Graph topology reframes the coherence of cell-state manifold inference under heterogeneous single-cell observations

Tomohiro Tamura[*,†,1,2], Yusuke Yamane[*,1], Yuji Okano[1,3], Tetsuo Ishikawa[1,3,4], Kazuhiro Sakurada[1,5]

* These authors equally contributed to this work.

† Corresponding author; tamurat626@keio.jp; Keio University School of Medicine and Graduate School of Medicine; Shinanomachi 35, Shinjuku, Tokyo, Japan, 160-8582; +81-3-6709-8107.

1. Department of Extended Intelligence for Medicine, The Ishii-Ishibashi Laboratory, Keio University School of Medicine and Graduate School of Medicine, Shinjuku, Japan.
2. Department of Diabetes, Endocrinology, and Metabolism, National Center for Global Health and Medicine, Japan Institute for Health Security, Shinjuku, Japan.
3. Division of Applied Mathematical Science, RIKEN Center for Interdisciplinary Theoretical and Mathematical Sciences (iTHEMS), RIKEN, Yokohama, Japan.
4. Collective Intelligence Research Laboratory, Graduate School of Arts and Sciences, The University of Tokyo, Meguro, Japan.
5. Predictive Medicine Special Project, RIKEN Center for Integrative Medical Sciences (IMS), Yokohama, Japan.

Abstract

Manifold-based single-cell omics analyses assume that high-dimensional observations can fall into a low-dimensional space encoding biological constraints. In practice, per-cell observation is highly heterogeneous: shallowly- and deeply-observed cells coexist. In an empirical scRNA-seq dataset, shallowly-observed cells cluster together to generate spurious hubs that can give rise to illusory loops in low-dimensional manifold skeletons in graph abstraction. Several imputation methods leave these artifacts largely intact, whereas graph abstraction restricted to homogeneously-observed cells alone recovers tree-like structures locally representing constrained cell state transitions. Simulations further demonstrate that realistic heterogeneous observation can create spurious subclusters and false branching. Grounded by these observations, we propose topological stability descriptors of low-dimensional manifold skeletons to delineate a regime in which manifold-based inference is trustworthy despite realistic heterogeneous observations. Our findings underscore that the heterogeneity of observations is not merely noise but a source of systemic distortion in manifold-based inference that must be addressed.

Introduction

Manifold-based analyses have become a central paradigm in single-cell transcriptomics[1–3]. In this view, each cell is represented as a point in a high-dimensional data space, but observations of biologically attainable states are assumed to lie near a low-dimensional manifold that encodes constraints such as differentiation paths and responses to stimuli. By examining local neighborhood relations in this space and constructing graphs to abstract cellular state similarity, one hopes to recover the coarse-grained, low-dimensional structure from snapshot data and to interpret it as a trajectory of state transitions.

This paradigm implicitly relies on an ideal regime of measurement in which cellular states are represented by uniformly informative observations. In practice, however, droplet-based scRNA-seq produces datasets with pronounced heterogeneous observation: shallowly- and deeply observed cells coexist, so that some cells possess rich molecular profiles to represent their states whereas others appear to have only sparse snapshots. Prior work has characterized how technical noise and gene-level *dropout* affect dimensionality reduction and clustering, and a wide range of imputation and normalization methods has been proposed and discussed to mitigate these effects[4–6]. Much less is known, however, about how heterogeneous single-cell observations themselves affects the geometry and topology of inferred data manifolds, and under what conditions manifold-based inferences remain reliable when shallow and deep observations are mixed.

Here we first show how realistic heterogeneous observations can systematically distort manifold-based inference. Using a peripheral blood mononuclear cell (PBMC) dataset, we demonstrate that shallowly-observed cells in heterogeneous observations form spurious hubs and illusory loops in graph abstraction, whereas restricting analyses

to homogeneously-observed cells recovers a tree-like structure locally representing constrained cell state transitions. Computational simulations reveal how realistic levels of heterogeneous observations can give rise to spurious subclusters, artifactual intermediates, and false branching. Motivated by these observations, we propose topological stability descriptors of low-dimensional manifold skeletons to delineate a regime in which manifold-based inference is trustworthy despite realistic heterogeneous observations. The unavoidable heterogeneity of observations is not merely noise but a major source of systemic distortion in manifold-based inference that must be addressed.

Results

Graph abstraction from heterogeneous single-cell observations generates a loop-rich structure

In general, scRNA-seq data are analyzed following the workflow, which we denote (#): log1p normalization, identification of highly variable genes (HVGs), scaling of gene expression across cells, principal component analysis (PCA), neighbor-graph construction, unsupervised clustering, and non-linear dimensionality regression uniform manifold approximation and projection (UMAP). To investigate whether heterogeneous observations empirically influences manifold-based inference in scRNA-seq data, we first applied the workflow (#) to viable cells in a PBMC scRNA-seq dataset[7]. As a result, Louvain clustering broadly classified cells into monocyte populations (clusters 0, 4, 5, 10, 12, 13, and 17), lymphocyte populations, and other cell types such as platelets (Figure 1A, Supplementary Figure 1A). When total unique molecular identifier (UMI) counts were projected onto the z-axis of these UMAP embeddings, shallowly-observed cells within monocyte populations were confined to specific regions

(Supplementary Figure 1B). To investigate this biased distribution more closely, we subsequently reanalyzed the seven clusters and defined a putative monocyte population M ($4{,}410$ cells) by removing small cell populations representing technical artifacts (Supplementary Figure 1C, D). Applying the workflow (#) and the three-dimensional visualization to M recapitulated the biased distribution of shallowly-observed cells in the manifold-based inference, as observed in the global analysis (Figure 1B–D, Supplementary Figure 1E). Notably, cluster 1 was predominantly composed of shallowly-observed cells, and gave rise to a pronounced heavy tail in the distribution of per-cell total UMI counts.

We next inferred PC neighbor graph-based connectivity among the clusters within M that were coarsely identified as distinct states using partition-based graph abstraction (PAGA)[8]. When we drew edges on the UMAP embedding between cluster pairs whose PAGA-inferred connectivity exceeded a given threshold (e.g., ≥ 0.05), a loop-rich structure emerged among clusters (Figure 1E). Interpreting this graph structure as a simplicial complex and analyzing it topologically revealed nine independent loops; particularly, the cluster 1, prominently composed of shallowly-observed cells, was connected by edges to clusters 0, 2, 4, 5, and 6, and thus, at face value, would be interpreted as a hub-like cell population in manifold-based inference.

Graph abstraction using homogeneous observation alone generates a low-dimensional skeleton locally representing constrained cell state transitions

To investigate whether the results of graph abstraction behave differently without extremely heterogeneous

88 observations containing shallowly-observed cells which formed the spurious hub-like subclusters, we extracted
89 1,986 cells with total UMI counts $> 10{,}000$ from M to define a subset M', and applied the workflow (#) to M'. As a
90 result, cells were partitioned into seven clusters, and any cluster with the pronounced heavy tail in the distribution of
91 per-cell total UMI counts observed in M was no longer apparent (Figure 2A, Supplementary Figure 2A, B).
92 Connectivity inference using PAGA further revealed a simpler, tree-like structure centered around cluster 2 as a hub
93 (Figure 2B). Moreover, differentially expressed gene analyses enabled biologically interpretable annotation of each
94 cluster, capturing several plausible local structures of monocyte state transitions consistent with prior experimental
95 and clinical work as follow (Figure 2C–G, Supplementary Figure 2C):

96 Annotations of each coarsely-discriminated cell state (CM: classical monocyte, IM: intermediate monocyte, NCM:
97 non-classical monocyte, ISG: interferon (IFN)-stimulated gene, moDC: monocyte-derived dendritic cell):

98 - cluster 0: IFN-stimulated CMs (*HLA genes*^{low}, *ISGs*^{high}, *CM markers*^{high})
99 - cluster 1: IM-NCM-lineage-primed CMs (*HLA genes*^{high}, *CM marker*^{high}, *NCM markers*^{low})
100 - cluster 2: HLA^{+} basal CMs (*HLA genes*^{moderate}, *CM markers*^{high})
101 - cluster 3: IMs (*HLA genes*^{high}, *CM markers*^{moderate}, *NCM markers*^{moderate})
102 - cluster 4: HLA$^{neg/low}$ CMs potentially immunosuppressive or exhausted (*HLA genes*$^{neg/low}$, *CM markers*^{high})
103 - cluster 5: NCMs (*HLA genes*^{moderate}, *NCM markers*^{high})
104 - cluster 6: moDCs (*HLA genes*^{high}, *DC markers*^{high})

Local structures consistent with prior work[9–14]:

- connection 4−0−2: IFNγ-mediated HLA induction during monocyte maturation and immune regulation
- connection 2−1−3−5: Differentiation to NCMs from CMs with temporal HLA upregulation at IMs
- connection 2−6: Differentiation to moDCs from CMs

These results suggest that graph abstraction using homogeneously-observed cells alone can uncover a low-dimensional manifold skeleton locally representing constrained cell state transitions that was obscured by heterogeneous observations.

Uncertainty in scRNA-seq data arising from limited observation depth has been widely discussed in the context of *dropout*, and numerous methods have been developed to impute or normalize expression profiles for genes that are not detected due to technical limitations. We next examined how the realistic levels of heterogeneous observations affects such correction procedures and results. We applied four commonly used imputation or normalization approaches: SCTransform[15], ALRA[16], SAVER[17], and scImpute[18] for the M, and applied the workflow (#) to the imputed or normalized gene expression data (Figure 2H–K). As observed in the uncorrected M (Figure 1C, D), the three-dimensional visualizations with UMAP embedding and total UMI counts showed that shallowly-observed cells were consistently confined to specific regions that were captured as distinct Louvain clusters with the heavy tail of total UMI count distribution across the four methods (Figure 2L–O, Supplementary Figure 2D–K). These results indicate that the heterogeneity of per-cell observation is difficult to correct by *dropout* imputation, because it intrinsically induces gene-wise differences in the expected distribution of UMI counts across cells.

