46–56 Nucleic Acids Research, 2022, Vol. 50, No. 1 *Published online 1 December 2021*
https://doi.org/10.1093/nar/gkab1132

MarkovHC: Markov hierarchical clustering for the topological structure of high-dimensional single-cell omics data with transition pathway and critical point detection

Zhenyi Wang[1], Yanjie Zhong[3,5], Zhaofeng Ye[2], Lang Zeng[6], Yang Chen[1], Minglei Shi[2], Zhiyuan Yuan[1], Qiming Zhou[7], 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 and [7]MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, BNRist; School of Life Sciences, Tsinghua University, Beijing 100084, China

Received February 23, 2021; Revised October 19, 2021; Editorial Decision October 25, 2021; Accepted October 27, 2021

ABSTRACT

Clustering cells and depicting the lineage relationship among cell subpopulations are fundamental tasks in single-cell omics studies. However, existing analytical methods face challenges in stratifying cells, tracking cellular trajectories, and identifying critical points of cell transitions. To overcome these, we proposed a novel Markov hierarchical clustering algorithm (MarkovHC), a topological clustering method that leverages the metastability of exponentially perturbed Markov chains for systematically reconstructing the cellular landscape. Briefly, MarkovHC starts with local connectivity and density derived from the input and outputs a hierarchical structure for the data. We firstly benchmarked MarkovHC on five simulated datasets and ten public single-cell datasets with known labels. Then, we used MarkovHC to investigate the multi-level architectures and transition processes during human embryo preimplantation development and gastric cancer procession. MarkovHC found heterogeneous cell states and sub-cell types in lineage-specific progenitor cells and revealed the most possible transition paths and critical points in the cellular processes. These results demonstrated MarkovHC's effectiveness in facilitating the stratification of cells, identification of cell populations, and characterization of cellular trajectories and critical points.

INTRODUCTION

High-throughput single-cell omics technologies, such as single-cell RNA sequencing (scRNA-Seq), single-cell ATAC sequencing (scATAC-Seq), and mass cytometry, provide a tremendous amount of data resource that can be used in studying functional cell subpopulations and their lineage relationships. Currently, cell clustering, cellular trajectory reconstruction, and critical point detection are typically done through separate analyses. Modern clustering algorithms include Seurat (1), SC3 (2), SIMLR (3), etc. (4–8) Classical ones include K-Means (9), hierarchical clustering (10), density-based methods (11–13) spectral method (14), and model-based method (15). After clustering, cellular trajectories analysis is performed to study relationships among cell populations. These trajectories correspond to biological processes such as lineage development and cell differentiation. Among existing trajectory reconstruction tools (16–22), monocle (16–18) is very popular. After this, an important downstream analysis is to detect critical or branching points on the cellular trajectory. Within the tipping-point theory (23,24), the transitions among cellular clusters occur when small perturbations at a cellular critical point result in moving from one stable cluster to another.

[*]To whom correspondence should be addressed. Tel: +1 972 883 2528; Fax: +1 972 883 4551; Email: michael.zhang@utdallas.edu
Correspondence may also be addressed to Minping Qian. Tel: +86 13718116995; Email: qianmp@math.pku.edu.cn

SGE (25) and scRCMF (26) are two useful tools for this task. These separate tools may perform reasonably well in their respective tasks, but we found it inconvenient to combine these independent tools in a joint analysis.

An ideal approach to performing these three analyses jointly could allow users to explore cell subpopulations and their lineage relationships more efficiently and more effectively with customized resolutions, this motivated us to develop MarkovHC. As shown in Figure 1A, MarkovHC can stratify common cell populations and their sub-populations by a hierarchy, and simultaneously detect trajectories and critical points among the cell populations on each resolution. The resolution of cell stratification unduly influences cell population identification in single-cell data analyses. For example, a resolution separating human ES cells from neuronal progenitor cells may not readily subdivide subtypes in neuronal progenitor cells. In the human cell atlas (27), cell lineage means the developmental relationship of cells; cell type implies a notion of homeostatic persistence; cell state refers to more inducible or transient properties. However, the boundaries among these concepts at the gene transcriptional level can be fuzzy, partly due to the limit of our knowledge and the understanding of cellular dynamics. A bottom-up cellular hierarchy is intrinsic in single-cell omics data (28). Recent works (4–8) have shown that hierarchies of cell populations can be used to effectively interrogate the echelons of cells. The primary aim of our algorithm is a hierarchical and interpretable clustering for systematic and multiscale single-cell omics data analysis. Reconstructing differentiation paths and detecting critical points can also be influenced by the clustering resolution. For instance, to detect the lineage path and critical points from the 8-cell embryo stage to the inner cell mass stage, an appropriate clustering resolution should be chosen to discriminate them. To this end, MarkovHC is designed to build a cellular hierarchy and solve the two related problems on each resolution simultaneously.

We demonstrate the effectiveness of MarkovHC through several benchmark analyses. Five simulated datasets and ten publicly available datasets including scRNA-Seq, scATAC-Seq, and mass cytometry datasets were used for comparative tests. MarkovHC performed equal to or better than the state-of-the-art methods in terms of clustering accuracy when compared to known labels. Additionally, MarkovHC correctly stratified mouse lineage-specific progenitor cells and was able to reconstruct the path in 'continuum' data and detect critical points between stages. We further used MarkovHC to explore cell differentiation in human preimplantation embryo development and disease progression in gastric cancer. MarkovHC reconstructed the correct lineage tree from the 8-cell stage to sub-populations in trophectoderm and inner cell mass. MarkovHC also revealed two potential 'hidden' trajectories from mesenchymal stem cells to early gastric cancer cells.

MATERIAL AND METHODS

Overview

The basic idea behind MarkovHC is intuitive. Waddington's epigenetic landscape (29,30) is a classical and metaphysical concept. For high-dimensional single-cell omics data that can be embedded in low-dimensional manifolds (31), it is feasible to explore the cellular landscape by utilizing these data. In Figure 1B, four basins correspond to four cell populations on the cellular landscape. Pouring water into this landscape, these basins will merge gradually as the water level increases from level1 (Lv.1) to level4 (Lv.4). The basins on these four levels form a four-level hierarchy which is consistent with the topology of the landscape deriving based on the geodesic distance. On each level (resolution), the bottom of the basin (cell cluster) is the attractor (cluster core), the water-flowing path among basins is the most possible transition path (cellular trajectory), and the tipping (critical) point is the critical point on the cellular trajectory. Moving cells across different basins will cost energy which we call pseudo-energy in this work. Further, an explanation of the basic idea from the perspective of the dynamics and intuitive illustration of concepts including basins, attractors, transition paths, and critical points are available in Supplementary Text S1 and Supplementary Figure S1.

Technically, we employed a Markov chain with an adjustable coarse-graining scale ('temperature') parameter to model the hypothetical random walk of a cell over possible gene expression states. The transition probability matrix is defined using similarity and density. Similarity characterizes the degree of flow conductance among cells and density measures the degree of cell concentration. As shown in Figure 1C, MarkovHC algorithm consists of five steps which are elaborated in the following section.

