-time[20,51,52,86] from the hierarchical structure. Though hierarchical ordering itself already has causal interpretation, one could refine it using the same framework as shown in this paper. Second, one could start out from a given relevant biological level, e.g. based on prior knowledge or other clustering method, and use MarkovHC to explore next level up or down in the hierarchy. Thirdly, the hierarchical structure obtained by MarkovHC could be used to integrating other or multi-omics data, e.g. by using UnionCom[87] recently developed by Cao *et. al*., for the unsupervised topological alignment of single-cell multi-omics integration.

Finally, our method, which is based on a general mathematics theory, is flexible and robust. In the preprocessing, we used sNN to derive the initial transition probability matrix in order to overcoming sparsity/dropout problems as well as computational burden. One could certainly use any other advanced manifold-learning methods[38,88] to build the ground similarity matrix. Although such preliminary processing step has little effect[63] on the topological structure at biologically relevant levels, it is possible to further optimize to best balance the tradeoff between efficiency and accuracy for a specific data set. Also, it is possible to apply metastability theory to more other types of data. For example, it would be interesting to adapt it to explore the hierarchical 3D genome structure. To our knowledge, there's still no research about modeling the landscape of 3D genome structure using a bottom-up hierarchical structure like ours.

Methods

Data preprocessing.

We employed functions in Seurat[34] to preprocess scRNA-Seq data matrixes. Read count matrixes were normalized by NormalizeData function with factor 10000, and 3000 highly variable features were selected using FindVariableFeatures function. Then, the normalized matrixes were scaled by ScaleData function. Next, RunPCA function was used to perform linear dimensional reduction and dimensions were selected by an heuristic method generating an "Elbow plot": principle components were ranked based on the percentage of variance explained by each one, and we always can observe an "elbow", suggesting that the top PCs before the "elbow" could capture the majority of true signal. And these dimension reduced matrixes were the input to MarkovHC algorithm.

In particular, the preprocessing steps in DC3[73] were used to normalize the data of 464 scRNA-Seq and 96 scATAC-seq samples (Fig. S8) from their work. And the cytometry data (Fig. S6) were scaled by log(x + sqrt(x^2 + 1)) and performed linear dimensional reduction by RunPCA function.

Overview of the MarkovHC algorithm

The input data of MarkovHC algorithm is the matrix of genes by cells or proteins by cells, representing the detected mRNA or proteins expression in each cell. Without loss of generality, let s_1, s_2,…, s_n be a set of d-dimensional data points. Traditionally, the clustering goal is to divide them into different basins that accords with intuition and are interpretable:

$B_1 = \{s_{11}, s_{12}, ..., s_{1p1}\}, ..., B_k = \{s_{k1}, s_{k2}, ..., s_{1pk}\}$ where $\{B_i, i = 1, ..., \mathrm{k}\}$ usually have no overlap. As we know, sometimes k, the number of clusters, is cannot be decide beforehand and may be varying depending on different goals. Therefore, the Markov algorithm is allowed to explore the intrinsic hierarchical structure in the data points. Fig. S14 shows the target hierarchical structure of MarkovHC, where every $\{B_i^j, i = 1, ..., k_j; j = 1, ..., r\}$ is the subset of $\{s_1, ..., s_n\}$ and B_i^p is the subset of some basins on level p+1. Note that we allow B_i^j on the same level have some overlaps defined exactly by the shared points. They are of great significance in the hierarchical structure. If there are points shared by B_i^j and B_k^j , once states in B_i^j reach to these shared points after a long trip, they can easily develop into points in B_k^j in a short time. Of course, sharing some points is not a sufficient condition for combining B_i^j and B_k^j together as. If it is very hard for core points of B_i^j and B_k^j to develop into points, B_i^j and B_k^j will not combine on the next level. Similar to the structure produced by traditional hierarchical clustering algorithms, when we select a level, there is a clustering result. Unlike those traditional hierarchical structures, structure produced by MarkovHC is more concise and meaningful. Every level's clustering result is interpretable. Different levels correspond to different δ mentioned in Fig. 2. If we want to obtain a more detailed clustering result, we can select a lower level.

We can define a measure of the relation between different sample points. Some measures are undirected, generally called similarity or dissimilarity. For example, k-means and standard hierarchical clustering usually use Euclidean distance as the dissimilarity between points. Some measures are directed. For example, DBSCAN[65] gives a directed definition of directly density-reachability. Obviously, a directed relation is more reasonable because in many cases, the status of two sample points is not equal. Let's take a biological example. Assume A is a point representing a hematopoietic stem cell and B is a point representing a red blood cell. A is able to develop into B while the inverse process is impossible. Therefore, we consider using Markov chains to describe the whole dataset. Continue with the example above, we can set the transition probability p(A, B) > 0 while p(B, A) = 0, which means that A can reach B but B can not return to A. If we have a Markov chain to describe the relations

of different points, it is very natural to see points belonging to a same recurrent class as core points of a cluster. Points that are not recurrent states but can reach to more than one recurrent class can be seen as shared points. However, a single Markov chain is not enough to describe the intrinsic structure of data points on the ground that one Markov chain only corresponds to one clustering criterion. As we have mentioned above, for many problems, building a hierarchical structure can completely show the information contained in data. As a result, we use a family of Markov chains to describe the dataset. We can regard the input dataset as a finite state-space $S=\{1,...,N\}$. $\{P^{\beta}:\beta\in[0,\infty)\}$ are a family of Markov chain transition matrices on S satisfying that there exists:

$$p_{i,j}=p(i\to j)=p(j|i)=\lim_{\beta\to\infty}p^{\beta}(j|i)=p^{\infty}(j|i),\forall i,j\in S$$

For a fixed β, if state i has a stronger directed relation towards state j, $p^{\beta}(j|i)$ would be bigger. The parameter β controls the movement intensity of states. When β becomes bigger, the whole state-space will become more inactive that the expected time taken from core states in one cluster to another cluster will exponentially rise. Finally, when β approaches infinity, the state movement is basically limited in core points of the same cluster. Taking advantage of the difference of movement time scale between points, we can build a hierarchical structure of the state-space. We assume the following limit exists:

$$c_{i,j}=c(i\to j)=\lim_{\beta\to\infty}-\frac{1}{\beta}\log p^{\beta}(j|i)$$

We assume that

$$c_{i,j}>0, if\ p^{\infty}(j|i)=0$$

$c_{i,j}$ can be viewed as pseudo-energy barrier from state i to move to state j.

Under the assumptions given above, Markov chains family, which is composed of Markov chains with transition matrices $\{P^{\beta}:\beta\in[0,\infty)\}$ is an exponential perturbation of the Markov chain with transition matrix P^{∞}. The core of our theory is the pseudo-energy matrix C, which concentrates all asymptotic information provided by the Markov chain family. Actually, the time scale for point ξ to directly move to point η is $e^{\beta C_{\xi\eta}}$. The time scale for points in one cluster to move to another cluster can also be expressed by entries of C. As a result, only with the help of C, we can construct the target hierarchical structure iteratively.

Level 1 clustering result should be the most detailed and basic one. Therefore, we consider the most inactive Markov chain, i.e. the Markov chain with transition matrix P^{β}. Firstly, we find out all recurrent classes and see them as core points of the

corresponding level 1 clusters. Actually, if ξ and η belong to a same recurrent class, there exists a path led from ξ to η such that the sum of pseudo-energy along this path is 0, which means that ξ can move to η effortlessly, and vice versa. We denote these level 1 recurrent classes as $\left\{A_i^1\right\}$, also called as "attractors". We denote the cluster whose attractor is $\left\{A_i^1\right\}$ as B_i^1, also called as a basin. Then, we need to assign non-recurrent states to appropriate clusters. Given an attractor A_i^1 and a non-recurrent state τ, we say τ belongs to basin B_i^1 if there exists a directed path from τ to A_i^1 that the sum of pseudo-energy on this path is 0. Though different attractors do no overlap, it is possible that a non-recurrent state τ shared by two or more basins. We give a simple example in Fig. S15. Assume data points are 1-dimensional and they are plot at the top of Fig. S15. The value of vertical coordinate is the pseudo-energy, which is defined as the negative logarithm of density here. The time scale for points in $A_i^1(A_j^1)$ to reach τ is $e^{T\left(B_i^1\right)\beta}\left(e^{T\left(B_j^1\right)\beta}\right)$ while the time scale for τ to reach $A_i^1(A_j^1)$ is a constant irrespective of β. ($T\left(B_i^1\right)$ can be viewed as the least pseudo-energy cost for points in A_i^1 to reach τ.) Therefore, it takes a long time for the points in attractors to move to the shared points while the shared points can jump into the attractors in a very short time. Of course, the time scales for points in A_i^1 and A_j^1 to move to τ are different. Such a difference is exactly the criterion used to define the relation between level 1 clusters (basins) and to construct higher-level Markov chains.

The construction of level 2 clustering result is based on the degree of difficulty for points in level 1 attractor A_i^1 to escape from their corresponding basin B_i^1, i.e. $T\left(B_i^1\right)$. $T\left(B_i^1\right)$ is the least pseudo-energy cost for points in A_i^1 to reach other basins. Considering the characteristic of critical points, if B_i^1 has critical points shared by other basins, then $T\left(B_i^1\right)$ equals to the least pseudo-energy cost for points in A_i^1 to reach B_i^1's shared points. We construct a new Markov chain to describe the relation between level 1 basins. The new state-space is $S_1=\left\{B_i^1 \middle| i=1,...,k_1\right\}$, where k_1 is the number of level 1 basins. We set the transition probability $p\left(B_i^1 \middle| B_j^1\right)>0$ if and only if

$T\left(B_i^1\right) = \min_{1 \le k \le k_1} T\left(B_k^1\right)$ and one of the following two conditions holds:

1. B_i^1 and B_j^1 have common shared points;

2. There exists a path from A_i^1 to B_j^1 that the sum of pseudo-energy equals to $T\left(B_i^1\right)$. After normalization, we can obtain a transition probability matrix and create a Markov chain on state-space S_1. Similar to level 1 clustering result, level 2 clustering results can be easily obtained on the basis of this new Markov chain. We illustrated it in Fig. S16. There are four basins on level 1 (Fig. S16A). We wrote down some time scales in the figure. For example, the time scale for points in A_1^1 to move to shared points shared by B_1^1 and B_2^1 is e^{β}. The rest of time scales not shown in the picture are all assumed to be bigger than $e^{4\beta}$. Based on our criteria, $B_1^1 \cup B_2^1 \cup B_4^1$ is a new level 2 basin (Fig. S16B), which is denoted as B_2^1. In addition, B_3^1 is another level 2 basin, which is denoted as B_2^2. We should note that the attractor of B_2^1 is $A_1^1 \cup A_2^1$, but not $A_1^1 \cup A_2^1 \cup A_4^1$, for the reason that moving from A_1^1 to B_4^1 is far more difficult than moving from A_4^1 to B_1^1.

