LMD: Multiscale Marker Identification in Single-cell RNA-seq Data

Ruiqi Li[1,4], Rihao Qu[1,4], Fabio Parisi[2], Francesco Strino[2], Xiuyuan Cheng[3], and Yuval Kluger[4,5]

[1] Computational Biology & Bioinformatics Program, Yale University, New Haven, CT, USA
[2] PCMGF Limited. 51 Clarendon Road, Watford, WD17 1HP, UK https://pcmgf.com/
[3] Department of Mathematics, Duke University, Durham, NC, USA
[4] Department of Pathology, Yale University School of Medicine, New Haven, CT, USA
[5] Program of Applied Mathematics, Yale University, New Haven, CT, USA

Abstract. Accurate cell marker identification in single-cell RNA-seq data is crucial for understanding cellular diversity and function. An ideal marker is highly specific in identifying cells that are similar in terms of function and state. Current marker identification methods, commonly based on clustering and differential expression, capture general cell-type markers but often miss markers for subtypes or functional cell subsets, with their performance largely dependent on clustering quality. Moreover, cluster-independent approaches tend to favor genes that lack the specificity required to characterize regions within the transcriptomic space at multiple scales.

Here we introduce Localized Marker Detector (LMD), a novel tool to identify "localized genes" - genes with expression profiles specific to certain groups of highly similar cells - thereby characterizing cellular diversity in a multi-resolution and fine-grained manner. LMD's strategy involves building a cell-cell affinity graph, diffusing the gene expression value across the cell graph, and assigning a score to each gene based on its diffusion dynamics.

We show that LMD exhibits superior accuracy in recovering known cell-type markers in the Tabula Muris bone marrow dataset relative to other methods for marker identification. Notably, markers favored by LMD exhibit localized expression, whereas markers prioritized by other clustering-free algorithms are often dispersed in the transcriptomic space. We further group the markers suggested by LMD into functional gene modules to improve the separation of cell types and subtypes in a more fine-grained manner. These modules also identify other sources of variation, such as cell cycle status.

In conclusion, LMD is a novel algorithm that can identify fine-grained markers for cell subtypes or functional states without relying on clustering or differential expression analysis. LMD exploits the complex interactions among cells and reveals cellular diversity at high resolution.

Keywords: scRNA-seq · Cluster-free · Localized markers detection · Gene modules.

1 Introduction

Single-cell RNA sequencing can currently reveal the transcriptome state of $10^4 - 10^7$ cells simultaneously at single-cell resolution [1]. These experimental approaches have enabled a better understanding of heterogeneity, cellular specialization and differentiation, cell-cell signaling, spatial organization, and temporal dynamics leading to e.g. senescence, apoptosis, or oncogenesis [2]. An important task in single-cell data analysis is the identification of genes with high specificity to certain cell populations, which may indicate distinct cellular identities or states in this biological system.

Most marker identification methods [3–10] first cluster cells based on their similarity in the transcriptomic space and then detect genes by differential expression analysis. This approach has been criticized for its unreliable cluster assignments [11–13]. Several state-of-the-art tools do not rely on an initial clustering step. For instance, SinglecellHaystack [11] and SEMITONES [13] utilize reference cells or grid points in the transcriptomic space to approximate cell geometry. They assess if a gene's expression aligns with the arrangement of cells within this coordinate system. These methods' limitation is their confinement to a specific range of length scales, potentially missing sensitivity in other length scales. Marcopolo[12] binarizes gene expression and examines whether the cells expressing the given gene are placed proximal to each other in a low-dimensional space. The proximity is defined by the average distance from each cell expressing the gene to their collective centroid. They assume that this proximity indicates the gene's informativeness in terms of its association with a meaningful biological status. However, the binarization procedure can lead to a loss of quantitative information. Additionally, the positioning of the centroid and proximity assessments can be sensitive to a few very distal cells, which are considered outliers in comparison to the majority of cells expressing the given gene. Hotspot [14] employs a cell-cell similarity graph to identify informative genes with non-random expression patterns, using a modified Geary's C. This 'non-randomness' is a general term that can include very different patterns: a gene being highly expressed in a certain region of the cell graph or showing a smooth expression gradient across the cell manifold. Although each pattern might suggest different biological interpretations, Hotspot does not provide an automated way to distinguish between these scenarios. This limitation, in turn, means it doesn't impose constraints to ensure that the genes it reports are exclusively expressed in specific regions. Consequently, this lack of specificity can undermine the reliability of Hotspot's genes for clear and definitive cell type classification.

To identify the informative genes that categorize distinct cellular identities or states, we consider that such a gene should exhibit a "localized expression pattern" in the cell graph - the graph that represents cell-cell transcriptomic similarities. This is based on the assumption that a cell population sharing similar roles or functions will exhibit highly similar transcriptomic profiles, thereby locating in the same region(s) of the cell graph. To accomplish this goal, we propose a novel algorithm, Localized Marker Detector (LMD), designed to identify all localized gene markers at multi-scale resolutions. Specifically, LMD finds marker genes by (i) constructing a cell-cell affinity graph, (ii) diffusing the gene expression value across the cell graph, and (iii) assigning a score to each gene based on the dynamics of its diffusion process. Intuitively, the diffusion process takes longer to spread across the entire cell-cell graph when seeded with the expression of a gene active in cells that are closer to each other on the cell-cell graph, or when the cells expressing the gene are a subset of a region that is less connected to the rest of the graph.