Model

1. **The input data.** A gene expression or other molecular quantities data matrix ($A = \{a_{gc}: g = 1,2,..,G; c = 1,2,\ldots,C\}$ is taken as a generic input (Figure 1C). We can view each distinct cell as a unique state in the G-dimensional 'expression' space.
2. **Shared nearest neighbours and density scores.** The similarity between two states can be calculated using the cells shared between nearest neighbour states (sNN and Jaccard index (32)), and the density can be defined as the node degree.
   In a high-dimensional space, sNN can explore the natural geodesic distance on the underlying manifolds and has been proved to be robust in recovering cell subpopulations and reconstructing trajectories (33,34). We define sNN similarity $s_{i,j}$ between node i and node j as
   $$s_{i,j} = \#\left\{\text{neighbours} \in \text{KNN}_i \cap \text{KNN}_j\right\}, \tag{1}$$
   where KNN_i is the K nearest neighbours of node i.
   For the robustness of our algorithm, the degree of each node (density score) in the sNN network is used to measure the cell aggregation density
   $$D_i = \sum_{j=1}^{n} a(i,j), \text{ where } a(i,j) = \begin{cases} 0, \text{edges}(i,j) = \emptyset, \\ 1, \text{edges}(i,j) \neq \emptyset. \end{cases} \tag{2}$$
   The density scores of the nodes are used together with the similarity matrix to define the basic Markov transition probability.
3. **Initial Markov transition matrix.** Intuitively, the basic idea of the transition probability is that the larger the

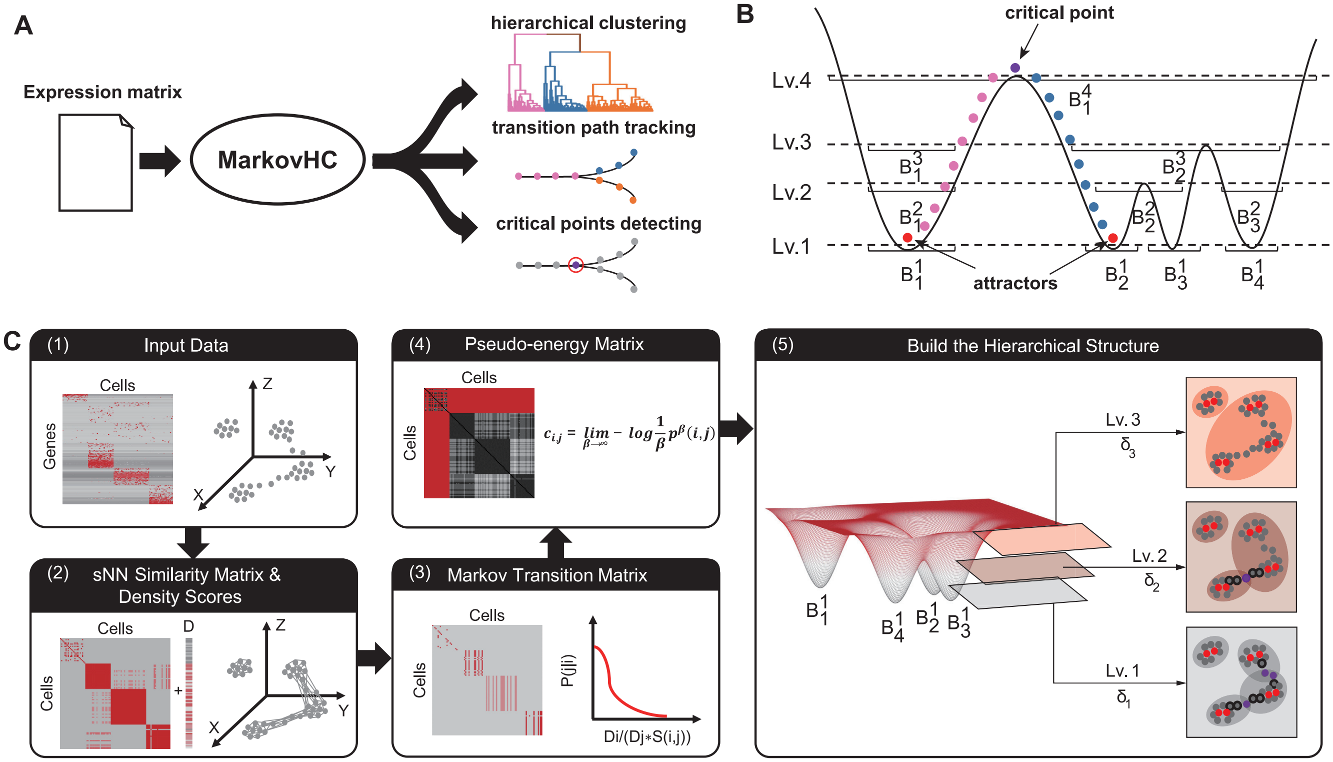

Figure 1. Overview of MarkovHC. **(A)** MarkovHC simultaneously performs hierarchical clustering, transition path tracking, and critical points detecting. **(B)** The intuitive idea behind MarkovHC. **(C)** The workflow of MarkovHC: (1) The original input data is the matrix of genes by cells. (2) We calculate sNN (shared Nearest Neighbours) 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 scores) in the network. (3) The Markov transition matrix is calculated using the similarity matrix and D scores. (4) The pseudo-energy matrix is calculated based on the Markov transition matrix. (5) The hierarchical structure is constructed based on attractors, basins, and critical points on each level.

similarity among nodes, the larger the probability of transition. Besides, nodes with high-density transit to nodes with low-density with low probabilities, vice versa. The transition probability from node i to node j is defined as

$$\mathrm{p}\left(\mathrm{j}\,|\,\mathrm{i}\right) = \lim_{\beta\to\infty} \frac{\mathrm{e}^{-\beta\frac{1}{\mathrm{s}^2_{\mathrm{i,j}}}\left(\frac{\mathrm{D_i}}{\mathrm{D_j}}\right)^2}}{\sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{e}^{-\beta\frac{1}{\mathrm{s}^2_{\mathrm{i,j}}}\left(\frac{\mathrm{D_i}}{\mathrm{D_j}}\right)^2}} \tag{3}$$

We denote the Markov transition matrix by P^1.

4. **Pseudo-energy matrix.** Based on the metastability theory developed by one of us (35), the pseudo-energy matrix is all we need to build the hierarchical structure. The pseudo-energy consumed by the direct transition from node i to node j is defined as

$$\mathrm{c_{i,j}} = \lim_{\beta\to\infty} -\frac{1}{\beta}\log \mathrm{p}^{\beta}\left(\mathrm{j}\,|\,\mathrm{i}\right) \tag{4}$$

This formula transforms the transition probability matrix P^1 to the pseudo-energy matrix C^1 representing the pseudo-energy cost. We can calculate the multi-step pseudo-energy cost between nodes by calculating the length of the shortest path, which is equal to geodesics on the surface of this 'energy landscape' (36) using the Dijkstra algorithm (37). If node i cannot transit to node j, the corresponding entry in matrix C^1 is set to be infinite.

5. **Build the hierarchical structure.** We term basins as clusters and attractors as highest-density 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 (38) modelling the nonlinear dynamics of organisms and biological phenomena as dynamics systems.

The hierarchical structure is built in an iterative manner. As shown in step 5 of Figure 1C, for any given pseudo-energy threshold $\delta > 0$, nodes with 'distances' (i.e. the asymmetric amount of pseudo-energy cost $\mathrm{c_{i,j}}$ in step 4) smaller than δ are partitioned into one basin. From Lv.1 to Lv.3, the basin merges as δ increases from δ_1 to δ_3, and thus forms a hierarchical structure. With a given δ, points in the core of each basin are attractors (red); points connecting attractors are transition paths (black border); tipping points on transition paths are critical points (purple). A more detailed description of this step is available in Supplementary Text S1. For easy level selection, we also provided an algorithm reconciling four indexes to automatically choose levels with probable and interpretable clusters from the hierarchy in Supplementary Text S2. In addition, we put the technical details of MarkovHC in Supplementary Text

S3 and the mathematical details of the metastability of exponentially perturbed Markov chains in Supplementary Text S4.