We can construct the higher-level clustering results in a similar way until the complete target hierarchical structure is finished. Actually, we just transform the information contained in matrix C into a hierarchical structure that can better show the intrinsic relations of different data points. A more detailed derivation, which contains some definitions and theorems regarding to metastability of exponentially perturbed Markov chain is in the Supplementary Materials (rigorous mathematical proofs can be found in the original Chen *et. al.* 1996 paper[4]).

Technical details of the MarkovHC algorithm

Step1: Calculate sNN similarity matrix and construct a sNN network

The sNN (shared Nearest Neighbors) between two points is the number of K nearest neighbor intersection, therefore sNN is a robust similarity measurement. Single-cell data is high dimensional and has serious dropout events, thus we used sNN to measure similarities among cells and computed a cell-by-cell sNN similarity matrix $s_{i,j}$ between node i and node j, which is also equal to the strength of connectivity between the two nodes:

$$s_{i,j} = \#\left\{neighbors \in KNN_i \cap KNN_j\right\}$$

where KNN_i is the K nearest neighbors of node i. Then we use the matrix to

construct a network. The nodes of the network are cells and the edges of the network are the similarity among cells.

Step2: Calculate density score of each node

Accurate density estimation is a challenging task in high dimensional data, especially in single-cell RNA sequencing data. Therefore, we use the degree of each node in sNN network to measure the node aggregation as our density score:

$$D_i = \sum_{j=1}^{n} a(i,j), a(i,j) = \begin{cases} 0, edges(i,j) = \varnothing \\ 1, edges(i,j) \neq \varnothing \end{cases}$$

Step3: Calculate the Markov transition matrix and the pseudo-energy matrix

The transition probability from node i to node j is defined:

$$p_{i,j} = p(i \to j) = p(j|i) = \lim_{\beta \to \infty} \frac{e^{-\beta \frac{1}{s^{w1}_{i,j}} \left(\frac{D_i}{D_j}\right)^{w2}}}{\sum_{j=1}^{n} e^{-\beta \frac{1}{s^{w1}_{i,j}} \left(\frac{D_i}{D_j}\right)^{w2}}}$$

where w1 and w2 are constant parameters determining the tradeoff between the importance of similarity and density score, and both of them are set to 2 by default. For the convenience of calculation, we can derive the concise formula of p(j|i) :

$$denote\, g_{i,j} = \frac{1}{s^{w1}_{i,j}} \left(\frac{D_i}{D_j}\right)^{w2}, where\, s_{i,j}\, is\, the\, similarity\, of\, i\, and\, j,$$

$$D_i\, is\, the\, centrality\, score\, of\, i.$$

$$p(j|i) = \lim_{\beta \to \infty} \frac{e^{-\beta g_{i,j}}}{\sum_{j=1}^{n} e^{-\beta g_{i,j}}}$$

$$= \lim_{\beta \to \infty} \frac{e^{-\beta g_{i,j}} e^{\beta g_{i,0}}}{\sum_{j=1}^{n} e^{-\beta g_{i,j}} e^{\beta g_{i,0}}}, \text{where } g_{i,0} = \min(g_{i,j}), \mathrm{j} = 1, \ldots, \mathrm{n}$$

$$= \lim_{\beta \to \infty} \frac{e^{-\beta g_{i,j} + \beta g_{i,0}}}{\sum_{j=1}^{n} e^{-\beta g_{i,j} + \beta g_{i,0}}}$$

$$= \lim_{\beta \to \infty} \frac{e^{-\beta (g_{i,j} - g_{i,0})}}{\sum_{j=1}^{n} e^{-\beta (g_{i,j} - g_{i,0})}}$$

$$= \begin{cases} 0, & g_{i,j} > g_{i,0} \\ m_i^{-1} & \end{cases}, m_i = \{ \#\, of\, j,\, st\, g_{i,j} = g_{i,0} \}$$

We denote the Markov transition matrix as P1 below.

The consumed pseudo-energy by the direct transition from node i to node j is:

$$c_{i,\mathrm{j}} = \lim_{\beta \to \infty} -\frac{1}{\beta} \log p^{\beta}(j|i)$$

We also can derive the concise formula for calculation convenience:

$$
\begin{aligned}
c_{i,\mathrm{j}} &= \lim_{\beta\to\infty} -\frac{1}{\beta}\log p^{\beta}\left(j|i\right) \\
&= \lim_{\beta\to\infty} -\frac{1}{\beta}\log\frac{e^{-\beta g_{i,j}}}{\sum_{j=1}^{n} e^{-\beta g_{i,j}}} \\
&= g_{i,j} + \lim_{\beta\to\infty}\frac{1}{\beta}\log\sum_{j=1}^{n} e^{-\beta g_{i,j}} \\
&= g_{i,j} + \lim_{\beta\to\infty}\frac{\sum_{j=1}^{n} -g_{i,j}e^{-\beta g_{i,j}}}{\sum_{j=1}^{n} e^{-\beta g_{i,j}}} \\
&= g_{i,j} + \lim_{\beta\to\infty}\frac{\sum_{j=1}^{n} -g_{i,j}e^{-\beta g_{i,j}}e^{-\beta g_{i,0}}}{\sum_{j=1}^{n} e^{-\beta g_{i,j}}e^{-\beta g_{i,0}}}, where\, g_{i,0} = \min(g_{i,j}), \mathrm{j} = 1,\ldots,\mathrm{n} \\
&= g_{i,j} + \frac{\sum_{0} -g_{i,0}}{\#\left\{m_i : g_{i,0} = \min(\mathrm{g}_{i,j}), \mathrm{j} = 1,2,\ldots,\mathrm{n}\right\}} \\
&= g_{i,j} - g_{i,0}
\end{aligned}
$$

This pseudo-energy matrix represents the pseudo-energy cost when making a direct transition between nodes. We can calculate geodesics (min-energy path) on the surface of this "energy landscape" by using Dijkstra algorithm[60]. If node i cannot transit to node j, the corresponding entry in the matrix is set to be infinite. We denote this reduced pseudo-energy matrix as C1 below.

Step4: Pre-clustering

To accelerate the speed of steps below, we can perform pre-clustering at the first stage to cluster all nodes into hundreds of small clusters. These small clusters are homogenous. The number of rough clusters is a customizable parameter in our tools and there are five optional pre- clustering methods, which are k-means, single linkage hierarchical clustering, complete linkage hierarchical clustering, average linkage clustering, and Louvain clustering. The mean density score of the nodes in the small cluster is used as the density score of the small cluster. The number of shared nearest neighbors among nodes is defined as the similarity of them, and the maximum similarity among nodes in two small clusters is used to define the similarity between cluster m and cluster n:

$$s_{m,n} = \max(s_{i,j}), i \in cluster_m, j \in cluster_n$$

Users could skip this pre-clustering step, but we recommend this for two reasons. First, such preliminary clustering does no harm and won't change the upper-level hierarchical structure, as long as the number of clusters is large enough after this step.

Second, pre-clustering can significantly reduce computational time. By default, we cluster all nodes into 200 small clusters using k-means.

If pre-clustering in this step was performed, we regard small clusters as nodes. For convenience, we still call these 'small clusters' as 'nodes' below.

Step5: Find the attractors and partition the basins of each level to build the hierarchical structure.

Firstly, we explain concepts of attractor, basin, the Markov transition matrix P, and the pseudo-energy matrix C. Attractors and basins are sets of nodes. And basins also can be regarded as clusters and attractors can be regarded as cores of clusters. Matrix P represents the transition probability between basins, and the pseudo-energy matrix represents the energy cost by making the transition between the basins.

On level 1, every node is an attractor and also is a basin according to a mathematical proof (Definition1.1 and Definition1.2 of part A in Supplementary Materials). Namely, there are n attractors and n basins for n nodes data. And there is only one node in each basin. Therefore, level 1 has been built directly according to this conclusion. The attractors are denoted as $\{A_i^1\}, i=1,2,...,\mathrm{n}$, the basins are denoted as $\{B_i^1\}, i=1,2,...,\mathrm{n}$, the $\mathrm{n}\times\mathrm{n}$ Markov transition matrix is denoted as P^1 that represents the transition probability between $\{B_i^1\}, i=1,2,...,\mathrm{n}$, and the $\mathrm{n}\times\mathrm{n}$ pseudo-energy matrix is denoted as C^1 representing the pseudo-energy consumed by transition among $\{B_i^1\}, i=1,2,...,\mathrm{n}$.

On a higher level, such as level h, the attractors are denoted as $\{A_i^h\}, i=1,2,...,\mathrm{n}^h$, the basins are denoted as $\{B_i^h\}, i=1,2,...,\mathrm{n}^h$, the $\mathrm{n}^\mathrm{h}\times\mathrm{n}^\mathrm{h}$ Markov transition matrix is denoted as P^h that represents the transition probability between $\{B_i^h\}, i=1,2,...,\mathrm{n}^h$, and the $\mathrm{n}^\mathrm{h}\times\mathrm{n}^\mathrm{h}$ pseudo-energy matrix is denoted as C^h representing the energy cost by transition among $\{B_i^h\}, i=1,2,...,\mathrm{n}^h$.

The pseudo code for building the hierarchical structure:

```
1 Input the Markov transition matrix P^1 and the pseudo-energy matrix C^1
2 P^h =P^1, C^h =C^1
3 while (TRUE){
#find attractors and partition basins
4 Find all basins belong to recurrent classes in P^h and we call them RB^h below
5 Use P^h to make a directed graph. In the graph, the vertexes are the basins on the
  current level and the edges are the transition probability between basins. Set i=1.
6     while (TRUE){
```

7 Select an untraversed recurrent vertex, say RB_i^h ,as the start vertex, find all reachable vertexes in the directed graph from RB_i^h and merge the nodes in the attractors of these reachable basins to be the attractor of the basin on the level h+1, denoted as A_i^{h+1}.