2 Results

In this section, we present the main results of the paper. First, we analyze the properties of the diffusion dynamics on the cell-cell affinity graph and define the LMD Score (Section 2.1). Next, we benchmark our methods against several algorithms and show that it identifies known markers with high accuracy (Section 2.2) and that it prioritizes local markers (Section 2.3). Finally, we group the genes into distinct functional modules and show that they represent different cell states (Section 2.4).

2.1 Overview of the LMD Method

Cell-cell affinity graph. A cell-cell affinity graph shows which cells are sufficiently similar to each other based on a certain metric. In our cell-cell graph, each cell is represented by a node, and two nodes are

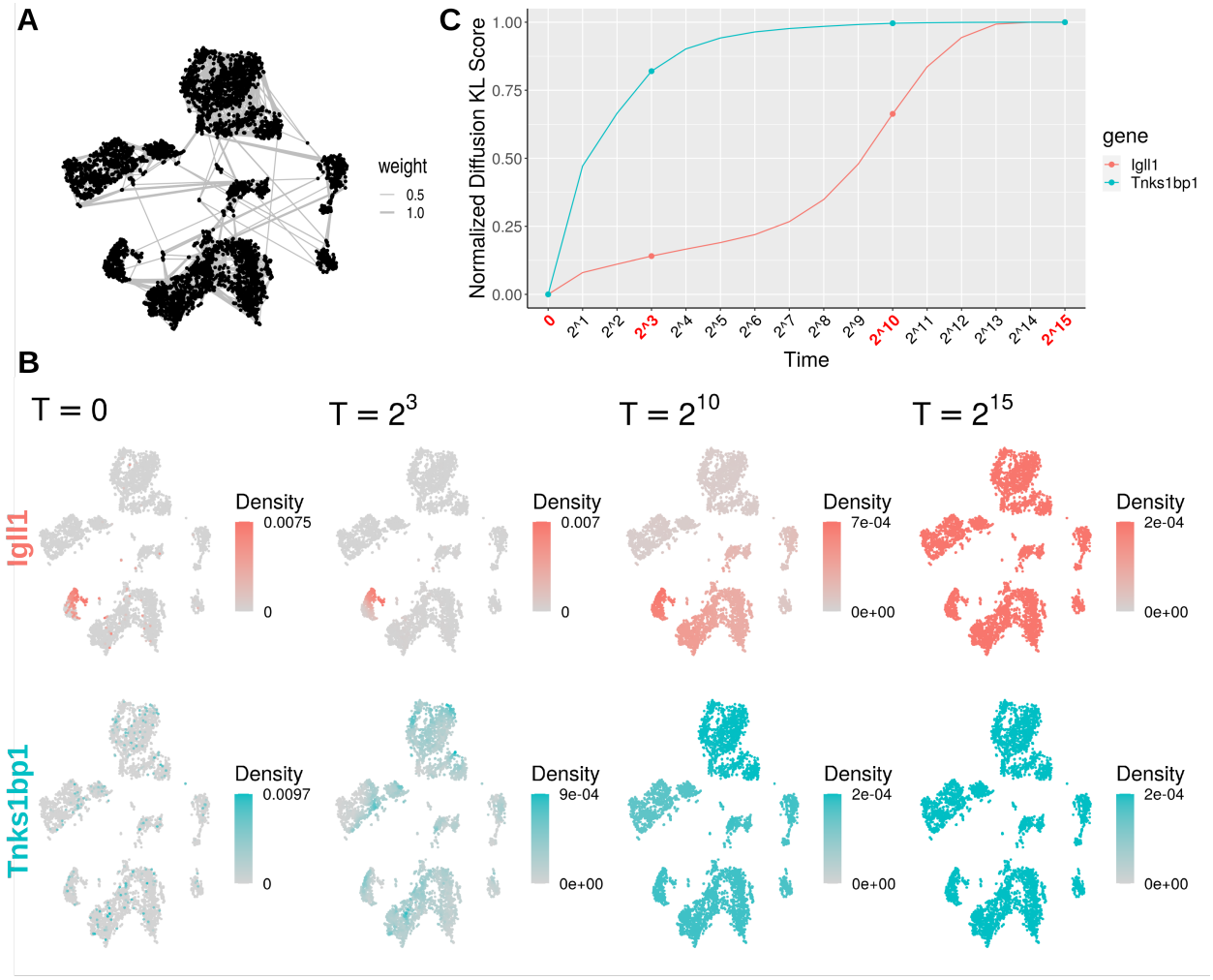

Fig. 1. LMD Workflow (A) t-SNE plot of the kNN (k=5) graph derived from the mouse bone marrow dataset. (B) Gene diffusion patterns for the localized gene Igll1 (red) and the non-localized gene Tnks1bp1 (cyan) overlaid on the t-SNE cell embedding at different time scales. (C) Display of the normalized KL-divergence pattern of Igll1 and Tnks1bp1 during the diffusion process. The y-axis indicates the normalized KL-divergence, while the x-axis represents the time scale. Time scales displayed in panel B are highlighted in red in panel C.

connected by an edge if either of their respective transcriptomes are among the closest $k > 1$ neighbours of the other according to a given metric. The metric can be any function that quantifies the similarity of cells, such as the Euclidean distance between their transcription profiles after reducing the dimensionality with Principal Component Analysis (PCA). A cell-cell affinity graph is undirected, i.e. the connections are symmetric. Moreover, every node in the graph has at least k edges, since each cell has at least k closest neighbors. Fig. 1A shows an example of a cell-cell affinity graph constructed from the Euclidean distances between the transcription profiles after PCA of mouse bone marrow cells obtained from single-cell RNA sequencing[15].