Simulated cell population shows how heterogeneous observations generate spurious subclusters

To simulate the influences of heterogeneous observations on various interpretations generally applied to scRNA-seq data, we first modeled cDNA (UMI) intracellular distribution among genes using several empirical UMI count data[7,19,20]. We calculated $N_i(k)$, the number of genes with exactly k observed UMIs, for cell i and computed $\bar{N}(k)$, the mean of $N_i(k)$, at each depth range (Supplementary Figure 3A–C). As a result, though dataset-specific differences were observed in the detailed behavior of $\bar{N}(k)$ such as a heavy- or light-tailed $\bar{N}(k)$ distribution, three common tendencies emerged:

1. $\bar{N}(k)$ monotonically decreased as k increased.
2. For $1 \leq k \leq 10$, $\bar{N}(1)$ was much higher than those of the other ks, and $\bar{N}(k)$ decreased sharply.
3. $\bar{N}(k)$ decreased gradually for $k > 10$, especially in higher-depth ranges.

We next simulated gene-specific cDNA fractions in the cDNA pool that could reproduce the observed $\bar{N}(k)$ profiles under random sampling. For each gene, we denote its fraction in the cDNA pool as p, and assign gene index i ($i = 1, 2, \ldots, 36{,}601$) in ascending order of p. It has been estimated that the total number of genes expressed in a single cell is approximately $10{,}000$; therefore, we assumed that only genes with indices $i \geq 26{,}602$ can be expressed[21,22]. In addition, the simulation was conducted under the assumption that all UMIs are equally probable and a distinct UMI is captured without replacements for each gene.

Consequently, we constructed three expression distribution models in which p_i increases as a function of i in a

141 linear, exponential, or combined manner (Figure 3A, Supplementary Figure 3D, E). Under Poisson or negative
142 binomial (NB) sampling models with various dispersion parameters, we computed $\mathbb{E}[N(k)]$, expected $N(k)$, and
143 compared them with the empirical $\bar{N}(k)$, which indicated that the sharp decrease in $\bar{N}(k)$ for $1 \leq k \leq 10$ and
144 gradual decrease in $\bar{N}(k)$ for $k > 10$ could be simultaneously explained by the combined (both linear and
145 exponential) increase of p_i (Figure 3B, C, Supplementary Figure 3F–I). From these results, we assumed that the
146 expression distribution model, which we denote (*), with the combined increase of p_i at least could reproduce the
147 UMI distribution similar to the empirical $\bar{N}(k)$, and overdispersion could underlie the differences in the detailed
148 behaviors of $\bar{N}(k)$ across datasets.

149 To evaluate how heterogeneous observations affect the generally used workflow (#), we first generated a simulated
150 cell population X consisting of 2,000 Cells 1 according to the model (*), and 1,000 Cells 2 and 3 modeling
151 completely distinct cell-types from Cell 1 (Figure 3D). For each simulated cell, the total UMI counts were randomly
152 drawn as an integer from a normal distribution with mean 5,000 and standard deviation 1,500, and clamped at a
153 minimum of 500. When applying the workflow (#) to X, Cell 1, Cell 2, and Cell 3 were embedded as discrete groups
154 in UMAP, but Louvain clustering interpreted spatial dispersion in Cell 1 as subclusters (Figure 3E, F). The three-
155 dimensional visualization shows that the spatial dispersion in Cell 1 was associated with heterogeneous observation
156 depth (Figure 3G, H, Supplementary Figure 3J).

157 We next generated a simulated cell population Y (Figure 3I) consisting of 2,000 Cells 1 and 2,000 Cells 4
158 modeling a pathway activation. When the workflow (#) was applied to Y, Cell 1 and Cell 4 formed an apparently
159 continuous distribution in UMAP, and Louvain clustering produced artifactual intermediates between Cell 1 and Cell

4 (Figure 3J, K). The three-dimensional visualization shows that distance in UMAP between Cell 1 and Cell 4 decreased as observations became shallow (Figure 3L, Supplementary Figure 3K). These results suggest that heterogeneous observations can generate spurious subclusters within certain cell-type populations or artifactual intermediates between subtypes.

Simulated cell lineage shows how heterogeneous observations distort state transition inference

Such transcriptional changes simulated by Y are crucial for cellular differentiation and responses to environmental stimuli within tissues. Here, instead of assuming intrinsic continuous state transitions interpreted by *pseudotime*, we simulated step-wise cell state transitions and investigated the impact of heterogeneous observations on PAGA-inferred connectivity. To this end, we generated a simulated cell population Z with an assumed lineage: Cell A → Cell B (pathway activation) → Cell C (further pathway activation)→ Cell D or Cell E (epigenetic de-repression) (Figure 3M). We assume that there are 1,000 cells at each state.

When the workflow (#) and the three-dimensional visualization were applied to Z, the cells formed a continuous distribution in UMAP, and the distances among different states became progressively closer as observations became shallow, as observed in the previous simulations (Figure 3N, O). Louvain clustering yielded clusters in which shallowly-observed cells from different underlying states were intermixed (Figure 3P, Supplementary Figure 3L). When we generated connectivity graph between these clusters using PAGA, the intermixed clusters remained connected to multiple neighboring clusters, giving rise to an overly complex and artifactual loop-rich structure

178 (Figure 3Q).

179 To assess whether the simple correction by enforcing a depth threshold can improve lineage inference, we
180 generated Z' by subsetting cells from Z with total UMI counts $\geq 4{,}000$, and applied the workflow (#). In Z', no
181 spurious intermixed clusters were formed, and PAGA recovered strong connectivity specifically along the step-wise
182 transition paths consistent with the assumed lineage (Figure 3R–T). These results are consistent with the results from
183 the empirical PBMC scRNA-seq dataset and suggest the mechanism by which heterogeneous observations drive
184 graph abstraction away from the manifold structure dictated by underlying biological constraints.

185

186 **Topology-guided diagnostics preserve the coherence of manifold inference while minimizing**
187 **sample loss**

188 Filtering for deeply-observed cells to generate a subset with homogeneous observations can uncover a low-
189 dimensional manifold skeleton reflecting intrinsic paths of cell state transitions, but at the same time threatens the
190 premise that each attainable state is sampled densely enough. Given prior reports that per-cell depth can influence
191 PCA[23], the plausibility of the latent space generated by the workflow (#) under heterogeneous observations needs to
192 be examined explicitly. Moreover, when the workflow (#) is applied to all viable PBMCs, the distribution of total
193 UMI counts vary substantially across cell types, indicating that although heterogeneous observations can distort
194 manifold-based inference, simply thresholding cells by observation depth is biologically questionable because it
195 ignores genuine differences in transcriptional activity (Supplementary Figure 4).

We defined cells included in M' as high-information cells and the others filtered simply by observation depth as low-information cells, and evaluated PAGA connectivity among the annotated clusters in M' (Figure 2A) and low-information cell population using the PCs learned on M. In this latent space, though the annotated clusters were connected with low-information cell population and generated numerous independent loops, the connectivity among the annotated clusters themselves was largely consistent with that observed when PCA was applied to M' (Figure 2B, 4A). Indeed, when we applied downstream procedures of the workflow (#) to M' using the PCs learned on M, the low-dimensional manifold skeleton was remarkably well preserved (Figure 4B).

Intuitively, if a low-information cell lies close, in terms of geodesic distance, to high-information cells, then a random walk starting from this cell should reach the high-information cell subset within a small number of steps. Conversely, cells that remain among low-information cell population for many steps are expected to reside in regions where low-information cells are spuriously connected to one another. To formalize this, we performed k-step random walks ($k = 5$) from each low-information cell along the weighted neighbor graph learned on M and, over many trials ($1{,}000$ per cell), computed the probability of hitting any high-information cell within k steps, which we termed the hit rate. Consequently, cells with low hit rates were confined to specific regions in UMAP, and though the hit rate was positively correlated with total UMI counts, at the same time, cells with comparable total UMI counts exhibited a wide range of hit rates (Figure 4C, D). These observations suggest that, while the hit rate partially depends on observation depth, it also reflects geodesic proximity to high-information cell population on the data manifold and can serve as a quantitative indicator of the reliability of each low-information cell.

Furthermore, we used the hit rate to order low-information cells to remove them progressively from M, and applied

the workflow (#) and PAGA under the same rules governing certain percentage (e.g. 1%) of neighborhoods in whole population and certain resolution at Louvain clustering. As low-information cells were removed according to the hit rate, the first Betti number (β_1), which counts independent loops in a low-dimensional manifold skeleton, decreased and then plateaued when approximately half of the low-information cells had been discarded, after which further removal no longer reduced the number of loops (Figure 4E). These results, therefore, indicate that thresholding M at the boundary allowed us to minimize sample losses while preserving the low-dimensional coarse skeleton of data manifolds.

Discussion

Changes in cellular state arise from numerous intercellular interactions and intracellular signaling events, and single-cell omics invite us to view each snapshot of a tissue not as a static picture, but as the aggregate outcome of them unfolding over time. Within a given cell lineage, if the repertoire of attainable molecular states is sampled densely enough, the neighbor graph should, in principle, encode which states can be reached from which others under biological constraints. This motivates a manifold perspective in which cells with high dimensional feature vectors are regarded as samples from a low-dimensional cell-state manifold, and intrinsic cell–cell interactions, together with external perturbations such as drug exposure where present, manifest as constrained trajectories. A central problem is therefore to reconstruct, from finite samples, the geometry and coarse topology of this manifold.

Graph abstraction of single-cell data such as PAGA provides a practical approach to this problem. Constructing a

233 coarse-grained neighborhood graph between clusters can be interpreted as a low-dimensional skeleton of data
234 manifold; in particular, the topology of the PAGA graph—its connected components and loops—summarizes the
235 degrees of freedom in state transitions inferred from the data. Our results support the view that such graph-based
236 representations can, when built under appropriate conditions, capture aspects of the constrained degrees of freedom
237 in biology that have classically been described as differentiation paths. This is especially valuable in medicine, where
238 longitudinal sampling of the same tissue over time is often ethically or practically infeasible and we are effectively
239 forced to interpret snapshot data about the underlying dynamical system.