8 Let RB_i^h to be the target vertex, find all vertexes in the directed graph that can reach RB_i^h and merge the nodes in these basins to be the basin on the level h+1, denoted as B_i^{h+1}. Mark all basins contained in B_i^{h+1} as traversed. i=i+1.

9 if (all B^h have been traversed) {break}
}

10 h=h+1

#Update the pseudo-energy matrix on level h+1

for(each basin pair on level h+1){

The energy consumed by basin i transiting to basin j is equal to the minimum energy consumed by the nodes in basin i's attractor transiting to the nodes in basin j. Update the values in pseudo-energy matrix C^h }

#Update the Markov transition matrix P^h on level h+1

Find the positions corresponding the values smaller than α% quantile of all values in C^h. α% is a customized parameter. P^h (positions) = 1, P^h (other positions) = 0, and convert P^h to be a Markov probability matrix by normalizing the sum of each row of it to 1.

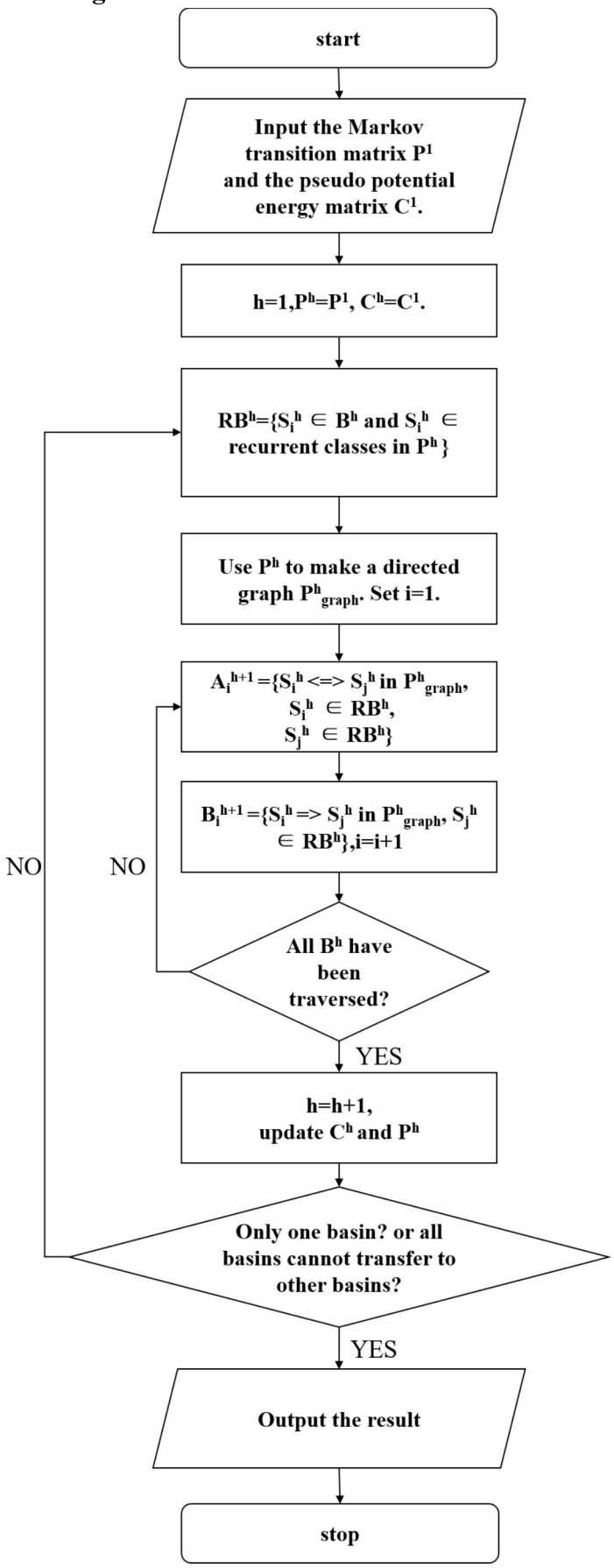
11 if(there is only one basin or all basins cannot transfer to other basins){

return the result

12 }

13 }

The workflow for building the hierarchical structure:



To build the level h+1 in the hierarchical structure, we need to use the Markov transition matrix P^h. At first, we must find all basins belong to recurrent states in the Markov transition matrix P^h. As the mathematical proof in part B of Supplementary Materials, that if a state i is a recurrent state in the Markov transition matrix, then there must be an eigenvalue 1 with a corresponding eigenvector whose i th entry is non-zero. If a state i is a transient state, then for all eigenvalues 1, the i th entry of the corresponding eigenvector must be 0. In this way, we can quickly identify all recurrent states.

The nodes in the attractors of these recurrent state basins can reach each other with positive transition probabilities are merged to be the attractor of the basin on level h+1. For example, suppose there are 10 basins on level h, and $\{B_i^h\}, i=1,2,...,5$ are recurrent states. $\{A_i^h\}, i=1,2$ have a positive transition probability path to each other and $\{A_i^h\}, i=3,4,5$ have a positive transition probability path to each other. Then, the nodes in $\{A_i^h\}, i=1,2$ are merged to be $\{A_1^{h+1}\}$ and the nodes in $\{A_i^h\}, i=3,4,5$ are merged to be $\{A_2^{h+1}\}$. The nodes in other basins, which can reach recurrent state basins with positive transition probability but recurrent state basins cannot reach them, are partitioned to the recurrent state basins. For example, $\{B_i^h\}, i=6,7,8$ have positive transition probability path to $\{B_i^h\}, i=1,2$ but $\{B_i^h\}, i=1,2$ cannot reach them, then all nodes in $\{B_i^h\}, i=6,7,8$ are merged to $\{B_i^h\}, i=1,2$ as $\{B_1^{h+1}\}$.

For implementation convenience, the Markov transition matrix P^h is converted to a directed graph. In the graph, the vertexes are the basins on level h and the edges are the transition probability between basins. Let each recurrent state to be the start vertex to find all reachable vertexes from it. The attractors in these basins are merged to be the attractor of one basin on level h+1.For example, $\{A_1^{h+1}\}$.

Next, these basins are used as target vertexes to find all other basins have a positive transition probability path to them. All nodes in these basins are merged to $\{B_1^{h+1}\}$. After all the cells have been traversed, we can partition all nodes into basins on level h+1.

To build the level h+2, we need to calculate the Markov transition matrix P^{h+1} representing the transition probability among each basin on level h+1 first. And the matrix P^{h+1} is derived from C^{h+1}. On level h+1, the pseudo-energy from $\{B_1^{h+1}\}$ to $\{B_2^{h+1}\}$ is the minimum pseudo-energy from the nodes in $\{A_1^{h+1}\}$ to the nodes in $\{B_2^{h+1}\}$. In details, we find the minimum value in the matrix C^{h+1} at first and set 1 to the

corresponding positions in the Markov transition matrix P^{h+1}. Then, we convert the updated P^{h+1} to a Markov probability matrix by normalizing the sum of each row of it to 1. After we obtain the matrix P^{h+1}. The details of building the level h+2 are same as that building level h+1 based on P^h.

There is an approximation in the Markov transition matrix P^h (h>1) calculation. Strictly according to theory, only the positions, which corresponding positions in the C^{h-1} matrix where the value is strictly equal to the minimum, is set to 1 in P^h. To accelerate basins merging, the entries whose corresponding position in the C^{h-1} matrix having a value smaller than the α% quantile of all values in C^{h-1}, are set to be 1 in P^h. The α% is a customizable parameter.

There's another optional approximation in this algorithm, which can provide considerable acceleration. On each level, we can merge some small basins to its closest qualified basin when some big basins already form. Naturally, we can finish the construction of a hierarchical structure with fewer iterations. This approximation is built on the belief that if a big basin has already existed on some level, it indicates that we've reached a macroscopically meaningful level and hence basins of small sizes can be considered as noise basins. This approximation shows great algorithm efficiency gain and leads to similar results compared with the accurate version when there's no huge density discrepancy in data. However, when there's a huge density discrepancy in data, which means that some clusters have obviously larger density over other clusters, we don't recommend this approximation. This approximation requires three customized parameters to specify the size of noise basins, qualified basins, and big basins respectively. Our empirical recommendation is as follows. We regard basins consisting of less than 10 nodes as noise basins, basins consisting of more than 10 nodes as qualified basins and basins consisting of more than 20 percentage nodes of all as big basins.

Iterating the process above until there is only one basin left or all pseudo-energy consumed by the transition among basins is equal to infinity, which indicates all basins cannot transit to others definitely, the hierarchical structure of the input data is constructed completely.

Comparison with other methods

It should be noted that there is a resolution parameter in Seurat which is critical to the result, with increased values leading to more clusters. For unknown data, users can only adjust this parameter based on experience and visualization of data. In the process of our evaluation, we fine-tuned this parameter in Seurat to cluster points into three clusters and used these results as the best results of Seurat. For comparison, we also used the default parameters of Seurat to get the default results. As for methods that needed to set the number of clusters, we set it equal to that in the ground truth so that we could get the best results of them. And the results of MarkovHC were taken from the level with the equal number of basins in the ground truth.

Level selection strategy

Selecting levels appropriately from the cluster hierarchy is important for

identifying meaningful cell basins.

For users who have enough prior knowledge about data, we recommend a top-down strategy to do it. Firstly, users can identify cell basins on high levels according to marker genes expression. Secondly, users should track the ambiguous basins from high levels to lower levels and identify the sub-basins by marker genes. Lastly, meaningful basins and the relationships among these basins can be extracted from the hierarchical structure by this strategy.

For users who do not have enough prior knowledge about their data, we suggest them to use the plot of "pseudo-energy vs level plot" to assist level selection. As the level increases, the threshold of pseudo-energy cost for basin transition increases accordingly. We divide the "pseudo-energy vs level plot" into two regimes, a noise regime and a signal regime. Low-level results could be swamped by a lot of noises, so the pseudo-energy change slowly in the noise regime. But the pseudo-energy dramatically changes in the signal regime for basins on the high-level may correspond to big cell clusters with biological significance. So levels around the boundary of the noise regime and the signal regime may have biological meaning. We show the "pseudo-energy vs level plots" of the simulation data in Fig. 2F (Fig. S21A) and four real data sets (Fig. S21B. data in Fig. 4; C. data in Fig. 5; D. data in Fig. 6; E. data in Fig. S20) used in this paper.