RESULTS

MarkovHC stratifies cells in agreement with known identities

To assess the performance of MarkovHC in stratification, we used it to stratify five simulation datasets and two scRNA-Seq datasets with known identities. For the five simulation datasets, four of these are 2 or 3-dimensional datasets (Figure 2A, B and Supplementary Figure S2) and the remaining one is a 1000 cells × 5000 genes dataset generated by splatter (39) (Figure 2C). MarkovHC was able to cluster hetero-density basins (Figure 2A), non-convex (the blue basin in Figure 2A and helixes in Supplementary Figure S2M), and continuum basins (Figure 2C), which correspond to hetero-density cell populations, complex cluster shapes, and differentiation states or lineages in single-cell omics data. Meanwhile, the hierarchy in Figure 2B is consistent with the distance among the basins in Figure 2A. Figure 2C shows the successfully detected trajectory and critical points. Supplementary Figure S2N and O show the differentially expressed genes (DEGs) along this trajectory.

For the two scRNA-Seq datasets, we used MarkovHC to interrogate the echelons of cells. From the bottom to the top, cell types, cell states, and cell lineages were presented on the cellular hierarchy. Chu *et.al.* (40) sorted and sequenced 1018 cells of six cell types including undifferentiated H1/H9 human ES cells, neuronal progenitor cells, definitive endoderm cells, endothelial cells, trophoblast-like cells, and human foreskin fibroblasts (Figure 2D). This dataset can be considered as a 'gold-standard' for stratification. Lv.7 was automatically chosen from the hierarchy (Figure 2F). There are six basins (Figure 2E) that perfectly match the six known cell types in Figure 2D. The H1/H9 ES basin and the neuronal progenitor basin separate into two basins on Lv.6 and Lv.5 respectively (Figure 2E, F). The patterns of top 50 DEGs (Supplementary Figure S3A, left; Supplementary Table S1) and 34 lineage-specific markers from Chu's paper (40) (Supplementary Figure S3A, right) suggested the cells of each Lv.5-basin were homogeneous. In H1/H9 ES cells, the up-regulated genes of the dark brown basin (Figure 2F) were enriched in cell division and cell cycle-related terms (Supplementary Figure S3B), which suggested these cells were in a more active division stage. In neuronal progenitor cells, the up-regulated genes of the light blue basin (Figure 2F) were enriched in cerebral cortex development-related terms (Supplementary Figure S3C), while those of the dark blue basin (Figure 2F) were enriched in forebrain development and synapse organization-related terms (Supplementary Figure S3D), which suggested that two sub-cell types might be found in these neuronal progenitor cells. The definitive endoderm cell and endothelial cell were two lineages that were merged on Lv.8.

The classification and hierarchical structures of human peripheral blood mononuclear cells (PBMCs) have been well studied previously (41–43). The echelon of PBMCs also can be considered as a 'gold-standard' for cellular stratification. We used MarkovHC to stratify a scRNA-Seq dataset of 33,000 PBMCs in Supplementary Figure S4A and Supplementary Text S5. To automatically identify cell population transition on the hierarchy with biological meaning, we provided a strategy, BHI selection, based on the biological homogeneity index (the details are available in Supplementary Text S6). Basins correspond to B cells, T cells, NK cells, megakaryocyte, monocyte, pDC, and their sub-cell populations were identified by MarkovHC. Thus, these results demonstrate that MarkovHC is able to automatically stratify common cell populations and their sub-populations simultaneously.

Benchmark MarkovHC against current methods for clustering

To assess the performance of MarkovHC in clustering, we benchmarked it against eight existing methods for clustering (Figure 2G): Seurat (1), SIMLR (3), SC3 (2), K-Means (9), hierarchical clustering with average linkage (HC) (10), Hdbscan (13), spectral clustering (Specc) (14) and model-based clustering (Mclust) (15). We used one simulation dataset (splatter) and seven previously analysed single-cell datasets including five scRNA-Seq datasets (Kolod (44), Pollen (45), Usoskin (46), Zeisel (8), and Celegans (47)), one mass cytometry dataset (cytof (48)) and one scATAC-Seq dataset (scATAC; detailed analyses of this dataset are available in Supplementary FigureS4 and Supplementary Text S7) from a variety of biological systems. For each of these datasets, cell labels have been well identified. In order to measure the agreement between known identities and clustering labels, we used adjusted rand index (ARI) (49) and normalized mutual information (NMI) (50) as test statistics.

Cells were coloured by the cluster labels from the original study and each clustering method in Supplementary FigureS5. Splatter is a 'continuum' dataset. Celegans have been considered 'gold-standard' to test lineage clustering algorithms as Packer *et al.* (47) carefully annotated these cells and lineages. In tissue development and cell differentiation, the single-cell data are in a 'continuum', which makes it harder to cluster. As MarkovHC depends on topological connectivity (51,52) and local density, it outperformed the other methods in clustering these two datasets ('splatter' and 'Celegans' in Figure 2G and Supplementary Figure S5). MarkovHC performed equal to or better than those methods in clustering scRNA-Seq datasets (Kolod (44), Pollen (45), Usoskin (46), Zeisel (8), and Celegans (47) in Figure 2G). Since MarkovHC is free of data distribution assumption, it can be applied to other omics data such as mass cytometry data and scATAC-Seq data, too. MarkovHC, Seurat, and Mclust performed better than the rest in clustering the cytof (48) and scATAC datasets. In Supplementary Text S7, we applied MarkovHC on two scATAC-Seq datasets with matching scRNA-Seq datasets. The results showed its good scalability in clustering scATAC-Seq data. We also noticed that using scRNA-Seq and scATAC-Seq datasets together could improve clustering accuracy and identify more and better sub-cell types. We also theoretically compared MarkovHC with eighteen popular clustering methods (1–7,10–14,53–62), topological data analysis (51), and four representative trajectory construction meth-