240 However, in biomedical omics the act of observation itself introduces a characteristic "ambiguity"[24]. In droplet-
241 based scRNA-seq, one of the most prominent sources of ambiguity is the strong heterogeneity in per-cell amounts of
242 observed molecules. As our computational simulations demonstrate, shallowly-observed cells tend to appear
243 ambiguous, in that they resemble multiple well-resolved states that are clearly separable among deeply-observed
244 cells. Such ambiguous cells act as *intermediate states* that create illusory bridges between clusters in graph abstraction.
245 In our monocyte example, when we restricted the analyses to the homogeneous observations from deeply-observed
246 cells alone, the PAGA graph formed several local structures consistent with prior experimental and clinical work
247 indicating HLA upregulation relevant to IFN stimulation[9,13,14] and differentiation toward IM-NCM lineage[11,12] or
248 moDC[10] from CMs, and also showed the sequence of these local structures. By contrast, under heterogeneous
249 observations, illusory loops generated by the spurious clusters made a low-dimensional manifold skeleton no longer
250 clearly identifiable.

251 This simple consequence of heterogeneous observations that shallowly-observed cells tend to connect to one

another and to multiple clusters in the neighborhood graph is not merely noise. This destabilizes the very notion of "local neighborhood" in the underlying metric space on which graph construction is based. In our analyses, imputation methods that rely on feature-space neighborhoods therefore show only limited benefit under realistic heterogeneous observations: when, at least locally, neighborhoods are already unreliable, propagating information along them can reinforce, rather than repair, the distortion of manifold inference. Restricting attention to sufficiently homogeneous observations can partially mitigate this issue. Yet, if a dataset already contains enough high-information cells to support a reliable neighborhood graph and a stable topology of graph abstraction, then using their observed expression values directly may be more consistent with the long-standing experimental ethos of biology that values raw measurements.

At the same time, our results do not imply that all structure in the data is hopelessly corrupted by the ambiguity due to heterogeneous observations. Prior work has shown that technical noise and *dropout* can substantially affect linear and non-linear compressions for low-dimensional embeddings[6,23,25–27]. Our analysis extends this view by showing that, even under realistic levels of observational ambiguity, certain global and topological aspects of the data manifold can remain relatively stable within subsets of cells that carry sufficient information (Figure 4A, B). Using topological descriptors of graph abstraction such as the first Betti number, we can explicitly flag regions where the inferred structure is likely driven by observational ambiguity rather than true dynamics, and consequently, carve out a trustworthy proxy for the underlying biological constraints.

Finally, we emphasize that these problems are not solely the responsibility of end-users of single-cell analysis pipelines. In interdisciplinary fields such as computational biology, methodological innovations are often packaged

and disseminated as broadly applicable solutions, while the underlying geometric and statistical assumptions—and the trade-offs they entail—are communicated only partially[28,29]. Our findings suggest that, for methods that operate on manifold approximations and neighborhood graphs, developers should explicitly articulate not only what additional "information" their models appear to recover, but also which assumptions about distances, neighborhoods and dynamics are being relaxed or violated in the process. The framework we introduce in this work is, of course, not exempt from these considerations. It presupposes that, within a given dataset, there exists a sufficiently large subset of high-information cells to anchor a faithful reconstruction which can be proxies locally for at least part of the underlying biological constraints. In this sense, it is less a recipe for extracting ever more information from single-cell data than a framework for deciding which biological narratives—namely, those concerning the degrees of freedom underlying differentiation and response to stimuli—we are prepared to trust.

Methods

Computational environment and package list

Computational simulations were conducted in Python 3.14.0 on macOS 15.5. Core numerical and data-handling libraries included NumPy (v2.3.5), pandas (v2.3.3), and SciPy (v1.16.3). Visualization was performed using Matplotlib (v3.10.7). HDF5-based input/output was handled through h5py (v3.15.1), and progress monitoring used tqdm (v4.67.1). To ensure reproducibility, all stochastic procedures were executed with fixed random seeds (SEED = 42) via both Python's random and NumPy's random number generators.

All R-based analyses were performed in R 4.5.0 on macOS 15.5. Core packages included Seurat (v5.3.1) and SeuratObject (v5.2.0) for single-cell analysis; tidyverse (v2.0.0; including ggplot2 v4.0.1, dplyr v1.1.4, tidyr v1.3.1, readr v2.1.6, purrr v1.2.0, tibble v3.3.0, stringr v1.6.0, and forcats v1.0.1) for data manipulation; plotly (v4.11.0), scales (v1.4.0), and htmlwidgets (v1.6.4) for visualization; hdf5r (v1.3.12) and SeuratDisk (v0.0.0.9021) for HDF5-based I/O and object conversion; and SeuratWrappers (v0.4.0) for auxiliary single-cell workflows. zellkonverter (v1.19.2) was used for interoperability between Seurat objects and AnnData. Python-based steps were accessed from R via reticulate (v1.44.1). Where applicable, imputation analyses used scImpute (v0.0.9) and SAVER (v1.1.2). To ensure reproducibility, we fixed the random seed in R (set.seed(42)). PAGA were executed from R using the reticulate interface. A dedicated conda environment named r-scanpy was created with Python 3.10, and core dependencies were installed from conda-forge (including scanpy, anndata, NumPy, SciPy, pandas, Matplotlib, h5py, umap-learn, python-igraph, leidenalg, and pynndescent).

Calculation of UMI distribution in empirical scRNA-seq datasets

Using the empirical UMI count matrix from real scRNA-seq datasets (in vivo, human: PBMC by 10x Genomics, $36{,}601$ genes $\times$ $11{,}996$ cells[7]; in vitro, human: lung progenitor organoids by Y. Miao *et al*, $36{,}601$ genes $\times$ $14{,}100$ cells[19]; mouse: complex sample by E. Mereu *et al*, $34{,}900$ genes $\times$ $10{,}000$ cells[20]), we calculated per-cell $N(k)$, the number of genes with exactly k observed UMIs. We then stratified cells by their total UMI counts into bins of width $1{,}000$ and computed $\overline{N}(k)$, the mean of $N(k)$, for each bin. For each dataset, we used the publicly available filtered UMI count matrix. The source of each dataset is in the **Data Availability**.

Manifold-based single cell analysis of empirical scRNA-seq data

In general, a workflow (#) has been used to analyze scRNA-seq data:

I. Log1p normalization of gene expression in each cell based on total UMI counts.

II. Identification of highly variable genes (HVGs).

III. Scaling of gene expression across cells (mean 0, variance 1).

IV. Principal component analysis (PCA).

V. Neighbor graph construction, clustering, and UMAP embedding in the PCA space.

We first extracted viable cells with $\%MT < 15$, where denotes the percentage of UMIs mapping to mitochondrial genes, from the PBMC scRNA-seq UMI count matrix and applied the workflow (#) to obtain a coarse classification

317 of cell types. Based on differential expression analysis between clusters, we identified clusters showing monocyte-
318 like gene expression patterns, lymphocyte-like patterns, and other patterns (e.g., platelets).

319 We next re-applied the workflow (#) to the set of clusters representing monocyte characteristics to assess
320 contamination by mRNA derived from other cell types, i.e., potential doublets. Differential expression analysis
321 among the resulting clusters revealed that cell populations forming discrete "islands" on the UMAP showed evidence
322 of lymphocyte-derived gene expression; these cells were therefore excluded as doublets. The remaining cells were
323 defined as the putative monocyte population M. Up to this point, parameters such as the number of neighbors and the
324 number of principal components in the standard analysis were chosen within commonly used ranges. For all
325 subsequent analyses, we adopted the following rule: among all observed cells, we constructed a weighted graph by
326 connecting each cell to the top 1% of cells with the smallest Euclidean distance in a thirty-dimensional PCA space,
327 and we interpreted the resulting Louvain modules using a fixed resolution parameter of 1. All the parameters used in
328 each workflow (#) were deposited in the corresponding author T.T.'s GitHub Page (https://github.com/Tomo-
329 koko/Heterogeneous-single-cell-observation).

330 Next, we applied the workflow (#) to M to obtain a PCA embedding and a Euclidean distance–based neighborhood
331 graph. We visualized the distribution of total UMI counts in each cluster using violin plots, and additionally generated
332 three-dimensional dot plots in which the UMAP embedding provided the x–y coordinates and total UMI counts were
333 mapped to the z-axis, in order to examine whether total UMI counts were unequally distributed within the UMAP-
334 derived low-dimensional manifold. To evaluate inter-cluster connectivity when interpreting Louvain clusters as a
335 coarse-grained partitioning of cell states on the manifold, we applied partition-based graph abstraction (PAGA)[8] to

M. The neighborhood graph used by PAGA was constructed with the same settings as in the workflow (#). Throughout this study, the PAGA connectivity threshold was fixed at 0.05; edges were drawn only between clusters whose connectivity exceeded this threshold, and the resulting graph was visualized as an overlay on the UMAP embedding and used for subsequent topological analyses.

Reconstruction of data manifold using high-information cells in empirical scRNA-seq data

M is a cell population with heterogeneous observation depth, that is, a mixture of shallowly- and deeply-observed cells. Among these, we defined high-information cells as those observed at sufficiently high depth and analyzed the data manifold in M', the subset consisting only of high-information cells. Note that the threshold defining M' is determined relative to the UMI count distribution of the parent population M and is not guaranteed to serve as a generally applicable threshold across all datasets. In this study, we set a threshold of $10{,}000$ total UMIs so as to include approximately the top 40% of cells, extracted high-information cells with total UMI counts $> 10{,}000$, and defined this subset as M'.

We then applied the workflow (#) to M' and, as for M, constructed a neighborhood graph and performed clustering to obtain a coarse-grained partitioning of cell states on the data manifold. After confirming that the localized enrichment of shallowly-observed cells seen in the UMAP of M was no longer apparent, we assigned biological annotations to clusters based on differential expression analysis and the expression of representative marker gene sets (e.g., classical monocyte (CM) marker genes, non-classical monocyte (NCM) marker genes, HLA-related genes,

interferon-stimulated genes (ISGs), and dendritic cell (DC) marker genes). Finally, using PAGA, we evaluated connectivity among the cell-state partitions obtained in M' and examined whether the local structure of the low-dimensional skeleton of the data manifold reflects phenomena that are supported by known biology.