Multi-scale diffusion process. In this step, we examine the gene expression pattern over the cell graph. We first normalize the expression of a gene into a probability distribution over cells, which we define as the initial state vector of the corresponding gene. Next, we apply a diffusion process with a series of time scales - a technique widely adopted in various fields [16–18] - and track how this state vector changes as a function of time. If the graph is connected, given enough time, all nodes will reach a non-zero value regardless of initial states, and the equilibrium state vector for our setup is a uniform distribution.

Some initial states are localized, meaning that most nodes in the graph have zero values except for the nodes in a specific region of the graph (Fig. 1B, top panel). We observe that for a more localized initial state, it takes a longer time to reach a non-zero value at every node and then reach the equilibrium state. (Fig. 1C). To quantify this effect, we measure changes in the state vector as a function of time by computing the Kullback-Leibler divergence (KL-divergence) between the state at time t and the initial state at time t_0. To adjust for the effect where initial states distributed across different numbers of nodes result in varying convergent values, we normalize this time-dependent divergence by the KL-divergence between the initial state and the equilibrium state, and term this quantity as normalized KL-divergence. We find that the normalized KL-divergence for more localized initial states increases at a slower rate during the early stages of the diffusion process, indicating that the redistribution of values primarily occurs between nodes with similar values. In contrast, for initial states that are not localized, the diffusion process quickly redistributes values from higher-value nodes to neighboring lower-value nodes at the beginning. This results in a rapid increase in normalized KL-divergence.

The LMD-score (LMDS). To summarize the time-dependent profile of the normalized KL-divergence, we calculate the area under the curve (Fig. 1C), which we term as the LMD-score (LMDS). A lower LMDS indicates a more localized initial state. In the context of gene expression profiles, we find that genes expressed in specific regions of the cell graph are assigned a lower LMDS, implying LMDS can distinguish localized genes from the genes non-specific to any local regions of the cell graph. This is the main idea of our LMD algorithm.

2.2 LMD identifies known markers with high accuracy

We tested whether LMD was able to recover known markers for mouse single-cell RNA sequencing of bone marrow data [15]. We compared LMD performance to a suite of marker gene identification algorithms consisting of singleCellHaystack [11], Hotspot [14], SEMITONES [13], MarcoPolo [12], and FindAllMarkers from Seurat v4 [3]. Known markers for murine bone marrow were taken from the CellMarker database [19]. While these markers are not necessarily local (e.g. Tnks1bp1, Fig. 1B), we can expect many of them to be reasonably confined to specific regions of the cell-cell affinity graph (e.g. Igll1, Fig. 1B) as they were chosen to be cell-type and cell-subtype specific.

We found that LMD yielded overall higher sensitivity compared to all other methods (AUROC 0.73, Fig. 2A, left panel) and correctly recovering 400 markers out of 1277 within the top 10% candidates, corresponding to a 10-fold enrichment (P-value $< 10^{-50}$, hypergeometric test). In comparison, MarcoPolo and Hotspot, the next best-performing algorithms (AUROC 0.69 and 0.68), recovered 350 and 300 markers respectively from within their top 10% of candidates. We also tested mouse single-cell RNA sequencing of pancreas and lung data [15] and found that LMD performed better (pancreas) or similarly (lung) compared to the other algorithms in recovering tissue-specific markers from the CellMarker database (Fig. 2A, left panel).

A potential limitation of this comparison is that CellMarker's markers are based on published data. This list may be biased by external factors unrelated to biology, such as research trends or intrinsic technical difficulties with validation experiments. To create a less biased set, albeit more prone to false positives, we defined the gold-standard marker set from the sequencing data itself as the top 100 genes with the largest expression fold change across cell types (see Methods). LMD yielded the highest AUROC score in the three tested tissues, 0.89, 0.96 and 0.98 for bone marrow, pancreas, and lung respectively (Fig. 2A, right panel).

2.3 LMD prioritizes localized markers

We compared the ability of markers from each method to enhance separability between cell groups. This can be quantitatively estimated using the density index [20], which evaluates cell distribution 'clumpiness' in the feature space by contrasting the average closeness of nearest neighbors against the average distance between random cell pairs. A higher density index value indicates a more compact and clumped cell distribution. For each method, we investigated how the density index changes as a function of the number of selected top-ranking markers. LMD showed the highest density index compared to the other methods (Fig. 2B). In addition, the density index for LMD decreased as we selected a larger number of markers indicating that the method prioritizes localized markers that can separate diverse cell populations.