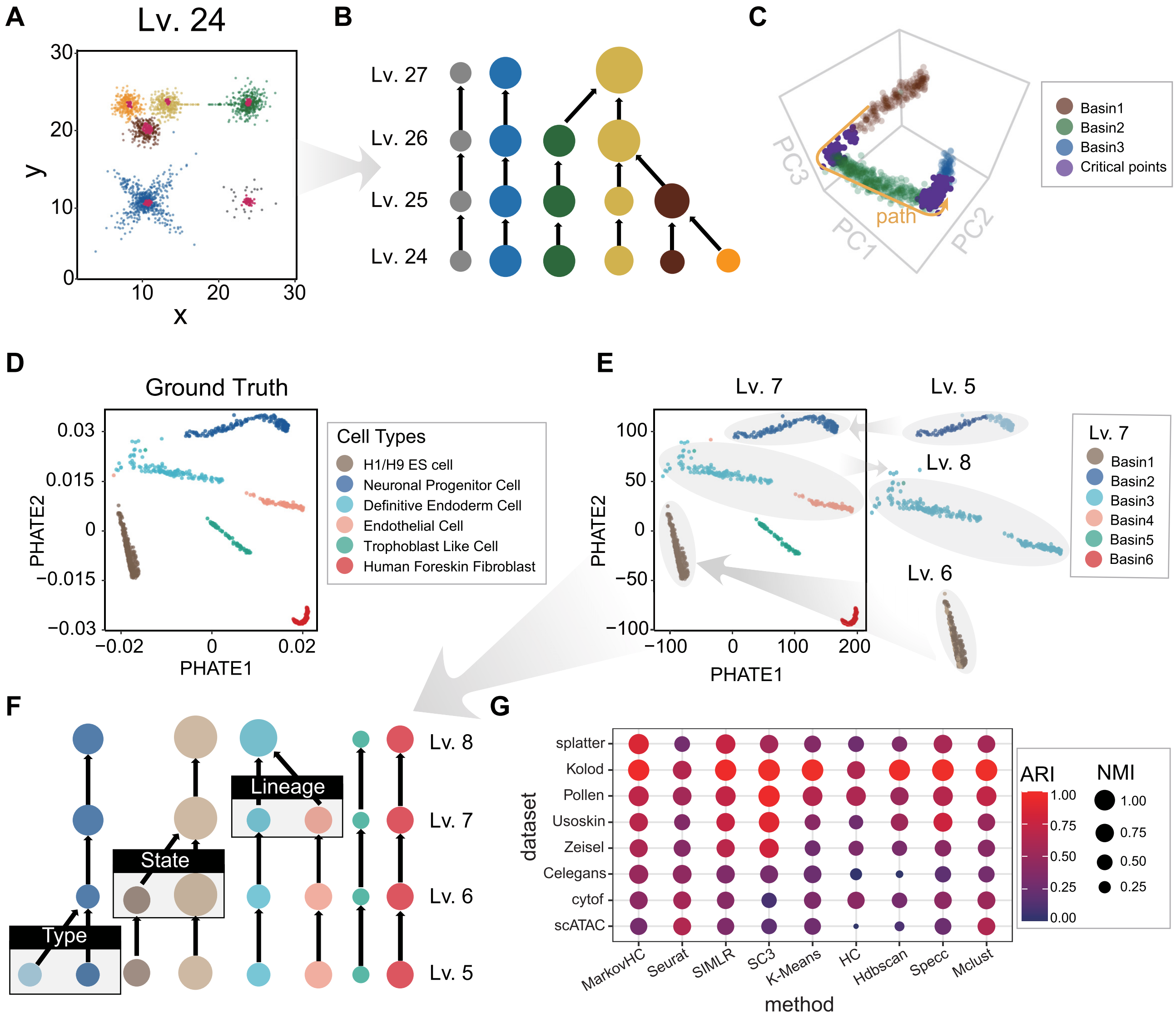

Figure 2. MarkovHC stratifies and clusters cells in agreement with known identities. **(A)** These 2-dimensional basins and attractors (red) found by MarkovHC are consistent with the topology. **(B)** The hierarchy from Lv.24 to Lv.27 of basins in (A). The sizes of basins represent the number of samples and the colors indicate different basins. **(C)** These 1000 cells × 5000 genes data were projected into 3-dimensional space by principal component analysis. Three basins were clustered (brown, green, and blue; the purple points are critical points; the yellow arrow shows the path from basin1 to basin 3). **(D)** scRNA-Seq data (40) of 1018 human ES cell-derived lineage-specific progenitors were projected into 2-dimensional space by phateR. **(E)** Basins from Lv.5 to Lv.8 reveal known cell types and sub-basins in the neuronal progenitor cells and H1/H9 ES cells. **(F)** From the bottom to the top, levels of the hierarchical structure correspond to cell types, cell states, and cell lineages. **(G)** ARI (Adjusted Rand Index) and NMI (Normalized Mutual Information) show MarkovHC performed equal to or better than these methods in clustering.

ods (16–18,20–22) in Supplementary Text S8 and Supplementary Table S2.

To sum up, overall, MarkovHC, Seurat, SIMLR, and SC3 exhibited superior performance in clustering against the others. As MarkovHC utilizes connectivity and density derived from data, it outperformed all other algorithms in clustering ‘continuum’ datasets (‘splatter’ and ‘Celegans’ in Figure 2G and Supplementary Figure S5). Besides, as MarkovHC is free of data distribution assumption, it can be used to cluster other omics data as long as the ‘similarity’ between cells can be reliably calculated using the data.

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

To reveal transition paths among lineage-related cell types and detect critical points in human preimplantation embryo development, we applied MarkovHC to analyse the scRNA-Seq dataset (63) of 1529 single-cells collected from 88 human preimplantation embryos at seven stages (Figure 3A). Petropoulos *et al.* (63) identified eight groups in these cells, including pre-lineages, trophectoderm (TE), inner cell mass (ICM), epiblast (EPI), primitive endoderm (PE), E5mid, mural, and polar in these data.