Imputation and normalization in scRNA-seq analysis

In scRNA-seq analysis, observed zeros are often regarded as a mixture of technical zeros (e.g., failures of cDNA reverse transcription/amplification or shallow sequencing) and biological zeros (e.g., true absence of expression), collectively referred to as *dropout*. Numerous normalization and imputation methods have been developed to address such technical zeros using neighborhood graphs, and we employed representative methods—SCTransform[15], ALRA[16], SAVER[17], and scImpute[18]—to test whether compensating for dropout can correct distortions of the data manifold caused by heterogeneous observation depth in scRNA-seq. UMI count data were used as input, and analyses were performed on the 4,410 cells contained in M. Parameters for each imputation method were set to their default values unless otherwise specified. Downstream analyses using the workflow (#) were then conducted on the normalized/imputed data generated by each method. For dot plots visualizing the three-dimensional distribution with UMAP coordinates on the x–y plane and total UMI counts on the z-axis, the z-axis corresponded to total UMI counts from the raw (pre-normalization/imputation) count data. For scImpute, we specified seven clusters, corresponding to the number of monocyte cell states that appeared appropriate based on analyses restricted to M'.

Simulation of expression distribution models

We considered a typical human gene panel comprising 36,601 genes. For each simulated cell, we defined a cDNA pool derived from its mRNA molecules and assigned gene-specific cDNA fractions as follows. Let p denote the fraction of gene-derived cDNA in the pool, and order genes in ascending order of p, assigning gene IDs i accordingly. The fraction for gene i is denoted p_i.

We next considered how p_i increase as gene IDs increase; a linear increase model, an exponential increase model, and a combinational increase model of p_i. Given that mRNA molecules, as the source of cDNA, are discrete physical entities with non-negative integer counts per gene and a total per cell on the order of 10^5 molecules, we computed p_i under this constraint. The simulation was performed under the assumption that all UMIs are equally probable and that each captured molecule is labeled with a distinct UMI. Based on previous estimates that the total number of genes expressed in a single cell is $\sim 10{,}000$, we further assumed that only genes with indices $i \geq 26{,}602$ can be expressed (i.e. $p_i \geq 0$ for $i \geq 26{,}602$ and $p_i = 0$ otherwise)[21,22]. p_i were calculated as below under each model assumption:

- Linear increase model:

$$p_i = \left\lfloor \frac{\max(0, i - 26{,}301)}{300} \right\rfloor, \tag{1}$$

where $\lfloor X \rfloor$ denotes a maximum integer which does not exceed X.

- Exponential increase model:

$$p_i = \left\lfloor \exp\left(\frac{i - 36{,}000}{75}\right)\right\rfloor. \tag{2}$$

- Combinational increase model

$$p_i = \left\lfloor \frac{\max(0, i - 26{,}301)}{300}\right\rfloor + \left\lfloor \exp\left(\frac{i - 36{,}000}{75}\right)\right\rfloor. \tag{3}$$

Under a Poisson sampling model for UMIs drawn from the cDNA pool, X_i, the observed UMI count for gene i and $\mathbb{E}[N(k)]$, the expected $N(k)$ are given by these equations, respectively:

$$X_i \sim Poisson(\lambda_i),\ \ \lambda_i = Dp_i,\ \ P(X_i = k) = \frac{e^{-\lambda_i}{\lambda_i}^k}{k!}, \tag{4}$$

$$\mathbb{E}[N(k)] = \sum_i P(X_i = k) = \sum_i \frac{e^{-\lambda_i}{\lambda_i}^k}{k!}, \tag{5}$$

where D denotes the total number of sampled UMIs.

To incorporate overdispersion assumed in scRNA-seq data, we also simulated $\mathbb{E}[N(k)]$ under a negative binomial (NB) sampling model:

$$X_i \sim NB(\mu_i, \theta),\ \ \mu_i = Dp_i,\ \ P(X_i = k) = \frac{\Gamma(k+\theta)}{\Gamma(\theta)k!}\left(\frac{\mu_i}{\mu_i+\theta}\right)^k\left(\frac{\theta}{\mu_i+\theta}\right)^\theta, \tag{6}$$

$$\mathbb{E}[N(k)] = \sum_i P(X_i = k) = \sum_i \frac{\Gamma(k+\theta)}{\Gamma(\theta)k!}\left(\frac{\mu_i}{\mu_i+\theta}\right)^k\left(\frac{\theta}{\mu_i+\theta}\right)^\theta, \tag{7}$$

where θ is a dispersion parameter, and computed $\mathbb{E}[N(k)]$ for $\theta = 0.1, 0.3, 0.5, 1, 3, 10$.

Simulation modeling cell-type discrimination, cell-state discrimination, and step-wise transition

Building on the model (*), we generated a simulated cell population X modeling cell-type discrimination as follows:

405 - Cell 1: 2,000 cells generated according to the model (*).

406 - Cell 2 and 3: 1,000 cells generated from a distribution obtained by randomly permuting the gene IDs of
407 model (*) with different random seeds.

408 Building on the model (*), we next generated a simulated cell population Y modeling transcriptomic changes such as
409 pathway activation as follows:

410 - Cell 1: 2,000 cells generated according to the model (*).

411 - Cell 4: 2,000 cells generated from Cell 1 by scaling the cDNA counts of 300 genes randomly selected from
412 gene IDs 26,602– 36,601 by either 2 or 0.5.

413 Building on the model (*), we further generated a simulated cell population Z modeling step-wise cell state transition
414 such as pathway activation and epigenetic de-repression with an assumed lineage as follows:

415 Assumed lineage: A → B (pathway activation) → C (further pathway activation)→ D or E (epigenetic de-repression)

416 - Cell A: 1,000 cells generated according to the model (*).

417 - Cell B: 1,000 cells generated from Cell A by scaling the cDNA counts of 300 genes randomly selected from
418 gene IDs 26,602– 36,601 by either 1.67 or 0.6, modeling signaling activation.

419 - Cell C: 1,000 cells generated from Cell B by further scaling the same set of 300 genes by either 1.67 or 0.6.

420 - Cell D and E: 1,000 cells generated from Cell C by selecting 10 genes randomly from gene IDs 1– 26,601
421 and increasing their cDNA molecule counts in the pool from 0 to 50 with different random seeds, modeling

epigenetic de-repression.

For each simulated cell, the total UMI counts were randomly drawn as an integer from a normal distribution with mean $5{,}000$ and standard deviation $1{,}500$, and clamped at a minimum of 500. We analyzed the simulated populations by the workflow (#), the three-dimensional visualization, and PAGA.

Topology-guided thresholding of data manifolds based on random walk metrics

When attempting to recover a data manifold that faithfully reflects biological constraints under heterogeneous observation depth, one inevitably faces the question of how to remove low-information cells. Simple thresholding by depth ignores biologically meaningful differences in transcriptional activity across cell types, so a quantitative metric is needed to evaluate low-information cells using the data manifold reconstructed from high-information cells as an anchor.

First, we transferred the annotations defined on the data manifold of M' (e.g., HLA^+ basal CMs) back to M, and labeled all cells that had been excluded from M' as "low-information cells." We then used PAGA to evaluate connectivity between each annotation and the low-information cells in the PCA space of M generated by workflow (#). Next, while keeping this PCA embedding fixed, we restricted the graph to high-information cells and again assessed inter-annotation connectivity using PAGA. These analyses confirmed that, in M, the low-dimensional skeleton generated by high-information cells is preserved even in the presence of low-information cells.

Next, to quantify how close each low-information cell lies to the data manifold spanned by high-information cells—that is, how easily it can "escape" the low-information cell population by following the neighbor graph—we considered k-step random walks on the weighted neighbor graph in PCA space, initiated from each low-information cell. We defined the Hit rate as the proportion of random-walk trials in which a k-step walk starting from a low-information cell reached any high-information cell at least once. In this study, we set $k = 5$ and performed $1{,}000$ trials per low-information cell to compute the Hit rate for all low-information cells.

Finally, we identified low-information cells falling into the bottom 10%, 20%, …, 90% of the Hit rate distribution, progressively removed these cells, and examined how the topological structure of the data manifold produced by workflow (#) and PAGA changed in terms of the number of clusters (nodes), the zeroth Betti number (β_0), and the first Betti number (β_1). By determining the Hit rate threshold at which the behavior of β_1 shifts—separating a loop-dominated regime from one dominated by tree-like structures—we obtained a threshold for low-information cells that does not directly depend on depth but is instead grounded in the topological properties of the data manifold.

Research integrity including usage of GPT models

In this study, we primarily used the generative model ChatGPT as an auxiliary tool for portions of the R and Python coding and for drafting the manuscript. All code and text generated by the model were reviewed and verified by the corresponding author, and, to further safeguard research integrity, we screened the manuscript for textual similarity against internet-available literature using Turnitin. All schematic representations in this manuscript were created

using BioRender and are published under an academic publishing license; the authors did not use generative AI for the creation of any schematic figures.

Acknowledgements

This work was supported by Keio University Grant-in-Aid for Encouragement of Young Medical Scientists (to T.T.), Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number JP25K21344 (to T.I.) and the Ishii-Ishibashi Foundation. T.T. was supported by Takeda Science Foundation Scholarship, and several scholarships in Keio University Graduate School of Medicine. The authors thank to Ms. M. Abe and the members of Department of Extended Intelligence for Medicine for dedicated administrative supports. The author T.T. also thanks to Dr. M. Uchihara in Japan Institute for Health Security and Dr. K. Sekine and Mr. K. Maekawa in National Cancer Center Research Institute for useful advices.

Authors Contribution

T.T. conceived and supervised the project. T.T. and Y.Y. designed the computational framework, performed the simulations and statistical analyses, and interpreted the results. Y.O. and T.I. performed validations of the results. T.T. drafted the manuscript, and Y.Y., Y.O., T.I., and K.S. contributed to revision and refinement of the text. All authors approved the final version of the manuscript.

Competing Interests

The authors declare no competing financial or non-financial competing interests in terms of this research.