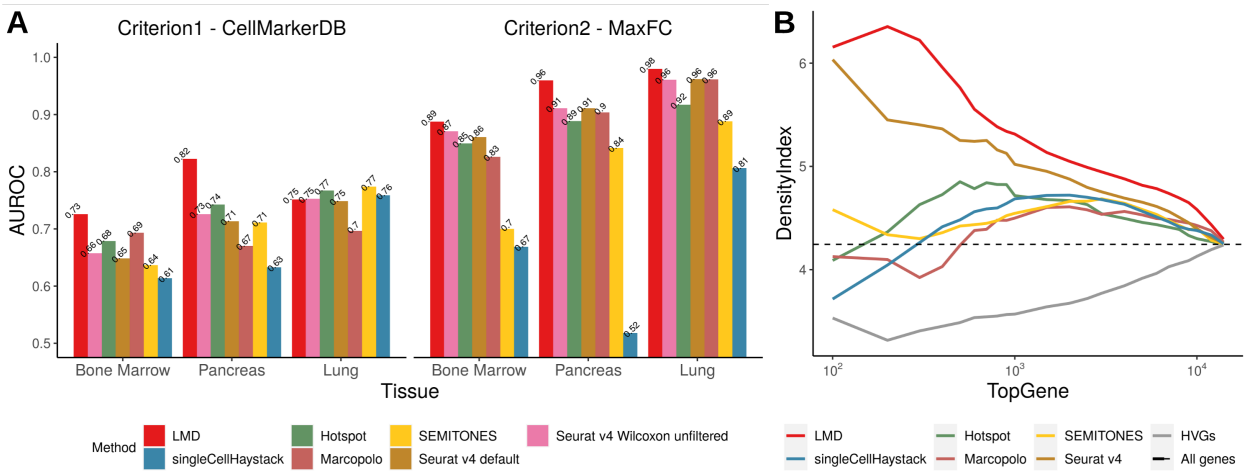

Fig. 2. Marker Identification and Separability Comparison between LMD and Other Methods (A) Benchmark of different methods in identifying cell type markers in three real datasets (Tabula Muris bone marrow, pancreas and lung): The y-axis displays the AUROC scores, while the x-axis represents true marker sets based on two criteria: markers derived from CellMarkerDB (left) and the top 100 genes with the maximum log fold change across cell types (right). (B) Density Index over a 20 Principal Components representation of scRNA-seq mouse Bone Marrow dataset. The density index is computed from the top-ranked genes identified using LMD, singleCellHaystack, Hotspot, Marcopolo, SEMITONES, and Seurat v4. Only one curve is shown for Seurat since the results of FindAllMarkers with and without its default filter produced the same results.

To better understand the difference between LMD and the other methods – Seurat v4 default, Hotspot, SEMITONES, Marcopolo, and singleCellHaystack – we compared their gene rankings. Among the methods, Seurat v4 default showed the highest concordance with LMD, while singleCellHaystack had the lowest (Fig. 3A). Candidate markers specific to LMD were typically expressed in compact cell neighborhoods (Fig. 3B, top row). In contrast, other methods prioritized markers lacking locality (Fig. 3B, bottom row). In detail, Seurat v4 identifies genes through differential expression and Hotspot prioritizes genes with global variation, thus both methods may overlook the confinement of markers to specific cell groups, i.e. they tend to lack locality; Marcopolo's markers often correspond to a noisy and less localized signal; SEMITONES and SingleCellHaystack missed markers that are expressed in cell subsets smaller than their pre-defined fixed length scale.