Find the transition paths and critical points

With the help of Markov chain's directivity, we can determine the transition path, which is the most probable path supported by large deviations theory[89]. The transition path from basin A to basin B constituted by points on the shortest path from A's attractors to B. And we employed Dijkstra algorithm to find the shortest path in implement. After the transition path found, the two closest points between basin A and B on the path are determined as critical points from A to B.

Differentially expressed genes and GO terms

Wilcoxon rank sum test in Seurat was employed to detect differentially expressed genes among groups of cells. Differentially expressed genes along the transition path were identified by differentialGeneTest function in monocle[20]. And GO terms of genes sets were enriched by enrichGO function in clusterProfiler[90].

Data availability

R package "MarkovHC" are available at GitHub (https://github.com/ZhenyiWangTHU/MarkovHC).

Supplementary data

The links for downloading datasets used in this paper are in the part.F of the

Supplementary Materials.

Author contributions

Conceptualization, M.P.Q., Z.Y.W. and Y.J.Z.;
Methodology, Z.Y.W. and Y.J.Z.;
Software, Z.Y.W.;
Formal Analysis, Z.Y.W.;
Investigation, Z.Y.W., Z.F.Y., L.Z. and M.L.S.;
Resources, M.Q.Z.;
Data Curation, Z.Y.W.;
Writing – Original Draft, Z.Y.W.;
Writing – Review & Editing, M.Q.Z., Z.F.Y., Z.Y.W., Y.J.Z., L.Z., Y.C., M.L.S. and M.P.Q.;
Visualization, Z.F.Y. and Z.Y.W.;
Supervision, M.Q.Z. and M.P.Q.;
Project Administration, M.Q.Z.;
Funding Acquisition, M.Q.Z.

Acknowledgement

We acknowledge Peng Zhang, Yuyao Song and Kui Hua of Tsinghua University for helpful discussion and suggestion.

References

1 Moris, N., Pina, C. & Arias, A. M. Transition states and cell fate decisions in epigenetic landscapes. Nat Rev Genet 17, 693-703, doi:10.1038/nrg.2016.98 (2016).
2 Chen, Z. et al. DensityPath: an algorithm to visualize and reconstruct cell state-transition path on density landscape for single-cell RNA sequencing data. Bioinformatics 35, 2593-2601, doi:10.1093/bioinformatics/bty1009 (2019).
3 Qian, H. Cycle kinetics, steady state thermodynamics and motors-a paradigm for living matter physics. J Phys Condens Matter 17, S3783-3794, doi:10.1088/0953-8984/17/47/010 (2005).
4 Chen, D., Feng, J. & Qian, M. Metastability of exponentially perturbed Markov chains. Science in China Series A-Mathematics 39, 7-28 (1996).
5 Wasserman, L. Topological data analysis. Annual Review of Statistics and Its Application 5, 501-532 (2018).
6 Chu, L. F. et al. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Genome Biol 17, 173, doi:10.1186/s13059-016-1033-x (2016).

7 Petropoulos, S. et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell 165, 1012-1026, doi:10.1016/j.cell.2016.03.023 (2016).
8 Zhang, P. et al. Dissecting the Single-Cell Transcriptome Network Underlying Gastric Premalignant Lesions and Early Gastric Cancer. Cell Rep 27, 1934-1947 e1935, doi:10.1016/j.celrep.2019.04.052 (2019).
9 Huang, S., Eichler, G., Bar-Yam, Y. & Ingber, D. E. Cell fates as high-dimensional attractor states of a complex gene regulatory network. Phys Rev Lett 94, 128701, doi:10.1103/PhysRevLett.94.128701 (2005).
10 Wright, S. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. Vol. 1 (1932).
11 Waddington, C. H. Principles of development and differentiation. (1966).
12 Waddington, C H. The strategy of the genes: A discussion of some aspects of theoretical biology. (1957).
13 Ao, P. Global view of bionetwork dynamics: adaptive landscape. Journal of Genetics and Genomics 36, 63-73, doi:10.1016/s1673-8527(08)60093-4 (2009).
14 Schiebinger, G. et al. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. Cell 176, 928-943 e922, doi:10.1016/j.cell.2019.01.006 (2019).
15 Ao, P., Galas, D., Hood, L. & Zhu, X. Cancer as robust intrinsic state of endogenous molecular-cellular network shaped by evolution. Med Hypotheses 70, 678-684, doi:10.1016/j.mehy.2007.03.043 (2008).
16 Ao, P. Laws in Darwinian evolutionary theory. Physics of Life Reviews 2, 117-156, doi:10.1016/j.plrev.2005.03.002 (2005).
17 Li, C. & Wang, J. Quantifying cell fate decisions for differentiation and reprogramming of a human stem cell network: landscape and biological paths. PLoS Comput Biol 9, e1003165, doi:10.1371/journal.pcbi.1003165 (2013).
18 Tan, Z. W., Guarnera, E. & Berezovsky, I. N. Exploring chromatin hierarchical organization via Markov State Modelling. PLoS Comput Biol 14, e1006686, doi:10.1371/journal.pcbi.1006686 (2018).
19 Qiu, P. et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol 29, 886-891, doi:10.1038/nbt.1991 (2011).
20 Trapnell, C. et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 32, 381-386, doi:10.1038/nbt.2859 (2014).
21 La Manno, G. et al. RNA velocity of single cells. Nature 560, 494-498, doi:10.1038/s41586-018-0414-6 (2018).
22 Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. bioRxiv, 820936 (2019).
23 Kiselev, V. Y. et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods 14, 483-486, doi:10.1038/nmeth.4236 (2017).

24 Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand-receptor complexes. Nat Protoc 15, 1484-1506, doi:10.1038/s41596-020-0292-x (2020).
25 Wasserman, L. Topological Data Analysis. (2016).
26 van Dijk, D. et al. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell 174, 716-729 e727, doi:10.1016/j.cell.2018.05.061 (2018).
27 Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. Bioinformatics 31, 2989-2998, doi:10.1093/bioinformatics/btv325 (2015).
28 Zappia, L. & Oshlack, A. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. Gigascience 7, doi:10.1093/gigascience/giy083 (2018).
29 Schwartz, G. W. et al. TooManyCells identifies and visualizes relationships of single-cell clades. Nat Methods 17, 405-413, doi:10.1038/s41592-020-0748-5 (2020).
30 Alcocer-Cuaron, C., Rivera, A. L. & Castano, V. M. Hierarchical structure of biological systems: a bioengineering approach. Bioengineered 5, 73-79, doi:10.4161/bioe.26570 (2014).
31 Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. in Int. Conf. Knowledge Discovery and Data Mining. 6.
32 Mihael Ankerst, M. M. B., Hans-Peter Kriegel, Jörg Sander. OPTICS: Ordering Points To Identify the Clustering Structure. (1999).
33 Campello, R. J. G. B., Moulavi, D., Zimek, A. & Sander, J. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. ACM Transactions on Knowledge Discovery from Data 10, 1-51, doi:10.1145/2733381 (2015).
34 Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 36, 411-420, doi:10.1038/nbt.4096 (2018).
35 Sibson, R. SLINK: an optimally efficient algorithm for the single-link cluster method. The computer journal 16, 30-34 (1973).
36 Sokal, R. R. A statistical method for evaluating systematic relationships. Univ. Kansas, Sci. Bull. 38, 1409-1438 (1958).
37 Defays, D. An efficient algorithm for a complete link method. The Computer Journal 20, 364-366 (1977).
38 Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. & Batzoglou, S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods 14, 414-416, doi:10.1038/nmeth.4207 (2017).
39 Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008, P10008, doi:10.1088/1742-5468/2008/10/p10008 (2008).

40 Rodriguez, M. Z. et al. Clustering algorithms: A comparative approach. PLoS One 14, e0210236, doi:10.1371/journal.pone.0210236 (2019).
41 Kiselev, V. Y., Andrews, T. S. & Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet 20, 273-282, doi:10.1038/s41576-018-0088-9 (2019).
42 Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol 19, 15, doi:10.1186/s13059-017-1382-0 (2018).
43 Lin, P., Troup, M. & Ho, J. W. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol 18, 59, doi:10.1186/s13059-017-1188-0 (2017).
44 Zurauskiene, J. & Yau, C. pcaReduce: hierarchical clustering of single cell transcriptional profiles. BMC Bioinformatics 17, 140, doi:10.1186/s12859-016-0984-y (2016).
45 Chen, J., Schlitzer, A., Chakarov, S., Ginhoux, F. & Poidinger, M. Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. Nat Commun 7, 11988, doi:10.1038/ncomms11988 (2016).
46 Guo, M., Wang, H., Potter, S. S., Whitsett, J. A. & Xu, Y. SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. PLoS Comput Biol 11, e1004575, doi:10.1371/journal.pcbi.1004575 (2015).
47 Ertöz, L., Steinbach, M. & Kumar, V. in Proceedings of the 2003 SIAM international conference on data mining. 47-58 (SIAM).
48 Bhargav, S. & Pawar, M. A review of clustering methods forming non-convex clusters with missing and noisy data. IJCSE, 39-44 (2016).
49 Hess, S., Duivesteijn, W., Honysz, P. & Morik, K. in Proceedings of the AAAI Conference on Artificial Intelligence. 3788-3795.
50 Van Dongen, S. M. Graph clustering by flow simulation, (2000).
51 Qiu, X. et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods 14, 979-982, doi:10.1038/nmeth.4402 (2017).
52 Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature 566, 496-502, doi:10.1038/s41586-019-0969-x (2019).
53 Farrell, J. A. et al. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science 360, doi:10.1126/science.aar3131 (2018).
54 Bendall, S. C. et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell 157, 714-725, doi:10.1016/j.cell.2014.04.005 (2014).
55 Levine, J. H. et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell 162, 184-197, doi:10.1016/j.cell.2015.05.047 (2015).
56 Shekhar, K. et al. Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. Cell 166, 1308-1323 e1330, doi:10.1016/j.cell.2016.07.054 (2016).