Data Availability

This study is based on previously published datasets. The human PBMC scRNA-seq dataset generated by 10x Genomics is available from 10x Genomics Web Page by Sample ID “10k_PBMC_3p_nextgem_Chromium_X”. The human lung progenitor organoid scRNA-seq dataset by Miao *et al*. is available from GSE250399. The mouse scRNA-seq dataset by E. Mereu *et al*. is available from GSE133535. No new sequencing data were generated in this study.

Code Availability (including supplementary information)

The source codes and supplementary files related to this research at this preprint version were deposited in the corresponding author T.T.’s GitHub Page (https://github.com/Tomo-koko/Heterogeneous-single-cell-observation). After a future peer-review, we will open updated supplementary information as soon as possible.

Figure Legends

Figure 1. Low-information cells confined to specific regions form spurious hub-like connectivity in graph-based abstraction. A For viable cells with $< 15\%$ mitochondrial RNA content in a PBMC scRNA-seq dataset obtained from 10x Genomics Web Page by Sample ID "10k_PBMC_3p_nextgem_Chromium_X", we applied a standard analysis workflow (#) consisting of log1p normalization, highly variable gene (HVG) selection, scaling, principal component analysis (PCA), nearest-neighbor graph construction, Louvain clustering, and UMAP embedding. The UMAP plot is shown with the resulting cluster annotations. Among these, clusters 0, 4, 5, 10, 12, 13, and 17 (red arrows) exhibit monocyte-like gene expression profiles and form a continuous distribution in UMAP space. The remaining clusters correspond to plausible peripheral blood populations such as lymphocytes and platelets; **B** After removing samples suspected of representing technical contamination by mRNA derived from multiple cells, we re-applied the workflow (#) to the putative monocyte population M. The UMAP plot is shown with the resulting cluster annotations; **C** Violin plot of total UMI counts per cell in each cluster. Cluster 1 (red arrow) is particularly enriched for shallowly-observed cells and exhibits a characteristic heavy-tailed distribution; **D** Dot plots visualizing the three-dimensional distribution of the putative monocyte population, with UMAP coordinates on the x–y plane and total UMI counts on the z-axis. Panels show views highlighting clusters 0, 1, and 3, and clusters 2, 4, 6, and 7, respectively. Shallowly-observed cells in cluster 1 (red arrow) are particularly confined to a specific region of the UMAP manifold; **E** UMAP embedding of M overlaid with partition-based graph abstraction (PAGA) edges representing inter-cluster connectivity. Only edges with connectivity ≥ 0.05 are shown, and the intensity of each line segment reflects the strength of connectivity. Cluster 1 (red arrow), which is predominantly composed of shallowly-

506 observed cells, shows strong connectivity with many other clusters; Interactive HTML-based visualization tools for

507 all three-dimensional plots are provided as **supplementary data**.

508

509 **Figure 2. Graph-based data abstraction using high-information cells alone generates plausible inference of cell**

510 **state transition. A** From the putative monocyte population M, we extracted high-information cells defined by total

511 UMI counts $\geq 10{,}000$ to obtain subset M', and applied the workflow (#) to M'. The UMAP plot is shown with the

512 resulting cluster annotations; **B** UMAP embedding of M' overlaid with PAGA edges representing inter-cluster

513 connectivity. Only edges with connectivity ≥ 0.05 are shown, and the intensity of each line segment reflects the

514 strength of connectivity; **C** Feature plots showing the expression of classical monocyte (CM) marker genes *CD14*,

515 *VCAN*, *S100A8*, *S100A9*, *CCR2*, and *LYZ* in M'. The color bar indicates normalized expression levels; **D** Feature

516 plots showing the expression of non-classical monocyte (NCM) marker genes *FCGR3A*, *SIGLEC10*, *NR4A1*, and

517 *CX3CR1* in M'. The color bar indicates normalized expression levels; **E** Feature plots showing the expression of

518 monocyte-derived dendritic cell (moDC) marker genes *CD1E* and *CLEC10A* in M'. The color bar indicates

519 normalized expression levels; **F** Feature plots showing the expression of HLA-related genes *HLA-DPA1*, *HLA-DQA1*,

520 *HLA-DRA*, *HLA-DMA*, *HLA-DPB1*, *HLA-DRB1*, *CD74*, and *CIITA* in M'. The color bar indicates normalized

521 expression levels; **G** Feature plots showing the expression of IFN-stimulated genes (ISGs) *IFIT1*, *IFIT2*, *OASL*, and

522 *MX1* in M'. The color bar indicates normalized expression levels; **H** After normalizing gene expression profiles in M

523 using SCTransform, we applied the workflow (#). The UMAP plot is shown with the resulting cluster annotations; **I**

524 After imputing *dropouts* in M using ALRA, we applied the workflow (#). The UMAP plot is shown with the resulting

525 cluster annotations; **J** After imputing *dropouts* in M using SAVER, we applied the workflow (#). The UMAP plot is

526 shown with the resulting cluster annotations; **K** After imputing *dropouts* in M using scImpute with a parameter $k =$

527 7 which is the same number of clusters as stratified in M', we applied the workflow (#). The UMAP plot is shown

528 with the resulting cluster annotations; **L** Dot plots visualizing the three-dimensional distribution of the SCTransform-

529 normalized M, with UMAP coordinates on the x–y plane and total UMI counts on the z-axis; **M** Dot plots visualizing

530 the three-dimensional distribution of the ALRA-imputed M, with UMAP coordinates on the x–y plane and total UMI

531 counts on the z-axis; **N** Dot plots visualizing the three-dimensional distribution of the SAVER-imputed M, with

532 UMAP coordinates on the x–y plane and total UMI counts on the z-axis; **O** Dot plots visualizing the three-

533 dimensional distribution of the scImpute-imputed M, with UMAP coordinates on the x–y plane and total UMI counts

534 on the z-axis; Interactive HTML-based visualization tools for all three-dimensional plots are provided as

535 **supplementary data**.

536

537 **Figure 3. Simulations demonstrate how heterogeneous single-cell observation distorts cell-state discrimination**

538 **and state trajectory inference. A** A mixed (linear + exponential) distribution model for the composition ratio p of

539 gene-derived products in the cDNA pool. The number of genes is set to $36{,}601$, corresponding to a typical gene panel.

540 For each gene, we model the molecule count as an integer so that the total number of molecules is on the order of

541 10^5, and we use the values normalized by the total molecule count as p. Genes are indexed such that $p_1 \leq p_2 \leq \cdots \leq$

542 $p_{36{,}601}$. It is generally considered that approximately $10{,}000$ genes are expressed in a single cell; in this model, gene

543 expression is allowed only for $i \geq 26{,}601$. For $26{,}602 \leq i \leq 36000$, p_i increases linearly, and for $i \geq 36{,}000$, an

544 additional term representing exponential increase is applied; **B** Line plot showing the expected number of genes
545 $N(k)$ with exactly k observed UMIs under a Poisson sampling model. For each gene, all UMIs are assumed to be
546 sampled with equal probability, and a different UMI is selected at each draw; **C** Line plots showing the expected
547 $N(k)$ under a negative binomial sampling model. Expectations are computed for the overdispersion parameter $\theta =$
548 0.1, 0.3, 0.5, 1, 3, and 10; **D** Schematic illustration of simulations modeling different cell-type discrimination in a
549 simulated cell population X. Created with BioRender; **E** UMAP embedding of cell population X after applying the
550 workflow (#), annotated with ground-truth cell-type labels; **F** UMAP embedding of cell population X after applying
551 the workflow (#), annotated with clusters obtained by Louvain clustering; **G** Violin plots of total UMI counts per cell
552 for each Louvain cluster in cell population X. Among clusters 3, 4, and 5, which are composed of observation samples
553 from Cell 1, cluster 4 (red arrow) contains many low-information cells and exhibits a pronounced heavy-tailed
554 distribution; **H** Dot plots visualizing the three-dimensional distribution of cell population X, with UMAP coordinates
555 on the x–y plane and total UMI counts on the z-axis; **I** Schematic illustration of simulations modeling transcriptional
556 changes (e.g., pathway activation) in a simulated cell population Y. Created with BioRender; **J** UMAP embedding of
557 cell population Y after applying the workflow (#), annotated with ground-truth cell-type labels; **K** UMAP embedding
558 of cell population Y after applying the workflow (#), annotated with clusters obtained by Louvain clustering; **L** Dot
559 plots visualizing the three-dimensional distribution of cell population Y, with UMAP coordinates on the x–y plane
560 and total UMI counts on the z-axis; **M** Schematic illustration of simulations modeling step-wise cell state transitions
561 assuming pathway activation and epigenetic de-repression in a simulated cell population Z. Created with BioRender;
562 **N** UMAP embedding of cell population Z after applying the workflow (#), annotated with ground-truth cell-type

labels; **O** UMAP embedding of cell population Z after applying the workflow (#), annotated with clusters obtained by Louvain clustering; **P** Dot plots visualizing the three-dimensional distribution of cell population Z, with UMAP coordinates on the x–y plane and total UMI counts on the z-axis; **Q** UMAP embedding of Z overlaid with PAGA edges representing inter-cluster connectivity. Only edges with connectivity ≥ 0.05 are shown, and the intensity of each line segment reflects the strength of connectivity. Cluster 2 (red arrow), which is predominantly composed of shallowly-observed cells, shows strong connectivity with many other clusters; **R** From the putative monocyte population Z, we extracted high-information cells defined by total UMI counts $\geq 5{,}000$ to obtain subset Z', and applied the workflow (#) to Z'. The UMAP plot is shown with the ground-truth cell-type labels; **S** UMAP embedding of cell population Z' after applying the workflow (#), annotated with clusters obtained by Louvain clustering; **T** UMAP embedding of Z' overlaid with PAGA edges representing inter-cluster connectivity. Only edges with connectivity ≥ 0.05 are shown, and the intensity of each line segment reflects the strength of connectivity; Interactive HTML-based visualization tools for all three-dimensional plots are provided as **supplementary data**.