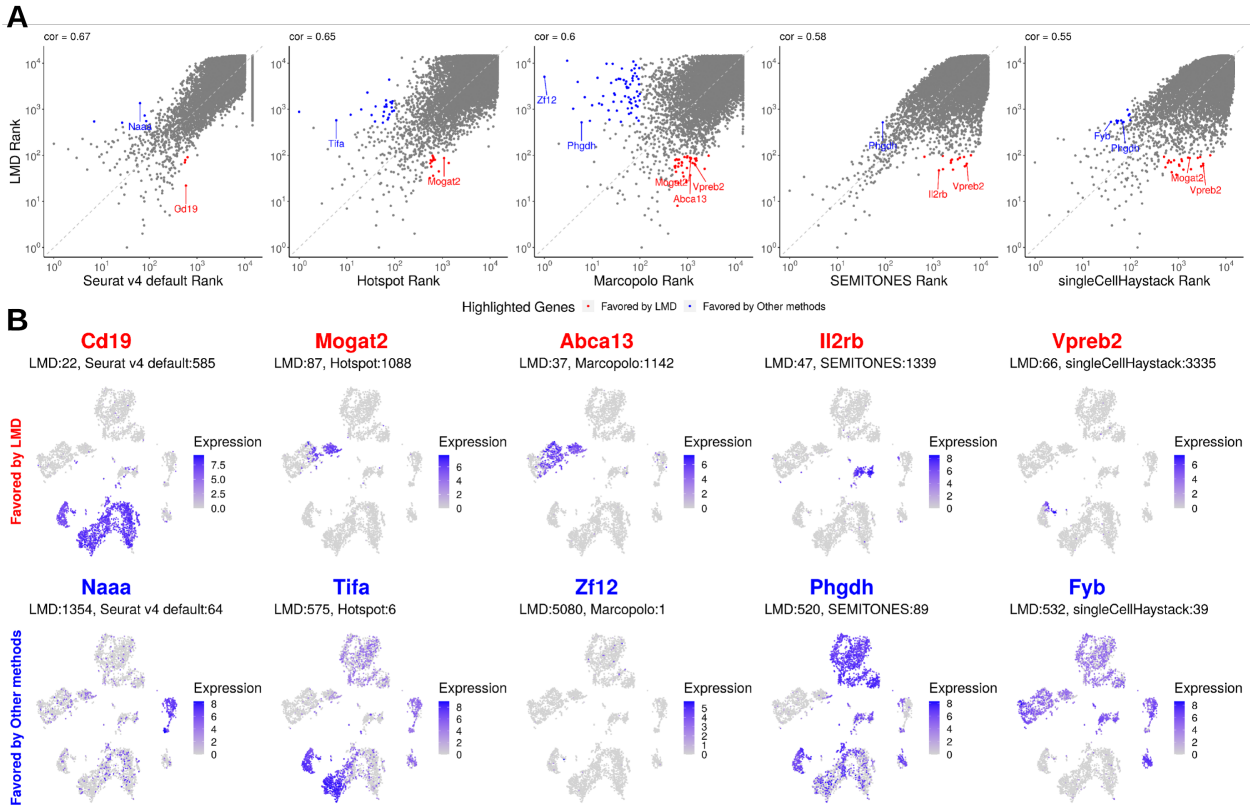

Fig. 3. Gene Ranking Comparisons between LMD and Other Methods. (A) Comparison of gene rankings by LMD against Seurat v4 default, Hotspot, Marcopolo, SEMITONES, and SingleCellHaystack, with concordance measured using the Pearson correlation coefficient of log-transformed rankings (base 10). Highlighted genes fall within LMD's top 100 ranked genes but are outside the top 500 ranked genes of the other method (red), and vice versa (blue). (B) Selected genes favored by LMD but neglected by other methods and vice versa. The detailed ranking of these genes across all methods can be found in Table S1.

2.4 LMD markers reveal distinct functional modules uncovering cell states at varied resolutions

Gene modules. To characterize the biological significance of LMD's candidate markers, we clustered the top 1,000 genes in 56 distinct gene modules (Fig. S2): 46 were enriched for at least one GO term (Benjamini-Hochberg adjusted p-value < 0.05) and 40 were enriched for at least one pathway from the Reactome Pathway Database[21] (Benjamini-Hochberg adjusted p-value < 0.05). For detailed information on each module, see Section S5.1. As expected, given the locality of the underlying markers, each module was associated with specific regions of the cell-cell affinity graph, which enables the characterization of the data at different length scales (Fig. S3). At a broader length scale, we found several modules that were jointly expressed across multiple local regions, for instance, modules associated with progression through the cell cycle. At cell-type resolution, LMD identified several modules that demonstrated high specificity to known annotations using just a small number of markers. Finally, at a finer length scale, LMD revealed several gene modules with each associated with a specific subpopulation in a pre-annotated cell type, suggesting the potential of using LMD to define novel cell types at a high resolution. Moreover, inspecting certain gene modules may elucidate the different cell states along a dynamic biological process (e.g., myeloid lineage differentiation) and facilitate the discovery of novel cell subtypes in this biological system.

Broad length-scale, larger data structures. Reactome pathway enrichment analysis for modules 32, 31, and 18 showed significant enrichment for G1/S transition phase, S phase, and M phase, respectively (Benjamini-Hochberg adjusted p-value < 0.05, Fig. 4A). We confirmed this result in an alternative mouse bone marrow dataset that employed microfluidic droplet-based 3'-end counting technique [15]. In both datasets (bone marrow FACS and droplet), the depiction of these gene modules on the cell t-SNE embedding closely matches cell annotations derived from Seurat cell cycle scores, which rely on established cell cycle markers. In contrast to Seurat annotations, which do not allow overlapping clusters, the modules derived from LMD's markers indicated incremental activation of markers, thus capturing the dynamics of progression through cell-cycle (Fig. 4B).

Cell-type resolution. LMD identified several modules that demonstrated specificity to pre-annotated cell types. The strength of this association was quantified using the Jaccard Index (see Methods). Due to limited cell numbers, we merged T cell subtypes into 'T cell' and NK cell subtypes into 'NK cell'. Of the resulting 18 cell types, 13 showcased at least one distinctly activated module, as depicted in Fig. S5. The relationship between these modules and cell types was further validated as the modules incorporated canonical cell type markers used for annotating this Tabula Muris bone marrow dataset (see Notes of Fig. S5). We again performed cross-dataset verification on the droplet dataset and found that modules identified as specifically activated in particular cell types in the FACS dataset also showed high specificity for the same cell types in the droplet dataset (Fig. S6).

High resolution, beyond cell types. We identified gene modules that trace the B lymphocyte differentiation process (Fig. 4C) or various developmental stages and subtypes of granulocytes (Fig. 4D). Modules associated with B lymphocytes are substantiated by the inclusion of genes with well-established roles in this differentiation process: Module 7, with genes Rag1 and Rag2, represents pro-B and pre-B cells, as these genes participate in the heavy and light chain rearrangement processes [22]; module 9, containing genes Ms4a1 and Cd22, pertains to B cell activation and regulation, marking both immature and mature B cells in the bone marrow [23, 24]; module 10, which features genes Fcer2a, Ms4a4c, and Cxcr5, characterizes naive B cells [25–27]. We similarly identified at least 4 modules associated with granulopoiesis: module 45, exclusively expressed in granulopoietic cells and including the granulopoietic marker Fcnb [11, 28]; module 44, featuring markers Camp and Ltf, both of which are found in granulocyte and granulopoietic cells [15]; module 41, specific to a subset of granulocytes and containing Retnlg and Cxcr2, which suggests a potential role in neutrophil migration and release [29, 30]; lastly, module 43, characterized by genes implicated in pyroptosis and antigen processing such as Stfa2l1, Il1b, Gm5483, H2-Q10, and Irg1 [31–34], which may represent a rare granulocyte subgroup engaged in the immune response against intracellular pathogens.