57 Setty, M. et al. Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat Biotechnol 34, 637-645, doi:10.1038/nbt.3569 (2016).
58 Haghverdi, L., Buttner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods 13, 845-848, doi:10.1038/nmeth.3971 (2016).
59 Huang, S., Li, F., Zhou, J. X. & Qian, H. Processes on the emergent landscapes of biochemical reaction networks and heterogeneous cell population dynamics: differentiation in living matters. J R Soc Interface 14, doi:10.1098/rsif.2017.0097 (2017).
60 Dijkstra, E. W. A note on two problems in connexion with graphs. Numerische mathematik 1, 269-271 (1959).
61 Jia, C., Qian, M., Kang, Y. & Jiang, D. Modeling stochastic phenotype switching and bet-hedging in bacteria: stochastic nonlinear dynamics and critical state identification. Quantitative Biology 2, 110-125, doi:10.1007/s40484-014-0035-5 (2015).
62 Mojtahedi, M. et al. Cell Fate Decision as High-Dimensional Critical State Transition. PLoS Biol 14, e2000640, doi:10.1371/journal.pbio.2000640 (2016).
63 Anchang, B. et al. Visualization and cellular hierarchy inference of single-cell data using SPADE. Nat Protoc 11, 1264-1279, doi:10.1038/nprot.2016.066 (2016).
64 Andrews, T. S. & Hemberg, M. Identifying cell populations with scRNASeq. Mol Aspects Med 59, 114-122, doi:10.1016/j.mam.2017.07.002 (2018).
65 Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. in Kdd. 226-231.
66 Jeitziner, R. et al. Two-Tier Mapper, an unbiased topology-based clustering method for enhanced global gene expression analysis. Bioinformatics 35, 3339-3347, doi:10.1093/bioinformatics/btz052 (2019).
67 Wolf, F. A. et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol 20, 59, doi:10.1186/s13059-019-1663-x (2019).
68 Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol 18, 174, doi:10.1186/s13059-017-1305-0 (2017).
69 Packer, J. S. et al. A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. Science, doi:10.1126/science.aax1971 (2019).
70 Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. Nat Rev Immunol 18, 35-45, doi:10.1038/nri.2017.76 (2018).
71 Villani, A. C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science 356, doi:10.1126/science.aah4573 (2017).

72 Bendall, S. C. et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science 332, 687-696, doi:10.1126/science.1198704 (2011).
73 Zeng, W. et al. DC3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data. Nat Commun 10, 4613, doi:10.1038/s41467-019-12547-1 (2019).
74 Regev, A. et al. The Human Cell Atlas. Elife 6, doi:10.7554/eLife.27041 (2017).
75 Moon, K. R. et al. Visualizing Structure and Transitions for Biological Data Exploration. doi:10.1101/120378 (2018).
76 Cheng, J. & Fan, X. M. Role of cyclooxygenase-2 in gastric cancer development and progression. World J Gastroenterol 19, 7361-7368, doi:10.3748/wjg.v19.i42.7361 (2013).
77 Boussioutas, A. et al. Distinctive patterns of gene expression in premalignant gastric mucosa and gastric cancer. Cancer Res 63, 2569-2577 (2003).
78 Alzahrani, S. et al. Effect of Helicobacter pylori on gastric epithelial cells. World J Gastroenterol 20, 12767-12780, doi:10.3748/wjg.v20.i36.12767 (2014).
79 Yuan, W., Yang, N. & Li, X. Advances in Understanding How Heavy Metal Pollution Triggers Gastric Cancer. Biomed Res Int 2016, 7825432, doi:10.1155/2016/7825432 (2016).
80 van der Flier, L. G., Haegebarth, A., Stange, D. E., van de Wetering, M. & Clevers, H. OLFM4 Is a Robust Marker for Stem Cells in Human Intestine and Marks a Subset of Colorectal Cancer Cells. Gastroenterology 137, 15-17, doi:10.1053/j.gastro.2009.05.035 (2009).
81 Wong, C.-M. et al. CEACAM6 Promotes Gastric Cancer Invasion and Metastasis by Inducing Epithelial-Mesenchymal Transition via PI3K/AKT Signaling Pathway. PLoS ONE 9, e112908, doi:10.1371/journal.pone.0112908 (2014).
82 Busuttil, R. A. et al. Role of p53 in the progression of gastric cancer. Oncotarget 5, 12016-12026, doi:10.18632/oncotarget.2434 (2014).
83 Jiaqing, L. et al. Role of cyclin E and p53 expression in progression of early gastric cancer. Gastric Cancer 1, 160-165, doi:10.1007/s101200050011 (1998).
84 Chen, Y. et al. MiR-596 down regulates SOX4 expression and is a potential novel biomarker for gastric cancer. Translational Cancer Research 9, 1294-1302, doi:10.21037/tcr.2020.02.02 (2020).
85 Wang, C. L., Wang, D., Yan, B. Z., Fu, J. W. & Qin, L. Long non-coding RNA NEAT1 promotes viability and migration of gastric cancer cell lines through up-regulation of microRNA-17. Eur Rev Med Pharmacol Sci 22, 4128-4137, doi:10.26355/eurrev_201807_15405 (2018).
86 Ji, Z. & Ji, H. Pseudotime Reconstruction Using TSCAN. Methods Mol Biol 1935, 115-124, doi:10.1007/978-1-4939-9057-3_8 (2019).

87 Cao, K., Bai, X., Hong, Y. & Wan, L. Unsupervised Topological Alignment for Single-Cell Multi-Omics Integration. doi:10.1101/2020.02.02.931394 (2020).

88 Kang, Z., Peng, C. & Cheng, Q. Kernel-driven similarity learning. Neurocomputing 267, 210-219, doi:10.1016/j.neucom.2017.06.005 (2017).

89 Varadhan, S. R. Asymptotic Probabilities and Differential Equations. Commun Pur Appl Math 19, 261-&, doi:DOI 10.1002/cpa.3160190303 (1966).

90 Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16, 284-287, doi:10.1089/omi.2011.0118 (2012).

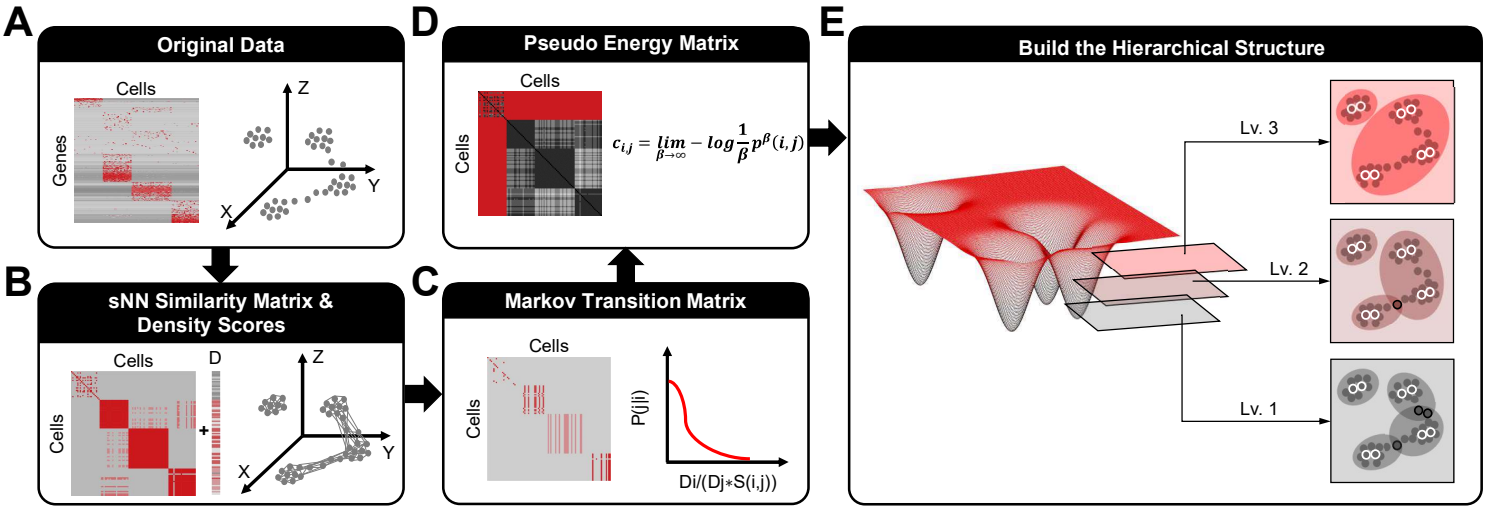

A
Original Data
Cells
Genes
Z
Y
X
B
sNN Similarity Matrix &
Density Scores
Cells
Cells
D
Z
Y
X
C
Markov Transition Matrix
Cells
Cells
P(j|i)
Di/(Dj*S(i,j))
D
Pseudo Energy Matrix
Cells
Cells
$c_{i,j} = \lim_{\beta \to \infty} -\log \frac{1}{\beta} p^{\beta}(i,j)$
E
Build the Hierarchical Structure
Lv. 3
Lv. 2
Lv. 1

Figure 1. Overview of MarkovHC.

(A) The original input data is the matrix of genes by cells.

(B) We calculate sNN (shared Nearest Neighbors) among cells to get the cell by cell similarity matrix. Then we construct a cellular network using the similarity matrix and calculate each cell's degree (D score) in the network.

(C) The Markov transition matrix is calculated using the similarity matrix and D scores.

(D) The pseudo-energy matrix is calculated based on the Markov transition matrix. (see Methods)

(E) The hierarchical structure is constructed based on attractors, basins, and critical points on each level.