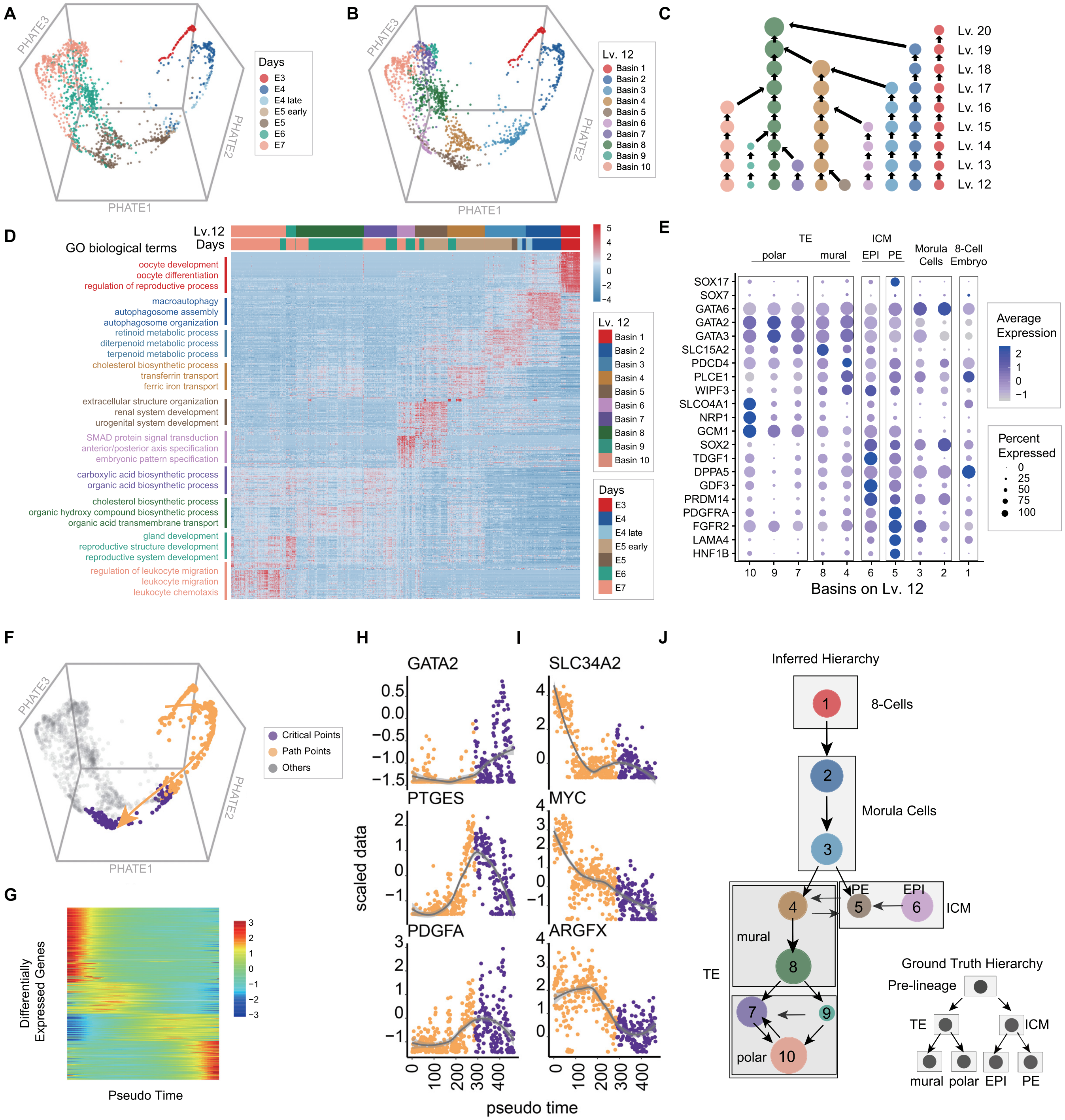

Figure 3. MarkovHC revealed transition paths and critical points in human preimplantation embryo development. **(A)** The scRNA-Seq data (63) of 1529 human preimplantation embryos cells from the E3 stage to the E7 stage were projected into 3-dimensional space by phateR. E3–E7 indicates the embryonic day. E4.late and E5.early indicate cells picked 4–6 hours later and earlier than that in the E4 stage and the E5 stage, respectively. **(B)** Ten basins on Lv.12 correspond to ten asynchronous development stages in human preimplantation embryos cells. **(C)** The cellular hierarchy from Lv.12 to Lv.20. **(D)** The heatmap of the top 50 DEGs and enriched GO terms per basin. **(E)** Four main cell types with sub-populations which are 8-cell embryo, morula cell, ICM (inner cell mass), and TE (trophectoderm) were identified according to marker genes expression. **(F)** The transition path (yellow arrow) from the 8-cell embryo to ICM was tracked. The yellow points indicate cells along the path and the purple points indicate the critical points from morula cells to ICM. **(G)** DEGs along the transition path in (F). **(H, I)** Important marker genes show increasing and decreasing 'gene-flow' trends along the path. Gene expression varies dramatically around critical points (purple points). **(J)** The inferred development hierarchy is consistent with the ground truth of the development hierarchy (in the lower right corner).

The basins on Lv.12 (Figure 3B) were automatically chosen from the cellular hierarchy (Figure 3C) by MarkovHC. To reveal the biological processes in each basin, the top 50 DEGs of each basin (Supplementary Table S3) were enriched in GO terms related to the process of embryonic development (Figure 3D; Supplementary Table S4). We used canonical marker genes from the paper of Petropoulos *et al.* (63) and LifeMap Discovery (https://discovery.lifemapsc.com/in-vivo-development/) to identify the basins in Figure 3E: four main cell types were 8-cell embryo (GATA6-, SOX7-, SOX17-), morula cell (GATA6+, SOX7-), ICM (inner cell mass; SOX2+, PDGFRA+) and TE (trophectoderm; GATA2+, GATA3+); sub-groups (63) in ICM and TE were polar cells (GCM1+, NRP1+, SLCO4A1+), mural (WIPF3+, PLCE1+, PDCD4+, SLC15A2+), EPI (SOX2, TDGF1, DPPA5, GDF3, PRDM14) and PE (PDGFRA, FGFR2, LAMA4, HNF1B).

We also used MarkovHC to reconstruct the transition path and detect the critical points from the 8-cell embryo to ICM as shown in Figure 3F. In this figure, yellow points indicate cells on the transition path and purple points indicate critical points from morula cells to ICM. There were 2476 DEGs (Supplementary Table S5) along this path (Figure 3G). The expression of significant genes dramatically changed around critical points (purple points in Figure 3H and I) indicating that these points may play pivotal roles in the developmental process. The development hierarchy (Figure 3J) was also inferred according to the transition probability matrix (Supplementary FigureS6A) and the trajectories (Supplementary FigureS6B), which is consistent with the ground truth hierarchy of human embryo development (in the lower right corner of Figure 3J).

MarkovHC detected clinical related pathways from MSCs to gastric cancer cells

The transition process from normal gastric cells to cancer cells is complex and not fully described. Many studies have been carried out to better understand carcinogenesis (64–66). For example, Zhang *et al.* (67) conducted a single-cell transcriptomic study on gastric antral biopsies and identified gastric cancer-related cell populations. We used MarkovHC to analyse the scRNA-Seq data of 831 mesenchymal stem cells (MSCs) and 695 MSC-origin early gastric cancer cells (EGCs) (Figure 4A) from their study.

To identify the cell types, we used MarkovHC to automatically choose two basins on Lv.21 (Figure 4B) and their sub-basins on Lv.18 (Figure 4C, D). GO analysis of the top 50 DEGs of each basin on Lv.18 (Supplementary Table S6) were related to gastric cancer progression (Supplementary Table S7; Figure 4E). To reconstruct the trajectories from MSCs to EGCs, we used MarkovHC to detect two transition paths with critical points from Basin 1 to Basin 4 (Figure 4C). MSC markers (OLFM4, EPHB2, SOX9) gradually decreased from MSCs to EGCs, while EGC markers (KLK10, SLC11A2, SULT2B1, KLK7, ECM1, LMTK3) gradually increased (Supplementary Figure S6C) (67). The inferred transitions among these basins based on the transition probability matrix (Supplementary FigureS6D) were shown in Figure 4D.

In Figure 4F, G, and Supplementary Table S8, the decreased genes of Path1 and Path2 were enriched in protein targeting, protein stabilization, and cell cycle arrest-related terms. The increased genes of Path1 were enriched in neutrophil-mediated immunity and alcohol metabolic process-related terms, suggesting that the potential disease progress might be driven by alcohol stimulus (68). Interestingly, this was partially supported by the fact that this patient (p8 in Zhang *et al.*'s paper (67)) had a chronic alcohol consumption history. In Path2, the increased genes were enriched in neutrophil-mediated immunity and response to metal ion-related terms suggesting this path could be driven by metal pollution stimulus (69). As shown in Figure 4H, we observed OLFM4, a marker for stem cells in the human intestine, decreased along Path1 (70), while CEACAM6, which plays important role in invasion and metastasis in Gastric Cancer, increased (71). In Figure 4I, SOX4 decreased along Path2, which was consistent with the results that MiR-596 down-regulates SOX4 expression and was a potential novel biomarker for gastric cancer (72), while NEAT1, a long non-coding RNA promoting viability and migration of gastric cancer cells through up-regulation of microRNA-17, increased (73). Furthermore, expression values of these genes dramatically changed around the critical points (purple points in Figure 4H, I), which suggested these were unstable cells in the trajectories. Thus, these two paths might be two potential routes from MSCs to EGCs for this patient, which could be valuable in revealing the underlying mechanisms of the disease progression.

DISCUSSION

In this paper, we developed MarkovHC based on the metastability of exponentially perturbed Markov chain to jointly perform hierarchical clustering, trajectory reconstruction, and critical point detection. We also developed a user-friendly R package, 'MarkovHC' (https://github.com/ZhenyiWangTHU/MarkovHC). For ease of use, we provided algorithms to automatically choose the number of PCs (Supplementary Text S3 and Supplementary Figure S7), to recommend levels with the reasonable number of clusters, and to identify levels with biologically meaningful basin transitions. The results showed that MarkovHC could accurately cluster cells into populations at different resolutions, in terms of the established knowledge. Besides, MarkovHC performed equal to or better than the state-of-the-art algorithms in clustering specific tasks. Since MarkovHC is free of data distribution assumption, it can be applied to other omics data such as mass cytometry data and scATAC-Seq data. Furthermore, the transition paths and the critical points among cell populations detected by MarkovHC could reveal certain developmental processes well, such as human embryonic development or cancer progression from MSCs to EGCs.