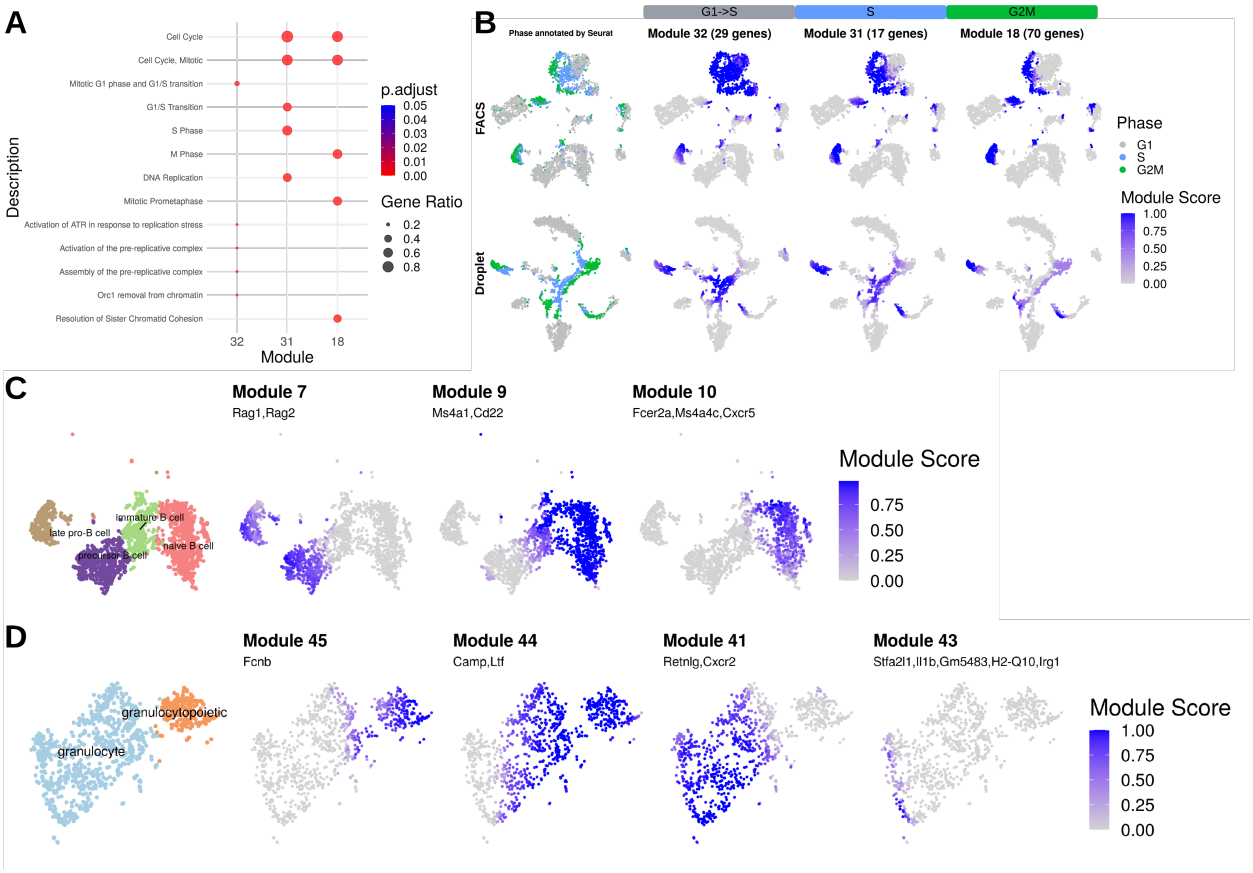

Fig. 4. Illustration of gene module specificity in cell cycle stages and within cell subsets. (A-B) Gene modules associated with the cell cycle. Gene modules were identified from the FACS dataset and their function was classified by Reactome pathway analysis. Modules 32, 31, and 18 were associated with G1→S transition, S phase, and G2/M phases, respectively. (A) Bubble plot of the 5 most significantly enriched pathways for each module. Each bubble is colored according to its adjusted p-value and scaled by the fraction of genes of the module that are associated with the pathway. (B) Module scores in two mouse bone marrow datasets, and cell cycle annotations obtained using Seurat are shown for comparison (left). (C) Gene modules related to B cells at different stages of maturation as annotated in the mouse bone marrow dataset. Module 7 is associated with late progenitor and precursor B cells, module 9 with immature and naive B cells, and module 10 with naive B cells. (D) Gene modules associated with the production of granulocytes. Module 45 contains genes expressed during granulocytopoiesis, module 44 contains genes common for granulocytopoietic cells and granulocytes, module 41 contains genes involved in granulocyte maturation, and module 43 may represent a rare granulocyte subgroup.

3 Discussion

We present LMD, a novel approach for identifying localized markers at multiple scales in single-cell RNA sequencing data. The core idea of LMD is that the dynamics of a diffusion process for each gene on a given cell-cell affinity graph reflects certain properties of the gene expression pattern such as locality. We have shown that expression patterns concentrated over a compact region of the cell-cell affinity graph yield a slower progression of diffusion at the beginning of this process. This property is captured by a characteristic LMDS for each gene, and it enables the prioritization of localized expression patterns.