Figure 4. Topological descriptors of the data-manifold skeleton resolve the trade-off between preserving low-dimensional structure and minimizing sample loss. A UMAP embedding of M overlaid with PAGA edges representing connectivity between annotated clusters inferred from the analysis of high-information cell population M' alone ($HLA^{neg/low}$ CMs potentially immunosuppressive or exhausted, IFN-stimulated CMs, HLA^{+} basal CMs, IM-NCM-lineage-primed CMs, IMs, NCMs, moDCs) and low-information cells defined by total UMI counts $< 10{,}000$. Only edges with connectivity ≥ 0.05 are shown, and the intensity of each line segment reflects the strength of

connectivity. Cluster 1 (red arrow), which is predominantly composed of shallowly-observed cells, shows strong connectivity with many other clusters; **B** UMAP embedding of M' which is computed by PCs calculated in the presence of low-information cells overlaid with PAGA edges representing connectivity between the annotated clusters. Only edges with connectivity ≥ 0.05 are shown, and the intensity of each line segment reflects the strength of connectivity; **C** To quantitatively assess, among low-information cells, those located near the data manifold spanned by high-information cells, we computed a hit rate to high-information cells on the neighbor graph of M constructed by the workflow (#). For each low-information cell, we performed $1{,}000$ k-step weighted random walks following the graph and calculated the proportion of walks that reached any high-information cell (Hit rate). The figure shows a feature plot of the Hit rate for each low-information cell at $k = 5$, with the color bar indicating the Hit rate. High-information cells are shown as gray dots; **D** Scatter plot of Hit rate versus total UMI counts for low-information cells. The x-axis shows total UMI counts and the y-axis shows the Hit rate at $k = 5$; the monotonic association between the two variables was assessed using Spearman's rank correlation (ρ and P value); **E** Changes in the topological properties of the low-dimensional skeleton as cells with decreasing reliability are progressively removed according to their Hit rate. From left to right, UMAP embeddings and PAGA graphs (connectivity threshold $= 0.05$) are shown for populations with no removal (same condition as **Figure 1E**) and after removing the bottom 10%, 20%, 30%, …, and 100% of cells ranked by Hit rate (100% corresponding to the condition in **Figure 2B**). For each step, we report the number of clusters in the PAGA graph and the number of independent loops (first Betti number, β_1). The resolution parameter for Louvain clustering on the neighbor graph was fixed at 1, and the number of neighbors was set to 1% of all cells; all topological properties were compared under these conditions.

601 # References

602 1. Rozenblatt-Rosen, O. *et al.* The Human Tumor Atlas Network: Charting Tumor Transitions across Space
603 and Time at Single-Cell Resolution. *Cell* 181, 236–249 (2020).

604 2. Lederer, A. R. *et al.* Statistical inference with a manifold-constrained RNA velocity model uncovers cell
605 cycle speed modulations. *Nat Methods* 21, 2271–2286 (2024).

606 3. Ding, J., Sharon, N. & Bar-Joseph, Z. Temporal modelling using single-cell transcriptomics. *Nat Rev Genet*
607 23, 355–368 (2022).

608 4. Bhuva, D. D. *et al.* Library size confounds biology in spatial transcriptomics data. *Genome Biol* 25, (2024).

609 5. Jiang, R., Sun, T., Song, D. & Li, J. J. Statistics or biology: the zero-inflation controversy about scRNA-seq
610 data. *Genome Biol* 23, (2022).

611 6. Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nat Rev Genet* 24, 550–572
612 (2023).

613 7. 10x Genomics. 10k Human PBMCs, 3' v3.1, Chromium X. *https://www.10xgenomics.com/datasets/10k-*
614 *human-pbmcs-3-ht-v3-1-chromium-x-3-1-high* (2021).

615 8. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a
616 topology preserving map of single cells. *Genome Biol* 20, 1–9 (2019).

617 9. Mengos, A. E., Gastineau, D. A. & Gustafson, M. P. The CD14+HLA-DrlO/NEG monocyte: An

618 immunosuppressive phenotype that restrains responses to cancer immunotherapy. *Front Immunol* 10,
619 (2019).

620 10. Boyette, L. B. *et al.* Phenotype, function, and differentiation potential of human monocyte subsets. *PLoS*
621 *One* 12, (2017).

622 11. Patel, A. A. *et al.* The fate and lifespan of human monocyte subsets in steady state and systemic
623 inflammation. *Journal of Experimental Medicine* 214, 1913–1923 (2017).

624 12. Williams, H. *et al.* Monocyte Differentiation and Heterogeneity: Inter-Subset and Interindividual
625 Differences. *Int J Mol Sci* 24, (2023).

626 13. de Metz, J. *et al.* Interferon-increases monocyte HLA-DR expression without effects on glucose and fat
627 metabolism in postoperative patients. *J Appl Physiol* 96, 597–603 (2004).

628 14. Virelizier, J. -L, Perez, N., Arenzana-Seisdedos, F. & Devos, R. Pure interferon gamma enhances class II
629 HLA antigens on human monocyte cell lines. *Eur J Immunol* 14, 106–108 (1984).

630 15. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using
631 regularized negative binomial regression. *Genome Biol* 20, (2019).

632 16. Linderman, G. C. *et al.* Zero-preserving imputation of single-cell RNA-seq data. *Nat Commun* 13, (2022).

633 17. Huang, M. *et al.* SAVER: Gene expression recovery for single-cell RNA sequencing. *Nat Methods* 15, 539–
634 542 (2018).

635 18. Li, W. V. & Li, J. J. An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat*
636 *Commun* 9, (2018).

637 19. Miao, Y. *et al.* Co-development of mesoderm and endoderm enables organotypic vascularization in lung and
638 gut organoids. *Cell* 188, 4295-4313.e27 (2025).

639 20. Mereu, E. *et al.* Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat Biotechnol*
640 38, 747–755 (2020).

641 21. Kuznetsov, V. A., Knott, G. D. & Bonner, R. F. General Statistics of Stochastic Process of Gene Expression
642 in Eukaryotic Cells. *Genetics* 161, 1321–1332 (2002).

643 22. Bengtsson, M., Ståhlberg, A., Rorsman, P. & Kubista, M. Gene expression profiling in single cells from the
644 pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res* 15, 1388–1392
645 (2005).

646 23. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for
647 single-cell RNA-Seq based on a multinomial model. *Genome Biol* 20, (2019).

648 24. Sarkar, A. & Stephens, M. Separating measurement and expression models clarifies confusion in single-cell
649 RNA sequencing analysis. *Nat Genet* 53, 770–777 (2021).

650 25. Bacher, R. & Kendziorski, C. Design and computational analysis of single-cell RNA-sequencing
651 experiments. *Genome Biol* 17, (2016).

652 26. Imoto, Y. *et al.* Resolution of the curse of dimensionality in single-cell RNA sequencing data analysis. *Life*
653 *Sci Alliance* 5, (2022).

654 27. Sun, T., Song, D., Li, W. V. & Li, J. J. scDesign2: a transparent simulator that generates high-fidelity single-
655 cell gene expression count data with gene correlations captured. *Genome Biol* 22, (2021).

656 28. Tritschler, S. *et al.* Concepts and limitations for learning developmental trajectories from single cell
657 genomics. *Development (Cambridge)* 146, (2019).

658 29. Herring, C. A., Chen, B., McKinley, E. T. & Lau, K. S. Single-Cell Computational Strategies for Lineage
659 Reconstruction in Tissue Systems. *CMGH* 5, 539–548 (2018).

660

Figure 1

*T. Tamura, *Y. Yamane, *et al.*

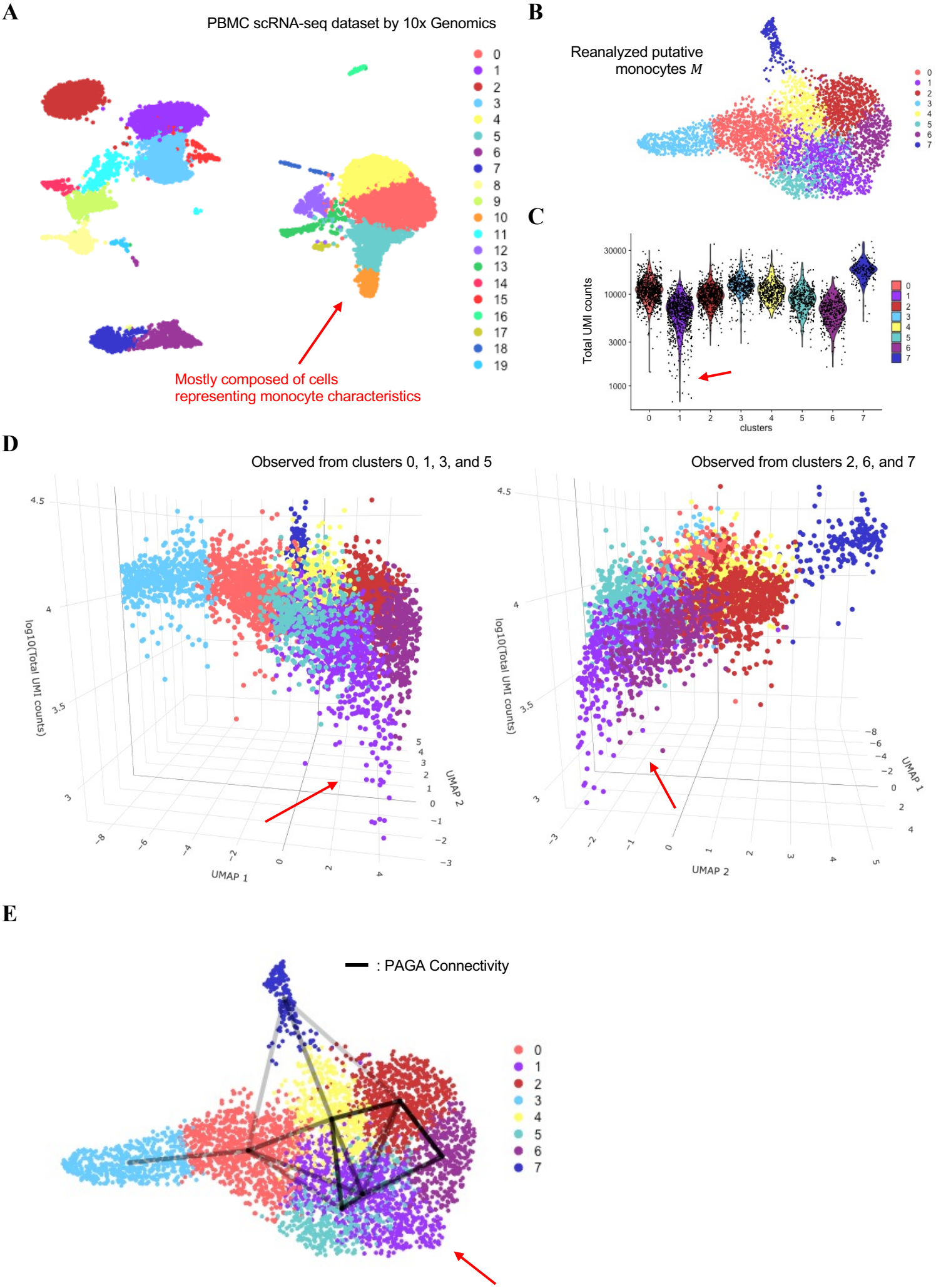


Figure 2 (Part 1)