Besides the analyses described above, there could be some other analyses where MarkovHC is helpful. Firstly, one could calculate pseudo-time (16–19) on a customized resolution (the details are available in Supplementary Text S3 and Supplementary Figure S8). Secondly, one could start out from a given relevant biological level, e.g. based on prior knowledge or other clustering methods, and use

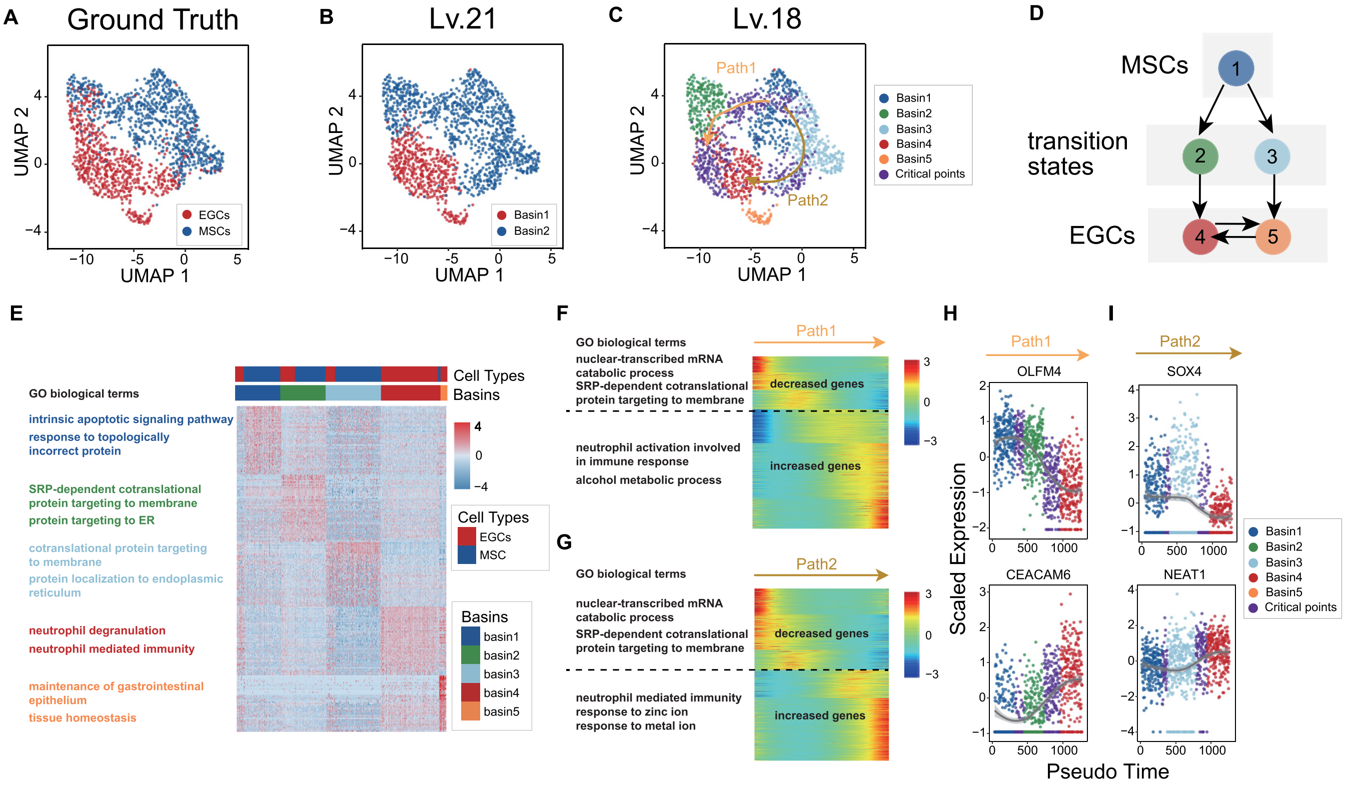

Figure 4. MarkovHC detected critical points from MSCs to gastric cancer cells. **(A)** 831 mesenchymal stem cells (MSCs) and 695 MSC-origin early gastric cancer cells (EGCs) (67) were projected into 2-dimensional space by UMAP. **(B)** MarkovHC found two basins on Lv.21. **(C)** Five basins were clustered and two transition paths from MSCs to EGCs were inferred by MarkovHC. Purple points indicate critical points on the transition paths. **(D)** The inferred transitions among basins in (C). **(E)** The heatmap and enriched GO terms of the top 50 DEGs per basin in (C). **(F, G)** DEGs along Path1 and Path2 in (C). **(H-I)** OLFM4 and CEACAM6 showed opposite 'gene-flow' trends along Path1 (H). SOX4 and NEAT1 showed opposite 'gene-flow' trends along Path2 (I). The expression values of these genes dramatically changed around the critical points.

MarkovHC to explore the next level up or down in the hierarchy. Thirdly, the hierarchical structure obtained by MarkovHC could be used to jointly analyse multi-omics data, e.g. by using UnionCom (74) recently developed by Cao *et al.* for the unsupervised topological alignment of single-cell multi-omics data. In addition, we put the detailed time complexity analysis of MarkovHC in Supplementary Text S8 and Supplementary Table S9. Although our method costs more time than Seurat, it can get more information from data including a cluster hierarchy, transition paths, and critical points among clusters. Compared with all the other methods, MarkovHC has a better computational complexity.

Finally, as our method is based on a general mathematics theory (Supplementary Text S4 and Supplementary Figure S9), it is robust as long as the input matrix reliably measures the 'similarity' among samples (Supplementary Text S7, Supplementary Text S8, and Supplementary Figure S10). It is possible to further apply the metastability of exponentially perturbed Markov chain to develop algorithms for spatial transcriptomic data analysis.

DATA AVAILABILITY

R package 'MarkovHC' is an open source available in the GitHub repository (https://github.com/ZhenyiWangTHU/MarkovHC).

The scRNA-Seq dataset used in Figure 2D was downloaded from GEO under accession number GSE75748. In these data, the labeled cell types include neuronal progenitor cells (NPCs, ectoderm derivatives, n = 173), DE cells (endoderm derivatives, n = 138), endothelial cells (ECs, mesoderm derivatives, n = 105), trophoblast-like cells (TBs, extraembryonic derivatives, n = 69), undifferentiated H1 (n = 212) and H9 (n = 162) human ES cells, and human foreskin fibroblasts (HFFs, n = 159).

The scRNA-Seq datasets and labels in Figure 2G ('Kolod', 'Pollen', 'Usoskin', 'Zeisel') were downloaded from 'SIMLR' repository (https://github.com/BatzoglouLabSU/SIMLR).

The scRNA-Seq dataset and lineage labels of C. elegans embryogenesis in Figure 2G ('Celegans') were downloaded from GEO under accession number GSE126954.

The mass cytometry dataset of PBMC in Figure 2G ('cytof') was downloaded from the Supplementary materials of Anchang's paper (48) (https://www.nature.com/articles/nprot.2016.066).

The scRNA-Seq dataset of 33k PBMCs in Supplementary Figure S4A was downloaded from 10X genomics support (https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc33k).

The scRNA-Seq and scATAC-seq datasets in Supplementary Figure S4B, C and D were downloaded from GEO under accession numbers GSE115968 and GSE107651.

The single-cell datasets in Supplementary Figure S4E, F, and G measuring both DNA accessibility and gene expression in the same cells were downloaded from 10x genomics support (Count matrix: https://cf.10xgenomics.com/samples/cell-arc/1.0.0/pbmc_granulocyte_sorted_10k/pbmc_granulocyte_sorted_10k_filtered_feature_bc_matrix.h5; ATAC fragment file: https://cf.10xgenomics.com/samples/cell-arc/1.0.0/pbmc_granulocyte_sorted_10k/pbmc_granulocyte_sorted_10k_atac_fragments.tsv.gz; ATAC fragment file index: https://cf.10xgenomics.com/samples/cell-arc/1.0.0/pbmc_granulocyte_sorted_10k/pbmc_granulocyte_sorted_10k_atac_fragments.tsv.gz.tbi)

The scRNA-Seq dataset of 1529 single-cells in Figure 3 was downloaded from EMBL-EBI under accession number E-MTAB-3929 (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-3929/).

The scRNA-Seq dataset of 831 mesenchymal stem cells (MSCs) and 695 MSC-origin early gastric cancer cells (EGCs) was downloaded from GEO under accession number GSE134520.

The scRNA-Seq dataset of 242,533 single-cells from mouse cell atlas (MCA) was downloaded from https://www.dropbox.com/s/8d8t4od38oojs6i/MCA.zip?dl=1

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to thank Peng Zhang, Yuyao Song and Kui Hua for the discussions.