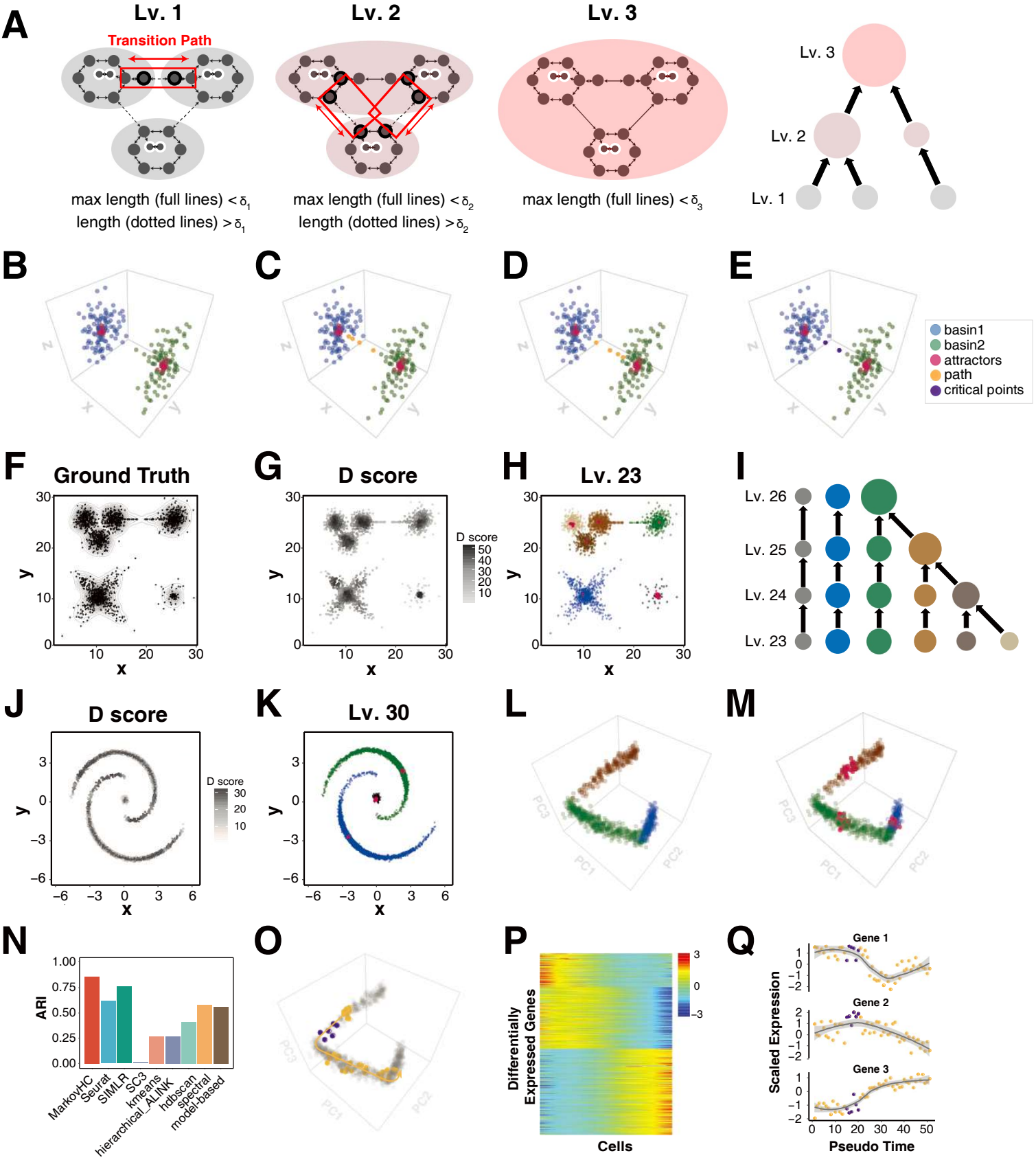

A
Lv. 1
Transition Path
Lv. 2
Lv. 3
max length (full lines) < δ_1
length (dotted lines) > δ_1
max length (full lines) < δ_2
length (dotted lines) > δ_2
max length (full lines) < δ_3
Lv. 3
Lv. 2
Lv. 1
B
C
D
E
basin1
basin2
attractors
path
critical points
F
Ground Truth
G
D score
D score
50
40
30
20
10
H
Lv. 23
I
Lv. 26
Lv. 25
Lv. 24
Lv. 23
J
D score
D score
30
20
10
K
Lv. 30
L
M
N
ARI
1.00
0.75
0.50
0.25
0.00
MarkovHC
Seurat
SIMLR
SC3
kmeans
hierarchical_ALINK
hdbscan
spectral
model-based
O
P
Differentially Expressed Genes
Cells
Q
Gene 1
Gene 2
Gene 3
Scaled Expression
Pseudo Time

Figure 2. MarkovHC is able to cluster non-convex data, hetero-density data, and differentiation stages in a continuum.

(A)The points that can be connected using lines (pseudo-energy) no longer than δn are partitioned into a basin on level n. In a basin, points in the highest density regions constitute attractors (points with white border). Points connecting basins' attractors constitute the transition path (red box), and the closest points (black border) between these two basins on the transition path are the critical points.

(B-E)These two group points are clustered into two basins (blue and green) by MarkovHC (B). To insight more details of these two basins, we highlighted the attractors (red), the transition paths (yellow) and critical points (purple) respectively. The points linking basin A's attractors and basin B's border constitute the transition path from A to B and vice versa (C, D). The transition path between A and B are the points linking their attractors, and the critical points which indicate the points jump at the boundary of these two basins are the two closest points belong to them on the transition path (E).

(F) The 2-dimensional simulated data with density contours.

(G) D score of each point.

(H) The basins and attractors on level 23 were plotted to show the 2-dimensional layouts. And we observed that MarkovHC can simultaneously detect non-convex basins and hetero-density basins.

(I) The hierarchical structure of these data. On this hierarchical structure, the sizes of basins represent the number of samples and the colors represent different basins.

(J) D score of each point.

(K) Blue, green and black points indicate three basins. And red points show the attractors of each basin.

(L) Simulated single-cell data in three stages (generated by splatter[68]). These data included 1000 cells belonging to three stages in the 5000 genes space and these cells were projected into principle component space.

(M) MarkovHC detected 3 basins on level 17 (brown, green, and blue) and the red points in each basin indicated the attractors.

(N) ARI (adjust rand index) was used to evaluate the clustering accuracy of MarkovHC, Seurat, SIMLR, sc3, k-means, hierarchical clustering with average linkage, HDBSCAN, spectral clustering, and model-based, which showed that MarkovHC performed best on these data.

(O) The orange points indicated the points on the transition path (orange line) from the brown basin to the blue basin. The purple points indicated the critical points.

(P) The heatmap of differentially expressed genes along the transition path in (O).

(Q) Three representative genes with different "gene-flow" trends in (P). The purple points indicated the critical points.

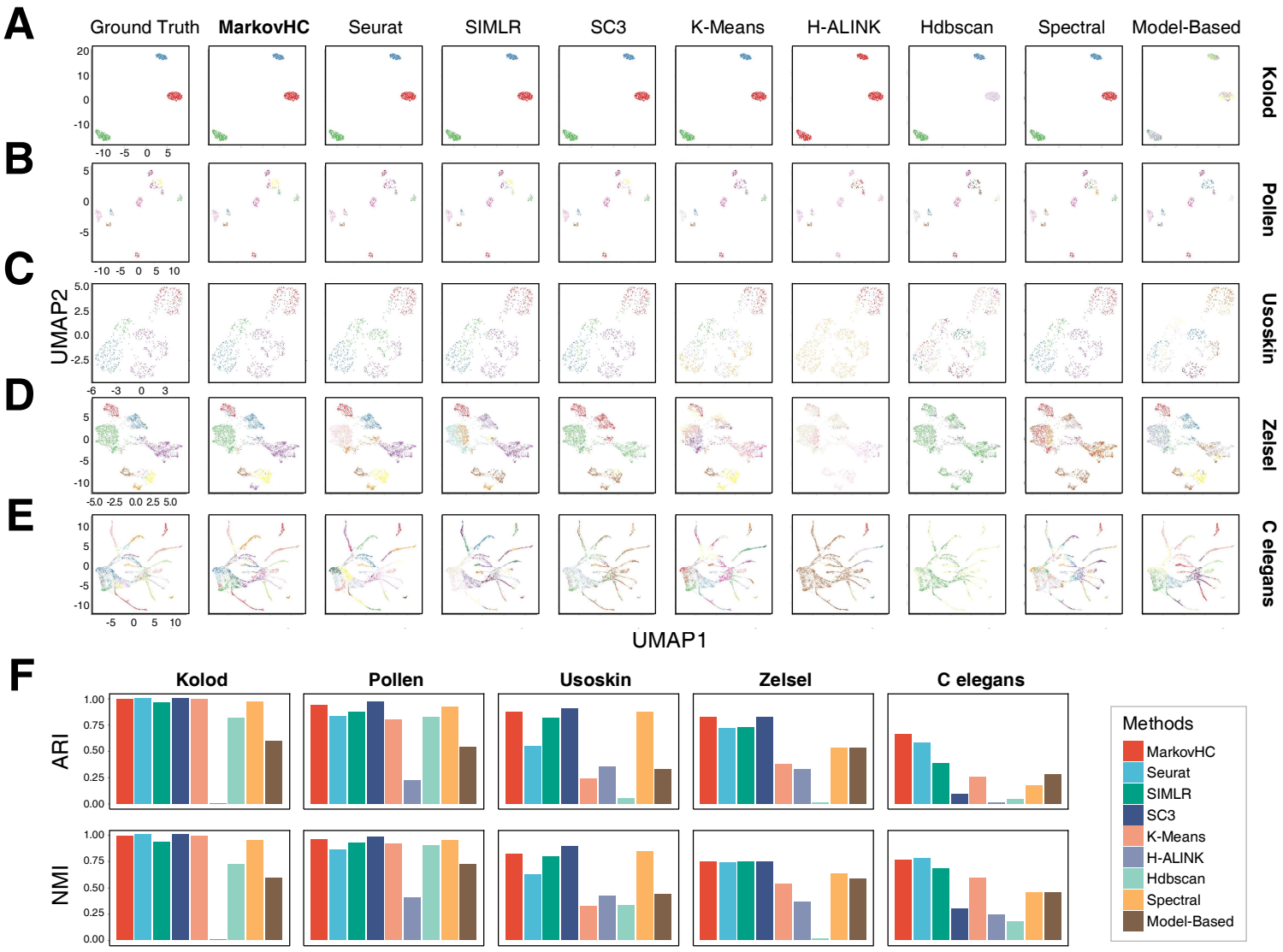

A
B
C
D
E
F
Ground Truth
MarkovHC
Seurat
SIMLR
SC3
K-Means
H-ALINK
Hdbscan
Spectral
Model-Based
Kolod
Pollen
Usoskin
Zelsel
C elegans
UMAP2
UMAP1
ARI
NMI
Methods
MarkovHC
Seurat
SIMLR
SC3
K-Means
H-ALINK
Hdbscan
Spectral
Model-Based

Figure3. MarkovHC can consistently achieve accuracy in clustering diverse types of single-cell omics data.

(A-E) Ground truths of five benchmark single-cell RNA-Seq data sets[38,69] and clustering results of MarkovHC, Seurat, SIMLR, SC3, k-means, hierarchical clustering with average linkage, HDBSCAN, spectral clustering, and EM.