LMD was able to recover known cell-type markers from bone marrow, lung, and pancreas single-cell murine RNA sequencing data, often better than other marker identification algorithms, which do not necessarily prioritize localization of candidate markers. While important for LMD, this result also implies that many known markers are indeed localized. The preference for localized markers further supports the need for algorithms like LMD over generic marker identification algorithms that cannot distinguish between local and global expression profiles.

LMD's candidate markers are also strongly associated with compact cell groups as demonstrated by our analyses of the density index. LMD's top candidate markers were consistently better at separating cells into different groups compared to the other tested algorithms, regardless of whether we considered the top 100, 500, 1000, or 5000 candidates. It follows that using LMD's top candidate markers it is possible to obtain low-dimensional representations of single-cell sequencing data that could facilitate further analyses and interpretation beyond the current state-of-the-art. We illustrate this point with examples of biological relevance: first, we compare a standard assignment of cells to cell-cycle states by Seurat with the dynamic view given by LMD, in which it is possible to discern the sequential activation of specific localized cell-cycle markers; second, we illustrate the dynamic progression of B-cell and Granulocyte maturation with just a handful of LMD's precisely localized markers.

The results presented are remarkable considering that LMD achieves them without relying on any knowledge about existing cell annotations or gene functions. LMD candidate markers are selected only based on their ability to localize to specific regions in the cell-cell affinity graph, and thus identify compact groups of similar cells, with high specificity, regardless of length scale, which is ultimately the goal of a useful marker.

4 Methods

4.1 Localized Marker Detector (LMD)

Input data. The input to LMD is a gene by cell matrix $X \in \mathbb{R}^{m \times n}$, where G is the set of all genes ($|G| = m$), C is the set of all cells ($|C| = n$), and the element $x_{i,j}$ is the measured expression level of gene i in cell j.

Cell-cell affinity matrix $\boldsymbol{P}$. First, we identify the k-nearest neighbors for each cell from the Euclidean distance between the transcription profiles after applying dimensional reduction, e.g. PCA (see Section 4.3). Next, the adjacency matrix A is constructed such that the element $a_{i,j}$ is 1 if cell i is a k-nearest neighbor of cell j and 0 otherwise. Finally, the doubly-stochastic cell-cell affinity matrix P is obtained by applying the Sinkhorn-Knopp algorithm [35] to the symmetrized matrix $\left(A + A^T\right)/2$.

Diffusion process. Given an initial state over the cells, propagate it over the nodes of the graph using a diffusion process. The initial state for gene i on the cell-cell affinity graph is the $(n \times 1)$ vector $\rho^{(i,0)}$ whose elements are $\rho_j^{(i,0)} = x_{i,j} / \sum_{j \in C} x_{i,j}$. The state of gene i at time t is estimated using a diffusion process: $\rho^{(i,t)} = (P^t)' \rho^{(i,0)}$. Note that for $t \to \infty$, $\rho^{(i,t)} \to 1/n \cdot \mathbf{1}, \forall i \in G$.

LMD score. For a given gene i, the Kullback-Leibler divergence (KL-divergence) between the initial state and the state at time t is $\delta(i,t) = D_{KL}(\rho^{(i,0)} || \rho^{(i,t)}) = \sum_{j \in C} \rho_j^{(i,0)} \log\left(\rho_j^{(i,0)} / \rho_j^{(i,t)}\right)$. The Normalized KL-divergence is computed as $d(i,t) = \delta(i,t)/\delta(i,\infty)$. We define the LMD score as LMDS $= \sum_t d(i,t)$. In LMD we compute LMDS over $t = 2^\tau, \tau = 1 \ldots T$ for a reasonably large T and rank candidate genes by ascending values of LMDS.

4.2 Gene modules

Identification. To identify gene modules, we first denoised the expression data using ALRA [36]. Next, we selected the top 1000 LMD candidate markers and computed the gene-gene pairwise distance matrix using the Jaccard distance [37] between gene expression profiles across all cells. A small Jaccard distance suggests that these two genes are co-localized in the cell graph. We then adopted a widely accepted strategy for gene module identification [38] where we processed the gene-gene pairwise distance matrix with hierarchical clustering with average linkage (hclust from R package stats, v. 4.1.3) [39] and we determined gene modules using the dynamic tree-cutting algorithm (dynamicTreeCut in R, v.1.63.1) [40].

Module Activity Scores. The module activity score (or module score) is the probability that we detect a module expressing in a given cell and it is estimated from repeating the following sampling procedure 100 times. First, we randomly selected half of the genes without replacement and used their average expression in each cell to cluster cells via a Gaussian Mixture Model (GMM). Next, we assigned these clusters to two distinct categories depending on whether the module was expressed via a hierarchical clustering of GMM cluster centroids with average linkage. Finally, we applied local smoothing to refine these groupings by classifying a cell as expressing the gene module based on majority voting of its 10 neighboring cells.

Cell type-specific activated modules Cell type-specific activated modules are modules for which the Jaccard index between the module's activity score and the one-hot vector representation of a given cell type was greater than 0.4.

4.3 Evaluation

Dataset and Processing. For each dataset, we retained genes detected in more than 10 cells but in fewer than 50% of cells. For measuring cell-cell distances we utilized PC coordinates from [41], which were derived using the default pipeline of Seurat. For visualization we employed t-SNE coordinates from [41]. For a fair comparison, for all methods, we used the same data, scaled or transformed as required by the algorithm's input specifics, e.g. either raw count data, log-normalized count data, or PC coordinates.