Authors' Contributions: M.P.Q., M.Q.Z., Z.Y.W. and Y.J.Z. conceived the project. Z.Y.W. and Y.J.Z. conducted the research. M.P.Q. and M.Q.Z. supervised the research. Z.Y.W. implemented the R package 'MarkovHC'. Z.Y.W. did the Formal Analysis. Z.Y.W., Z.F.Y., L.Z. and M.L.S. did investigation. Z.Y.W. wrote the original draft; M.Q.Z., Z.F.Y., Z.Y.W., Y.J.Z., L.Z., Y.C., M.L.S., Z.Y.Y., Q.M.Z. and M.P.Q. contributed to the writing of the manuscript. Z.F.Y. and Z.Y.W. did the visualization. All authors read and approved the final manuscript.

FUNDING

This work was supported by Natural Science Foundation of China [81890991, 62050152] to M.L.S; the Natural Science Foundation of China [81890994], National Key Research and Development Program [2018YFA0801402] to Y.C; M.Q.Z acknowledges the support by the Cecil H. and Ida Green Distinguished Chair.

Conflict of interest statement. None declared.

REFERENCES

1. Butler,A., Hoffman,P., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.
2. Kiselev,V.Y., Kirschner,K., Schaub,M.T., Andrews,T., Yiu,A., Chandra,T., Natarajan,K.N., Reik,W., Barahona,M., Green,A.R. *et al.* (2017) SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods*, **14**, 483–486.
3. Wang,B., Zhu,J., Pierson,E., Ramazzotti,D. and Batzoglou,S. (2017) Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nat. Methods*, **14**, 414–416.
4. Lin,P., Troup,M. and Ho,J.W. (2017) CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biol.*, **18**, 59.
5. Zurauskiene,J. and Yau,C. (2016) pcaReduce: hierarchical clustering of single cell transcriptional profiles. *BMC Bioinformatics*, **17**, 140.
6. Chen,J., Schlitzer,A., Chakarov,S., Ginhoux,F. and Poidinger,M. (2016) Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nat. Commun.*, **7**, 11988.
7. Schwartz,G.W., Zhou,Y., Petrovic,J., Fasolino,M., Xu,L., Shaffer,S.M., Pear,W.S., Vahedi,G. and Faryabi,R.B. (2020) TooManyCells identifies and visualizes relationships of single-cell clades. *Nat. Methods*, **17**, 405–413.
8. Zeisel,A., Munoz-Manchado,A.B., Codeluppi,S., Lonnerberg,P., La Manno,G., Jureus,A., Marques,S., Munguba,H., He,L., Betsholtz,C. *et al.* (2015) Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, **347**, 1138–1142.
9. MacQueen,J. (1967) Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. **1**, pp. 281–297.
10. Sokal,R.R. and Michener,C.D. (1958) A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.*, **38**, 1409–1438.
11. Ester,M., Kriegel,H.P., Sander,J. and Xu,X. (1996), A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, **96**, 226–231.
12. Ankerst,M., Breunig,M.M., Kriegel,H-P and Sander,J. (1999) OPTICS: Ordering Points To Identify the Clustering Structure. *ACM Sigmod record*, **28**, 49–60.
13. Campello,R.J.G.B., Moulavi,D., Zimek,A. and Sander,J. (2015) Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *TKDD*, **10**, 1–51.
14. Ng.,A.Y., Jordan,M.I. and Weiss,Y. (2002) On spectral clustering: Analysis and an algorithm. *Adv. Neural Inf. Process. Syst.*, **2**, 849–856.
15. Fraley,C. and Raftery,A.E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*, **97**, 611–631.
16. Trapnell,C., Cacchiarelli,D., Grimsby,J., Pokharel,P., Li,S., Morse,M., Lennon,N.J., Livak,K.J., Mikkelsen,T.S. and Rinn,J.L. (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, **32**, 381–386.
17. Qiu,X., Mao,Q., Tang,Y., Wang,L., Chawla,R., Pliner,H.A. and Trapnell,C. (2017) Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods.*, **14**, 979–982.
18. Cao,J., Spielmann,M., Qiu,X., Huang,X., Ibrahim,D.M., Hill,A.J., Zhang,F., Mundlos,S., Christiansen,L., Steemers,F.J. *et al.* (2019) The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, **566**, 496–502.
19. Ji,Z. and Ji,H. (2016) TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic. Acids. Res.*, **44**, e117.
20. Farrell,J.A., Wang,Y., Riesenfeld,S.J., Shekhar,K., Regev,A. and Schier,A.F. (2018) Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, **360**, 6392.
21. Schiebinger,G., Shu,J., Tabaka,M., Cleary,B., Subramanian,V., Solomon,A., Gould,J., Liu,S., Lin,S., Berube,P. *et al.* (2019) Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, **176**, 928–943.
22. Qiu,P., Simonds,E.F., Bendall,S.C., Gibbs,K.D., Bruggner,R.V., Linderman,M.D., Sachs,K., Nolan,G.P. and Plevritis,S.K. (2011) Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.*, **29**, 886–891.
23. Lenton,T.M. and Livina,V.N. (2016) Detecting and anticipating climate tipping points. *Geophys. Monogr.*, **214**, 51–62.
24. Clements,C.F., McCarthy,M.A. and Blanchard,J.L. (2019) Early warning signals of recovery in complex systems. *Nat. Commun.*, **10**, 1681.
25. Zhong,J., Han,C., Zhang,X., Chen,P. and Liu,R. (2020) Predicting cell fate commitment of embryonic differentiation by single-cell graph

entropy. bioRxiv doi: https://doi.org/10.1101/2020.04.22.055244, 24 April 2020, preprint: not peer reviewed.