(F) ARI (Rand Adjust Index) and NMI (Normalized Mutual Information) showed MarkovHC can consistently achieve top accuracy in clustering real data.

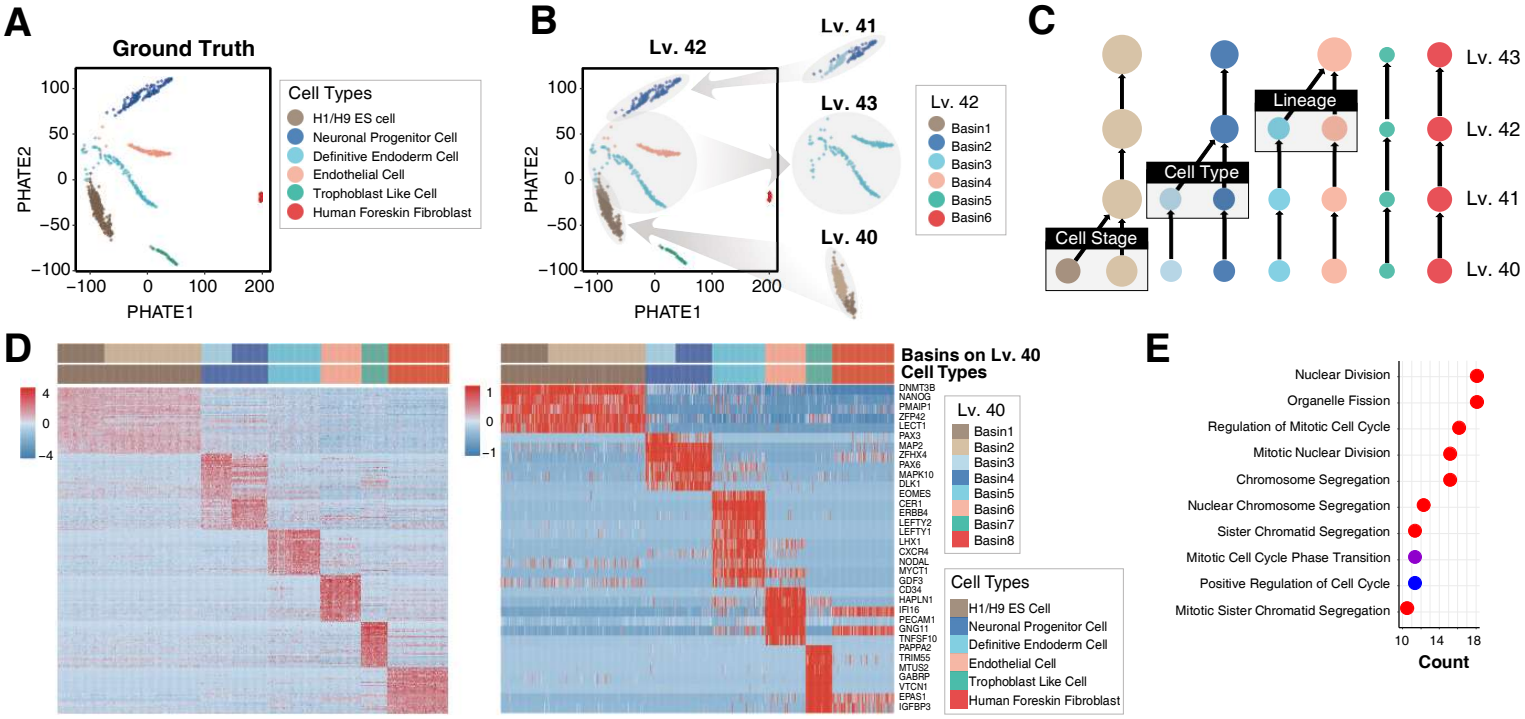

A
Ground Truth
PHATE2
PHATE1
Cell Types
H1/H9 ES cell
Neuronal Progenitor Cell
Definitive Endoderm Cell
Endothelial Cell
Trophoblast Like Cell
Human Foreskin Fibroblast
B
Lv. 42
Lv. 41
Lv. 43
Lv. 40
Lv. 42
Basin1
Basin2
Basin3
Basin4
Basin5
Basin6
C
Lineage
Cell Type
Cell Stage
Lv. 43
Lv. 42
Lv. 41
Lv. 40
D
Basins on Lv. 40
Cell Types
Lv. 40
Basin1
Basin2
Basin3
Basin4
Basin5
Basin6
Basin7
Basin8
Cell Types
H1/H9 ES Cell
Neuronal Progenitor Cell
Definitive Endoderm Cell
Endothelial Cell
Trophoblast Like Cell
Human Foreskin Fibroblast
E
Nuclear Division
Organelle Fission
Regulation of Mitotic Cell Cycle
Mitotic Nuclear Division
Chromosome Segregation
Nuclear Chromosome Segregation
Sister Chromatid Segregation
Mitotic Cell Cycle Phase Transition
Positive Regulation of Cell Cycle
Mitotic Sister Chromatid Segregation
Count

Figure 4. MarkovHC's hierarchical structure revealed cell lineages, cell types, and cell stages simultaneously.

(A) Single-cell RNA-Seq data[6] of human ES cell-derived lineage-specific progenitors were projected into 2-dimensional space by phateR. And seven cell types are identified using marker genes.

(B) On level 40, two different cell stages were found in H1 and H9 ES cells. On level 41, two cell types in the neuronal progenitor cells were found. On level 42, cells of different lineages were separated.

(C) On the hierarchical structure, the difference among basins on level 42 might correspond to different cell lineages, while that among basins on level 41 and 40 might correspond to cell types and cell stages

(D) The expression pattern of the 50 most significantly up-regulated genes in each basin (left) and 34 selected marker genes (right).

(E) The differentially expressed genes between basin1 and basin2 on level 40 in H1 and H9 ES cells were enriched in cell division related GO terms, which suggested that these two cell stages relate to the cell cycle.