*T. Tamura, *Y. Yamane, *et al.*

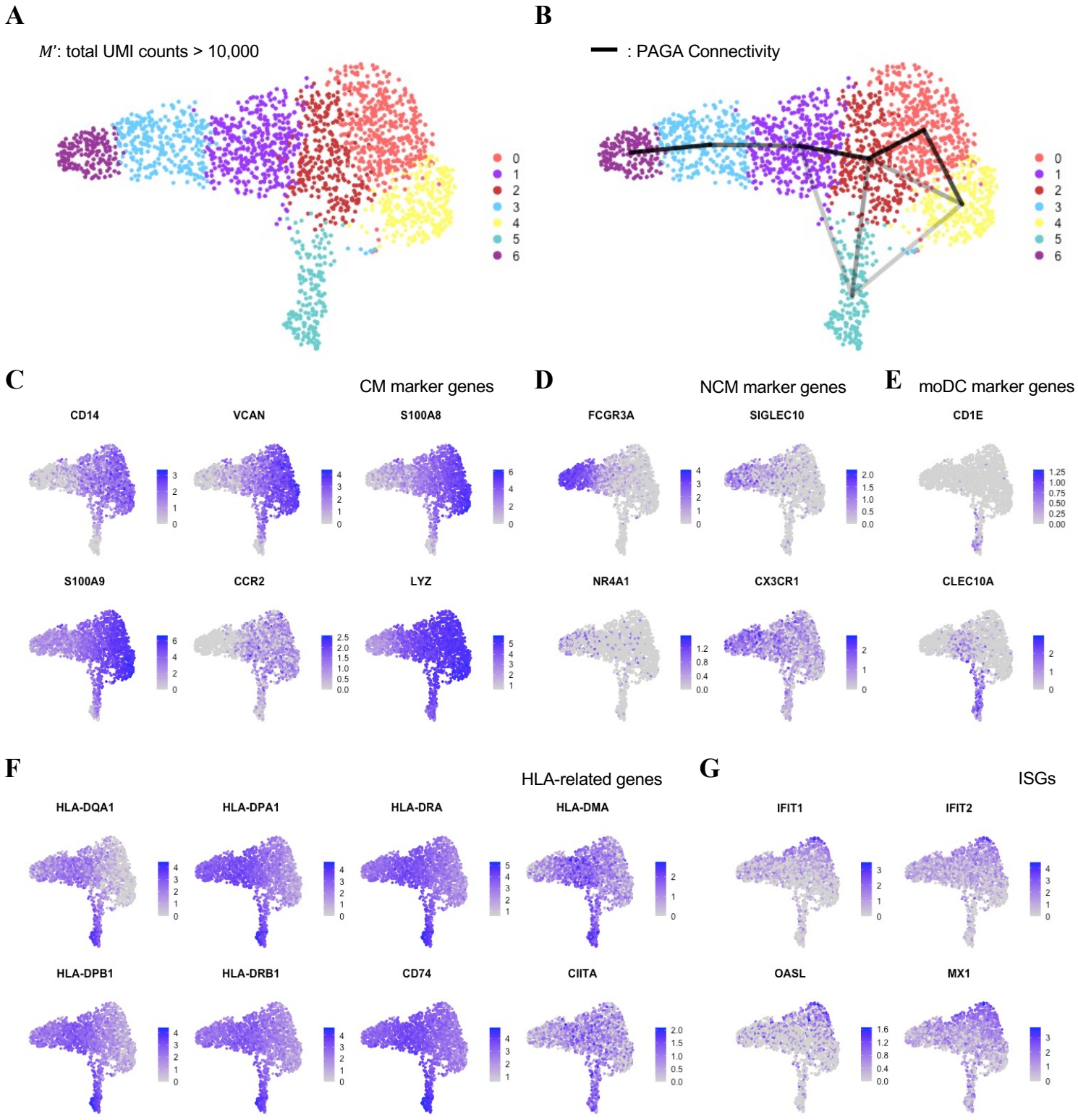


Figure 2 (Part 2, continued)

*T. Tamura, *Y. Yamane, *et al.*

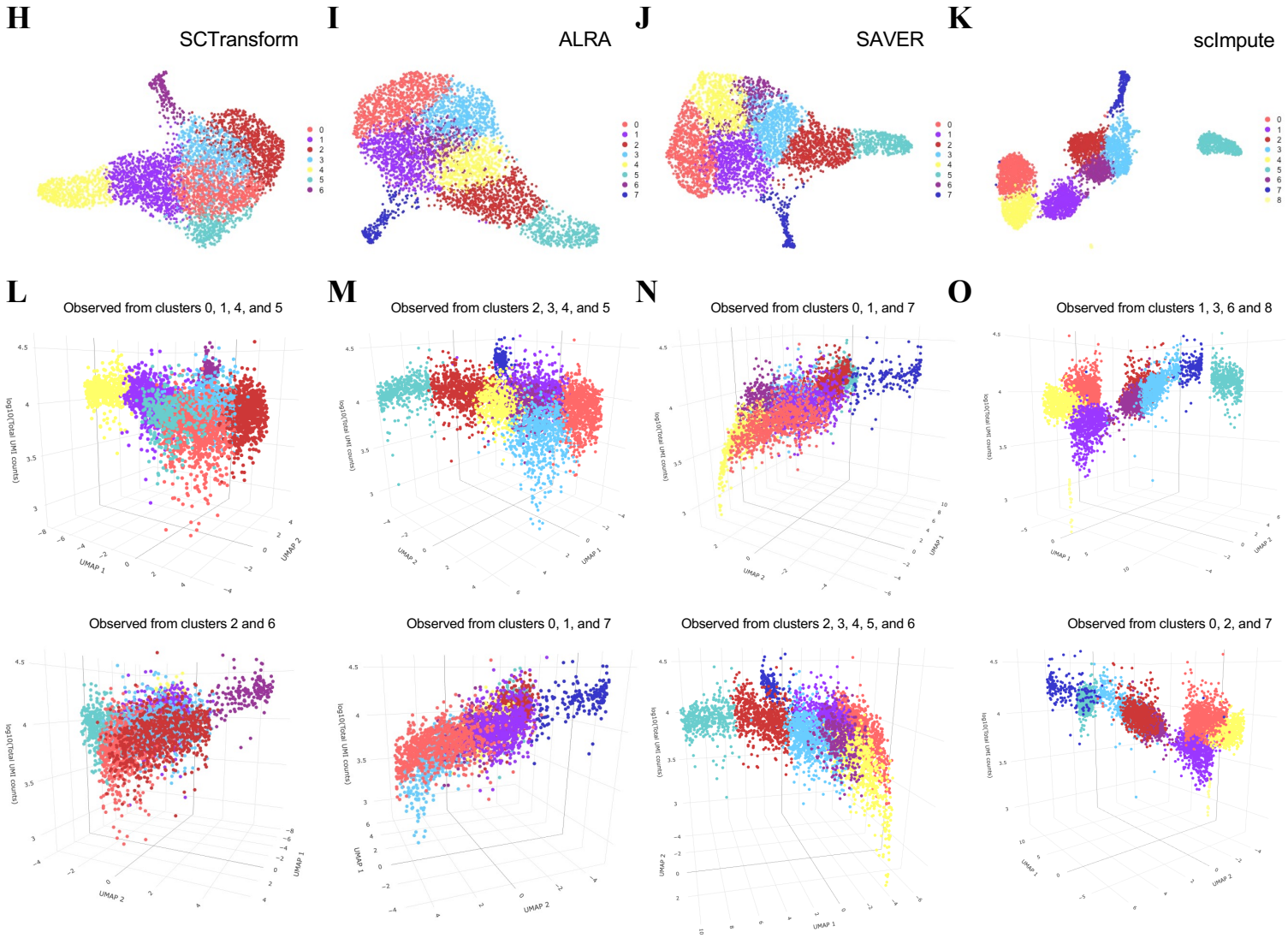


Figure 3 (Part 1)