26. Zheng,X., Jin,S., Nie,Q. and Zou,X. (2019) scRCMF: Identification of cell subpopulations and transition states from Single-Cell transcriptomes. *IEEE. Trans. Biomed. Eng.*, **67**, 1418–1428.
27. Regev,A., Teichmann,S.A., Lander,E.S., Amit,I., Benoist,C., Birney,E., Bodenmiller,B., Campbell,P., Carninci,P., Clatworthy,M. *et al.* (2017) Science forum: the human cell atlas. *Elife*, **6**, e27041.
28. Miller,J.G. (1965) Living systems: Basic concepts. *Behav. Sci.*, **10**, 193–237.
29. Wright,S. (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. XI Int. Congr. Genet.*, **1**, 356–366.
30. Waddington,C.H. (1957) In: *The Strategy of the Genes*. Allen & Unwin, London.
31. van Dijk,D., Sharma,R., Nainys,J., Yim,K., Kathail,P., Carr,A.J., Burdziak,C., Moon,K.R., Chaffer,C.L., Pattabiraman,D. *et al.* (2018) Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, **174**, 716–729.
32. Jaccard,P. (1912) The distribution of the flora in the alpine zone. *New Phytol.*, **11**, 37–50.
33. Haghverdi,L., Buttner,M., Wolf,F.A., Buettner,F. and Theis,F.J. (2016) Diffusion pseudotime robustly reconstructs lineage branching. *Nat. Methods*, **13**, 845–848.
34. Setty,M., Tadmor,M.D., Reich-Zeliger,S., Angel,O., Salame,T.M., Kathail,P., Choi,K., Bendall,S., Friedman,N. and Pe'er,D. (2016) Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.*, **34**, 637–645.
35. Chen,D., Feng,J. and Qian,M. (1996) Metastability of exponentially perturbed Markov chains. *Science in China Series A: Mathematics*, **39**, 7.
36. Chen,Z., An,S., Bai,X., Gong,F., Ma,L. and Wan,L. (2019) DensityPath: an algorithm to visualize and reconstruct cell state-transition path on density landscape for single-cell RNA sequencing data. *Bioinformatics*, **35**, 2593–2601.
37. Dijkstra,E.W. (1959) A note on two problems in connexion with graphs. *Numer. Math.*, **1**, 269–271.
38. Ao,P., Galas,D., Hood,L. and Zhu,X. (2008) Cancer as robust intrinsic state of endogenous molecular-cellular network shaped by evolution. *Med. Hypotheses*, **70**, 678–684.
39. Zappia,L., Phipson,B. and Oshlack,A. (2017) Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.*, **18**, 174.
40. Chu,L.F., Leng,N., Zhang,J., Hou,Z., Mamott,D., Vereide,D.T., Choi,J., Kendziorski,C., Stewart,R. and Thomson,J.A. (2016) Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.*, **17**, 173.
41. Papalexi,E. and Satija,R. (2018) Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.*, **18**, 35–45.
42. Villani,A.C., Satija,R., Reynolds,G., Sarkizova,S., Shekhar,K., Fletcher,J., Griesbeck,M., Butler,A., Zheng,S., Lazo,S. *et al.* (2017) Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, **356**, eaah4573.
43. Bendall,S.C., Simonds,E.F., Qiu,P., Amir el,A.D., Krutzik,P.O., Finck,R., Bruggner,R.V., Melamed,R., Trejo,A., Ornatsky,O.I. *et al.* (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science*, **332**, 687–696.
44. Kolodziejczyk,A.A., Kim,J.K., Tsang,J.C., Ilicic,T., Henriksson,J., Natarajan,K.N., Tuck,A.C., Gao,X., Buhler,M., Liu,P. *et al.* (2015) Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell*, **17**, 471–485.
45. Pollen,A.A., Nowakowski,T.J., Shuga,J., Wang,X., Leyrat,A.A., Lui,J.H., Li,N., Szpankowski,L., Fowler,B., Chen,P. *et al.* (2014) Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.*, **32**, 1053–1058.
46. Usoskin,D., Furlan,A., Islam,S., Abdo,H., Lonnerberg,P., Lou,D., Hjerling-Leffler,J., Haeggstrom,J., Kharchenko,O., Kharchenko,P.V. *et al.* (2015) Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nat. Neurosci.*, **18**, 145–153.
47. Packer,J.S., Zhu,Q., Huynh,C., Sivaramakrishnan,P., Preston,E., Dueck,H., Stefanik,D., Tan,K., Trapnell,C., Kim,J. *et al.* (2019) A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. *Science*, **365**, 6459.
48. Anchang,B., Hart,T.D., Bendall,S.C., Qiu,P., Bjornson,Z., Linderman,M., Nolan,G.P. and Plevritis,S.K. (2016) Visualization and cellular hierarchy inference of single-cell data using SPADE. *Nat. Protoc.*, **11**, 1264–1279.
49. Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classification*, **2**, 193–218.
50. Vinh,N.X., Epps,J. and Bailey,J. (2010) Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, **11**, 2837–2854.
51. Wasserman,L. (2018) Topological data analysis. *Annu. Rev. Stat. Appl.*, **5**, 501–532.
52. Wolf,F.A., Hamey,F.K., Plass,M., Solana,J., Dahlin,J.S., Gottgens,B., Rajewsky,N., Simon,L. and Theis,F.J. (2019) PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.*, **20**, 59.
53. Sibson,R. (1973) SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, **16**, 30–34.
54. Defays,D. (1977) An efficient algorithm for a complete link method. *The Computer Journal*, **20**, 364–366.
55. Blondel,V.D., Guillaume,J.L., Lambiotte,R. and Lefebvre,E. (2008) Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, **10**, P10008.
56. Rodriguez,M.Z., Comin,C.H., Casanova,D., Bruno,O.M., Amancio,D.R., Costa,L.D.F. and Rodrigues,F.A. (2019) Clustering algorithms: A comparative approach. *PLoS One*, **14**, e0210236.
57. Kiselev,V.Y., Andrews,T.S. and Hemberg,M. (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.*, **20**, 273–282.
58. Wolf,F.A., Angerer,P. and Theis,F.J. (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, **19**, 15.
59. Guo,M., Wang,H., Potter,S.S., Whitsett,J.A. and Xu,Y. (2015) SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis. *PLoS Comput. Biol.*, **11**, e1004575.
60. Ertöz,L, Steinbach,M and Kumar,V. (2003) Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In: *Third SIAM International Conference on Data Mining(SDM)*. San Fransico. CA, pp. 47–58.
61. Bhargav,S. and Pawar,M. (2016) A review of clustering methods forming non-convex clusters with missing and noisy data. *IJCSE*, **4**, 39–44.
62. van Dongen,S. (2000) *A Cluster Algorithm for Graphs*. Technical Report. INS-R0010 National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam. (Stichting Mathematisch Centrum, Amsterdam).
63. Petropoulos,S., Edsgard,D., Reinius,B., Deng,Q., Panula,S.P., Codeluppi,S., Plaza Reyes,A., Linnarsson,S., Sandberg,R. and Lanner,F. (2016) Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell*, **165**, 1012–1026.
64. Cheng,J. and Fan,X.M. (2013) Role of cyclooxygenase-2 in gastric cancer development and progression. *World J. Gastroenterol.*, **19**, 7361–7368.
65. Boussioutas,A., Li,H., Liu,J., Waring,P., Lade,S., Holloway,A.J., Taupin,D., Gorringe,K., Haviv,I., Desmond,P.V. *et al.* (2003) Distinctive patterns of gene expression in premalignant gastric mucosa and gastric cancer. *Cancer Res.*, **63**, 2569–2577.
66. Alzahrani,S., Lina,T.T., Gonzalez,J., Pinchuk,I.V., Beswick,E.J. and Reyes,V.E. (2014) Effect of Helicobacter pylori on gastric epithelial cells. *World J. Gastroenterol.*, **20**, 12767–12780.
67. Zhang,P., Yang,M., Zhang,Y., Xiao,S., Lai,X., Tan,A., Du,S. and Li,S. (2019) Dissecting the Single-Cell Transcriptome Network Underlying Gastric Premalignant Lesions and Early Gastric Cancer. *Cell Rep.*, **27**, 1934–1947.
68. Zali,H., Rezaei-Tavirani,M. and Azodi,M. (2011) Gastric cancer: prevention, risk factors and treatment. *Gastroenterol. Hepatol. Bed Bench*, **4**, 175–185.
69. Yuan,W., Yang,N. and Li,X. (2016) Advances in Understanding How Heavy Metal Pollution Triggers Gastric Cancer. *Biomed. Res. Int.*, **2016**, 7825432.
70. van der Flier,L.G., Haegebarth,A., Stange,D.E., van de Wetering,M. and Clevers,H. (2009) OLFM4 Is a Robust Marker for Stem Cells in Human Intestine and Marks a Subset of Colorectal Cancer Cells. *Gastroenterology*, **137**, 15–17.

71. Zang,M., Zhang,B., Zhang,Y., Li,J., Su,L., Zhu,Z., Gu,Q., Liu,B. and Yan,M. (2014) CEACAM6 Promotes Gastric Cancer Invasion and Metastasis by Inducing Epithelial-Mesenchymal Transition via PI3K/AKT Signaling Pathway. *PLoS One*, **9**, e112908.
72. Chen,Y., Gong,W., Dai,W., Pan,Z., Xu,X. and Jiang,H. (2020) MiR-596 down regulates SOX4 expression and is a potential novel biomarker for gastric cancer. *Translational Cancer Research*, **9**, 1294–1302.
73. Wang,C.L., Wang,D., Yan,B.Z., Fu,J.W. and Qin,L. (2018) 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.
74. Cao,K., Bai,X.Q., Hong,Y.G. and Wan,L. (2020) Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics*, **36**, i48–i56.