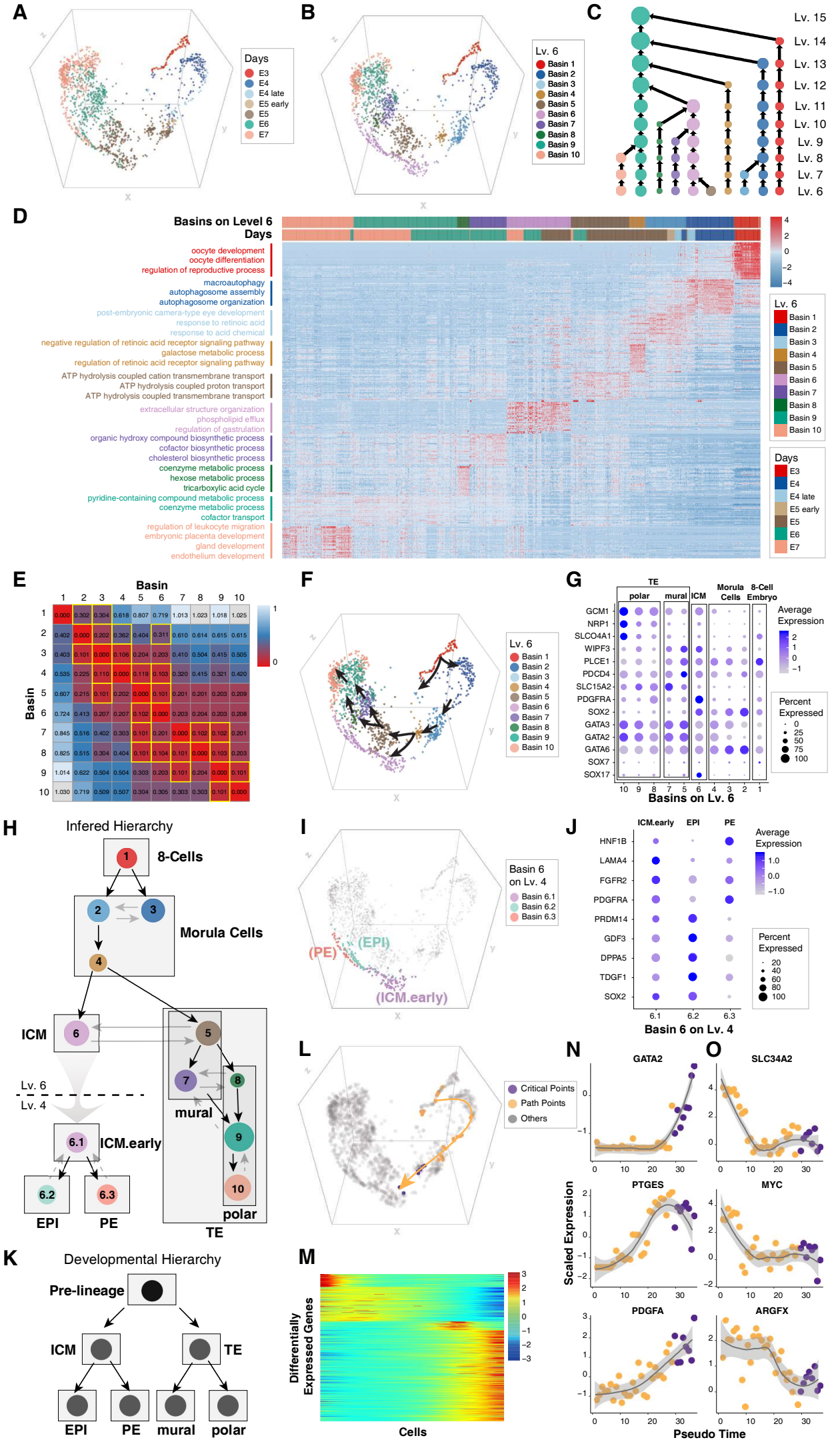

A
Days
E3
E4
E4 late
E5 early
E5
E6
E7
B
Lv. 6
Basin 1
Basin 2
Basin 3
Basin 4
Basin 5
Basin 6
Basin 7
Basin 8
Basin 9
Basin 10
C
Lv. 15
Lv. 14
Lv. 13
Lv. 12
Lv. 11
Lv. 10
Lv. 9
Lv. 8
Lv. 7
Lv. 6
D
Basins on Level 6
Days
oocyte development
oocyte differentiation
regulation of reproductive process
macroautophagy
autophagosome assembly
autophagosome organization
post-embryonic camera-type eye development
response to retinoic acid
response to acid chemical
negative regulation of retinoic acid receptor signaling pathway
galactose metabolic process
regulation of retinoic acid receptor signaling pathway
ATP hydrolysis coupled cation transmembrane transport
ATP hydrolysis coupled proton transport
ATP hydrolysis coupled transmembrane transport
extracellular structure organization
phospholipid efflux
regulation of gastrulation
organic hydroxy compound biosynthetic process
cofactor biosynthetic process
cholesterol biosynthetic process
coenzyme metabolic process
hexose metabolic process
tricarboxylic acid cycle
pyridine-containing compound metabolic process
coenzyme metabolic process
cofactor transport
regulation of leukocyte migration
embryonic placenta development
gland development
endothelium development
E
Basin
F
G
TE
polar
mural
ICM
Morula Cells
8-Cell Embryo
GCM1
NRP1
SLCO4A1
WIPF3
PLCE1
PDCD4
SLC15A2
PDGFRA
SOX2
GATA3
GATA2
GATA6
SOX7
SOX17
Average Expression
Percent Expressed
Basins on Lv. 6
H
Infered Hierarchy
8-Cells
Morula Cells
ICM
Lv. 6
Lv. 4
ICM.early
mural
polar
EPI
PE
TE
I
Basin 6 on Lv. 4
Basin 6.1
Basin 6.2
Basin 6.3
(PE)
(EPI)
(ICM.early)
J
ICM.early
EPI
PE
HNF1B
LAMA4
FGFR2
PDGFRA
PRDM14
GDF3
DPPA5
TDGF1
SOX2
Basin 6 on Lv. 4
L
Critical Points
Path Points
Others
N
GATA2
O
SLC34A2
PTGES
MYC
PDGFA
ARGFX
Scaled Expression
Pseudo Time
K
Developmental Hierarchy
Pre-lineage
ICM
TE
EPI
PE
mural
polar
M
Differentially Expressed Genes
Cells

Figure 5. MarkovHC revealed transition paths and critical points in human preimplantation embryo development.

(A) The scRNA-Seq data[7] of 1529 human preimplantation embryos cells from E3 stage to E7 stage were projected into 2-dimensional space by phateR. In the data, E3–E7 indicates the embryonic day. E4.late and E5.early indicate cells picked 4–6 hours later and earlier than that in E4 stage and E5 stage, respectively.

(B) All cells were separated into ten basins on level 6 corresponding to ten asynchronous development stages in human preimplantation embryos cells.

(C) The hierarchical structure of these cells.

(D) Enriched GO terms of basin-specific gene sets suggest the main biological processes in each development stage.

(E) Pseudo-energy matrix (yellow boxes indicate low energy cost) could be used to infer transition among basins step by step.

(F) Three trajectories (black arrows) were inferred by the pseudo-energy matrix in (E).

(G) Four main cell types which are 8-cell embryo, morula cell, ICM (inner cell mass), and TE (trophectoderm) were identified according to marker genes expression.

(H) The development hierarchy was built according to the inferred trajectories in (F). The black arrows indicate development directions, the grey arrows indicate transition relations, and the grey dashed arrows indicate reversed links.

(I) ICM (inner cell mass) was split into three basins corresponding to PE (primitive endoderm), EPI (epiblast), and ICM.early (the early stage of inner cell mass) on level 4.

(J) Marker genes were used to identify the three basins in (I).

(K) The ground truth of the development hierarchy.

(L) The transition path (yellow line with the arrow) from 8 cell embryo to ICM were tracked. The yellow points indicate cells along the path and purple points indicate the critical points from morula cells to ICM.

(M) Differentially expressed genes along the transition path in (L).

(N) Representative genes show increasing “gene-flow” trends along the path. And gene expression varies dramatically around critical points (purple points).

(O) Representative genes show decreasing “gene-flow” trends along the path. And gene expression varies dramatically around critical points (purple points).

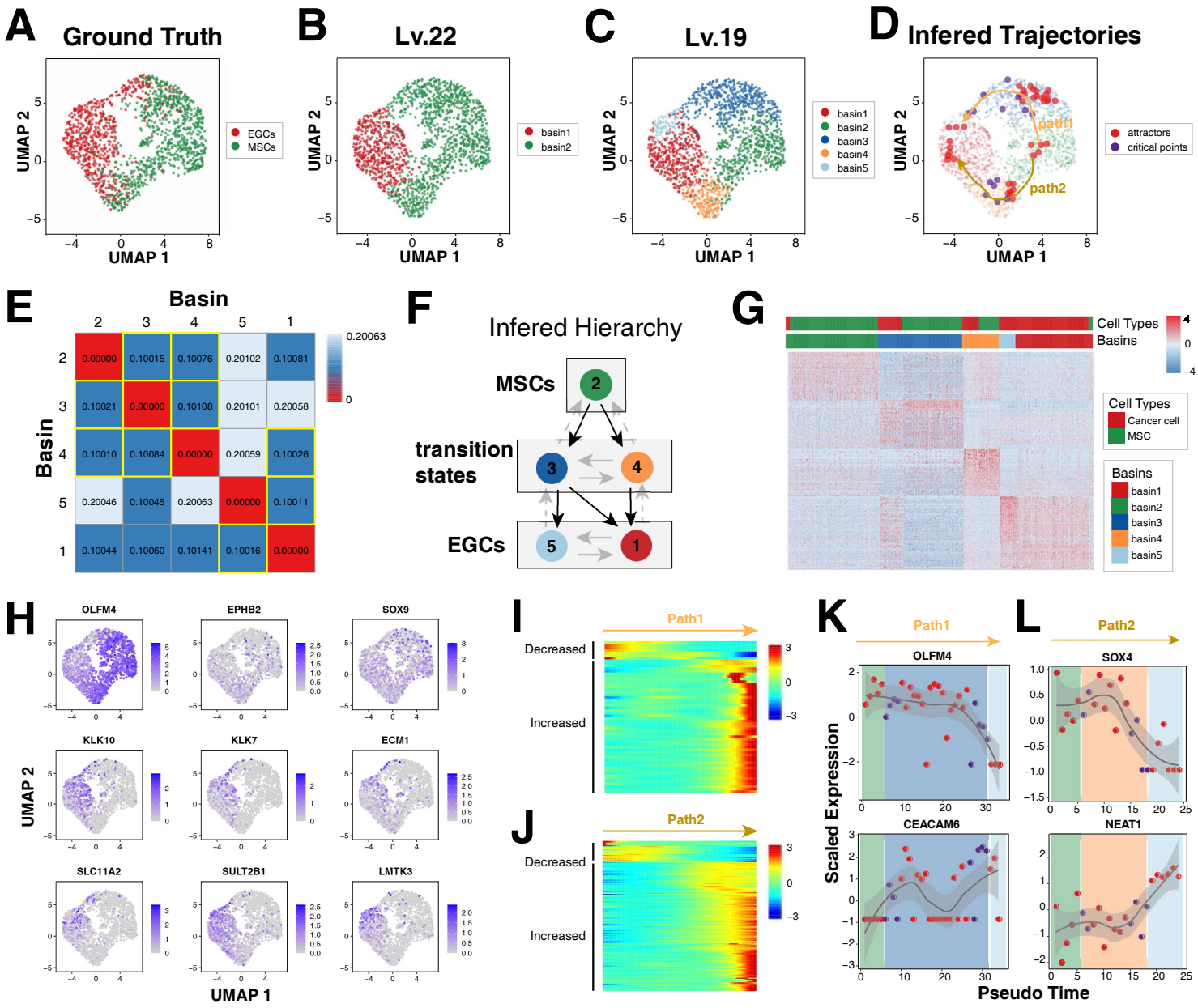

A
Ground Truth
UMAP 2
UMAP 1
EGCs
MSCs
B
Lv.22
basin1
basin2
C
Lv.19
basin1
basin2
basin3
basin4
basin5
D
Infered Trajectories
path1
path2
attractors
critical points
E
Basin
2 3 4 5 1
2 0.00000 0.10015 0.10076 0.20102 0.10081
3 0.10021 0.00000 0.10108 0.20101 0.20058
4 0.10010 0.10084 0.00000 0.20059 0.10026
5 0.20046 0.10045 0.20063 0.00000 0.10011
1 0.10044 0.10060 0.10141 0.10016 0.00000
0.20063
0
F
Infered Hierarchy
MSCs
transition states
EGCs
G
Cell Types
Basins
Cancer cell
MSC
basin1
basin2
basin3
basin4
basin5
H
OLFM4
EPHB2
SOX9
KLK10
KLK7
ECM1
SLC11A2
SULT2B1
LMTK3
UMAP 2
UMAP 1
I
Path1
Decreased
Increased
J
Path2
K
Path1
OLFM4
CEACAM6
Scaled Expression
L
Path2
SOX4
NEAT1
Pseudo Time

Figure 6. MarkovHC detected critical points from MSCs to gastric cancer cells.

(A) 830 mesenchymal stem cells (MSCs) and 649 MSC-origin early gastric cancer cells (EGCs) were projected into 2-dimensinal space using UMAP.

(B) MarkovHC found two basins on level 22.

(C) The basin corresponding to MSCs and EGCs split into three basins and two basins respectively on level 19.

(D) Two potential separating trajectories from MSCs to EGCs were inferred. Red points indicate attractors of each basin and purple points indicate critical points on boundaries of neighboring basins.

(E) Pseudo-energy matrix (yellow boxes indicate low energy cost) could be used to infer transition among basins step by step.

(F) The hierarchy was inferred according to the pseudo-energy matrix in (E). The black arrows indicate differentiation directions, the grey arrows indicate transition relations, and the grey dashed arrows indicate reversed links.

(G) The heatmap of basin-specific gene sets.

(H) MSC markers (OLFM4, EPHB2, and SOX9) and EGC markers (KLK10, SLC11A2, SULT2B1, KLK7, ECM1 and LMTK3) were gradually decreasing and increasing respectively along the trajectories from MSCs to EGCs.

(I) Differentially expressed genes along path 1 in (D).

(J) Differentially expressed genes along path 2 in (D).

(K-L) OLFM4 and CEACAM6 shown different "gene-flow" trends along path 1(K), and SOX4 and NEAT1 shown different "gene-flow" trends along path 2. Red points are attractors, purple points are critical points, and colors of background panels indicate basins on level 19 (C). Meanwhile, expression values of these genes dramatically changed around the critical points.