*T. Tamura, *Y. Yamane, *et al.*

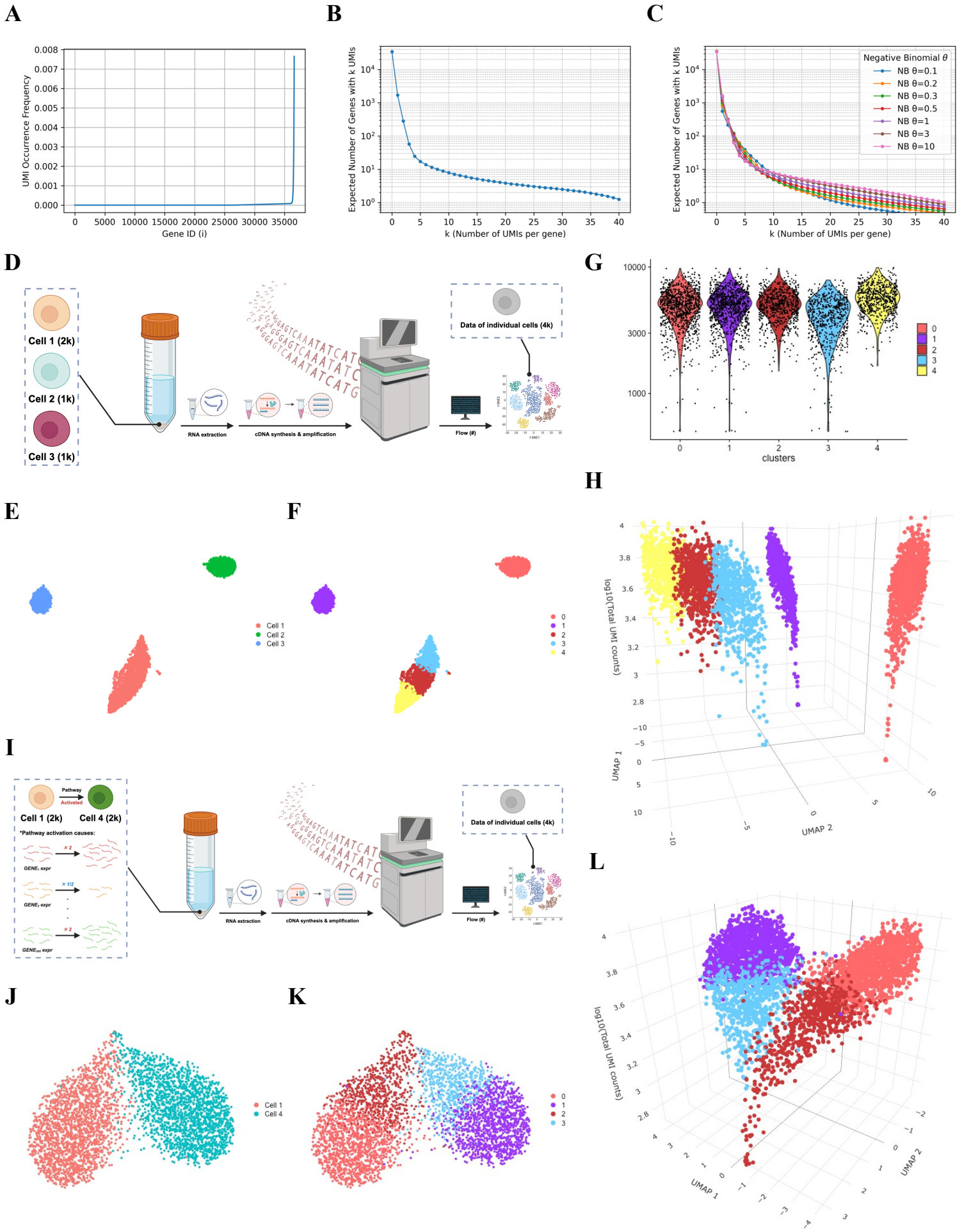
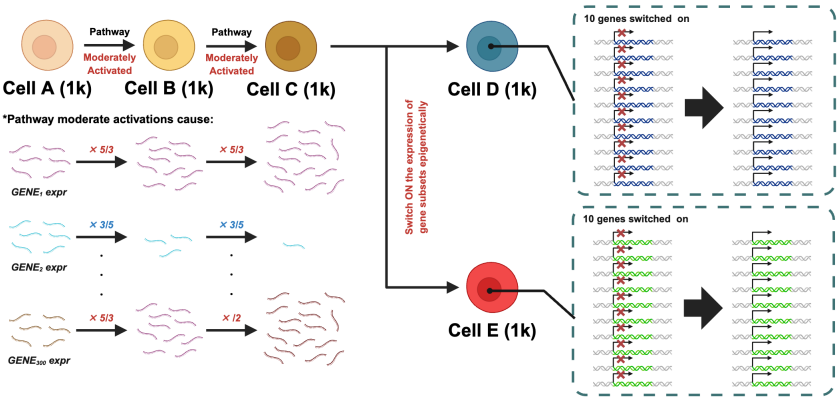


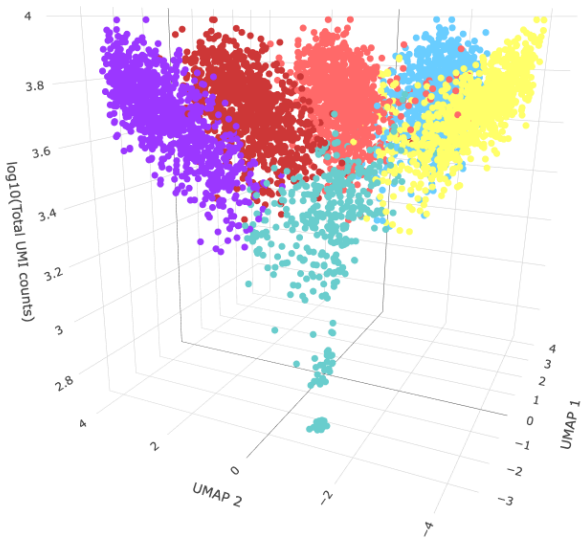
Figure 3 (Part 2, continued)

*T. Tamura, *Y. Yamane, *et al.*

M



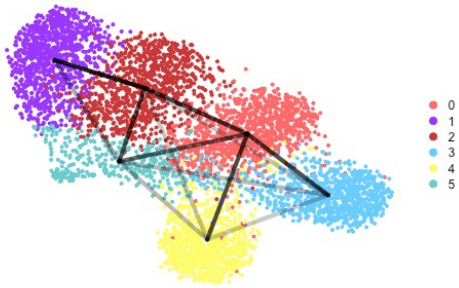
P



N

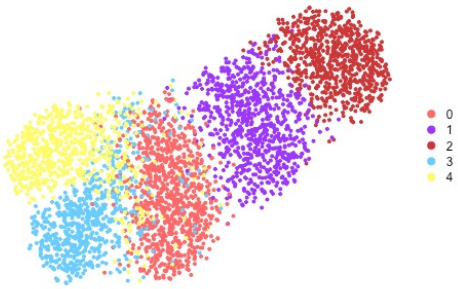
O

Q



R

S



T

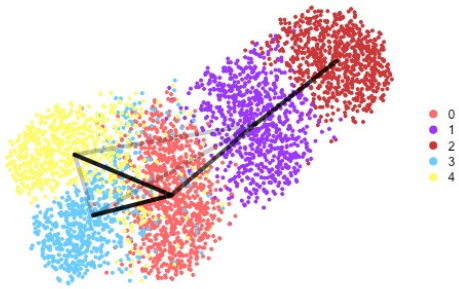
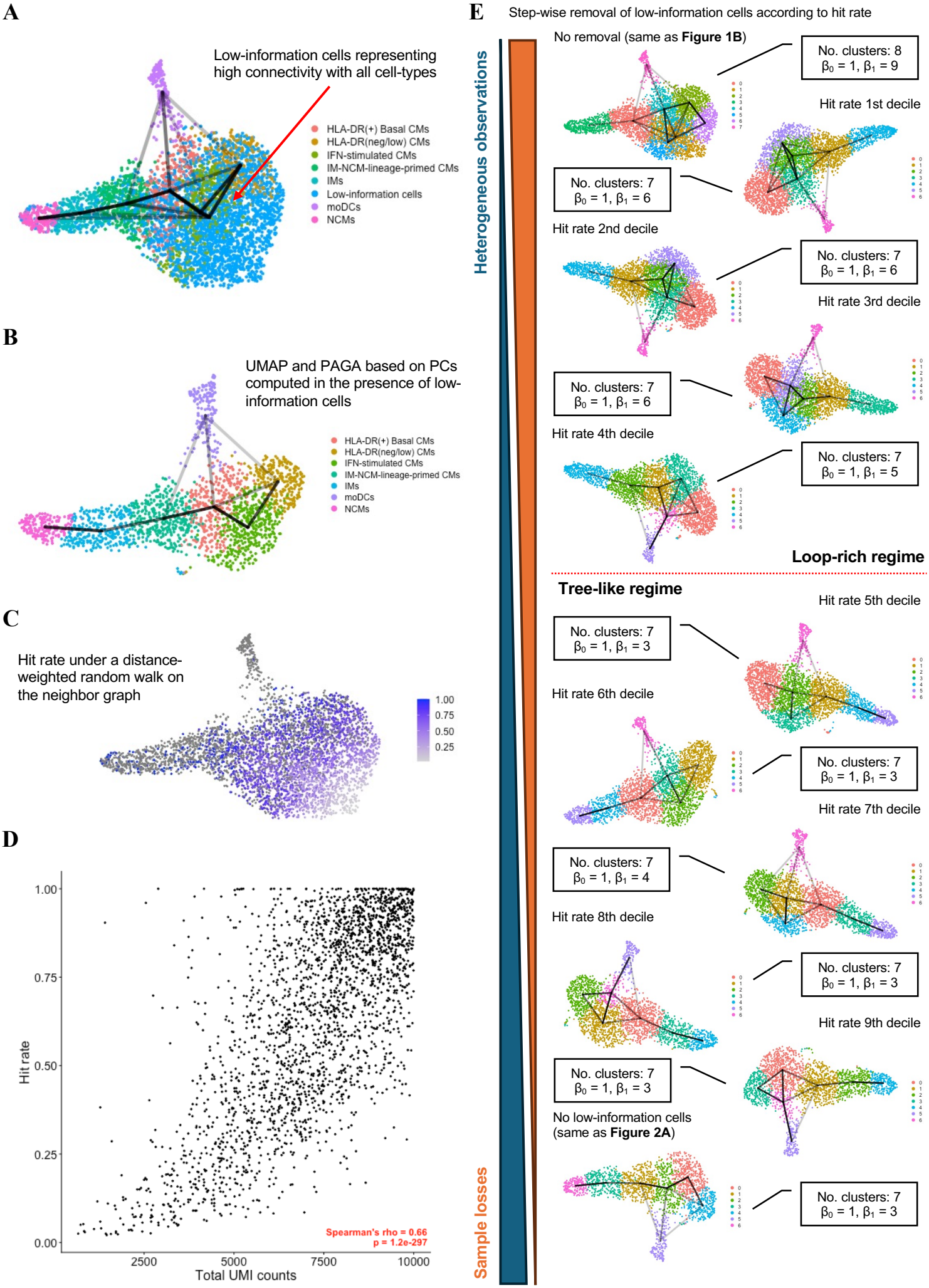


Figure 4

*T. Tamura, *Y. Yamane, *et al.*