Gold-standard cell-type marker sets. We employ two distinct gold-standard marker sets, as suggested in [12]. The first gold-standard identified marker genes from the CellMarker database [19] which were specific to the relevant tissue and present in our datasets. The second gold standard consists of the top 100 genes with the largest differential expression across cell types [12], using cell-type annotations provided by the original studies.

Gene Ranking by Competing Method LMD was qualitatively compared to other marker gene identification methods: singleCellHaystack, Hotspot, SEMITONES, MarcoPolo, Seurat v4, and HighlyVariableGenes. For a detailed explanation of the gene ranking procedures of other methods, refer to Section S1.3.

4.4 Code availability

LMD's source code and tutorial covering the present work in R implementation will be available soon on GitHub.

4.5 Acknowledgment

The authors thank Ronald Coifman, Will Yaochen Zhu, Junchen Yang for fruitful discussions. Y.K., F.P. and F.S. disclose support for the research of this work from NIH [U54AG076043, U54AG079759]. Y.K. is also supported by [R01GM131642, UM1PA051410, U54AG076043, U54AG079759, U01DA053628, P50CA121974, and R33DA047037]. X.C. is also partially supported by NSF [DMS-2237842, DMS-2007040] and Simons Foundation.

References

1. J. S. Fleck, J. G. Camp, and B. Treutlein, "What is a cell type?," *Science*, vol. 381, no. 6659, pp. 733–734, 2023.
2. C. Cheng, W. Chen, H. Jin, and X. Chen, "A review of single-cell rna-seq annotation, integration, and cell–cell communication," *Cells*, vol. 12, no. 15, p. 1970, 2023.
3. Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, *et al.*, "Integrated analysis of multimodal single-cell data," *Cell*, vol. 184, no. 13, pp. 3573–3587, 2021.
4. R. Patrick, D. T. Humphreys, V. Janbandhu, A. Oshlack, J. W. Ho, R. P. Harvey, and K. K. Lo, "Sierra: discovery of differential transcript usage from polya-captured single-cell rna-seq data," *Genome biology*, vol. 21, no. 1, pp. 1–27, 2020.
5. V. Svensson, S. A. Teichmann, and O. Stegle, "Spatialde: identification of spatially variable genes," *Nature methods*, vol. 15, no. 5, pp. 343–346, 2018.
6. Z. Miao, K. Deng, X. Wang, and X. Zhang, "Desingle for detecting three types of differential expression in single-cell rna-seq data," *Bioinformatics*, vol. 34, no. 18, pp. 3223–3224, 2018.
7. S. Nabavi, D. Schmolze, M. Maitituoheti, S. Malladi, and A. H. Beck, "Emdomics: a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes," *Bioinformatics*, vol. 32, no. 4, pp. 533–541, 2016.
8. K. D. Korthauer, L.-F. Chu, M. A. Newton, Y. Li, J. Thomson, R. Stewart, and C. Kendziorski, "A statistical approach for identifying differential distributions in single-cell rna-seq experiments," *Genome biology*, vol. 17, pp. 1–15, 2016.
9. M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.
10. X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell, "Reversed graph embedding resolves complex single-cell trajectories," *Nature methods*, vol. 14, no. 10, pp. 979–982, 2017.
11. A. Vandenbon and D. Diez, "A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data," *Nature communications*, vol. 11, no. 1, p. 4318, 2020.
12. C. Kim, H. Lee, J. Jeong, K. Jung, and B. Han, "Marcopolo: a method to discover differentially expressed genes in single-cell rna-seq data without depending on prior clustering," *Nucleic acids research*, vol. 50, no. 12, pp. e71–e71, 2022.
13. A. H. C. Vlot, S. Maghsudi, and U. Ohler, "Cluster-independent marker feature identification from single-cell omics data using semitones," *Nucleic Acids Research*, vol. 50, no. 18, pp. e107–e107, 2022.
14. D. DeTomaso and N. Yosef, "Hotspot identifies informative gene modules across modalities of single-cell genomics," *Cell systems*, vol. 12, no. 5, pp. 446–456, 2021.
15. T. M. Consortium, O. coordination Schaum Nicholas 1 Karkanias Jim 2 Neff Norma F. 2 May Andrew P. 2 Quake Stephen R. quake@ stanford. edu 2 3 f Wyss-Coray Tony twc@ stanford. edu 4 5 6 g Darmanis Spyros spyros. darmanis@ czbiohub. org 2 h, L. coordination Batson Joshua 2 Botvinnik Olga 2 Chen Michelle B. 3 Chen Steven 2 Green Foad 2 Jones Robert C. 3 Maynard Ashley 2 Penland Lolita 2 Pisco Angela Oliveira 2 Sit Rene V. 2 Stanley Geoffrey M. 3 Webber James T. 2 Zanini Fabio 3, and C. data analysis Batson Joshua 2 Botvinnik Olga 2 Castro Paola 2 Croote Derek 3 Darmanis Spyros 2 DeRisi Joseph L. 2 27 Karkanias Jim 2 Pisco Angela Oliveira 2 Stanley Geoffrey M. 3 Webber James T. 2 Zanini Fabio 3, "Single-cell transcriptomics of 20 mouse organs creates a tabula muris," *Nature*, vol. 562, no. 7727, pp. 367–372, 2018.
16. R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods," *Proceedings of the National Academy of Sciences*, vol. 102, no. 21, pp. 7432–7437, 2005.
17. R. R. Coifman and M. Maggioni, "Diffusion wavelets," *Applied and computational harmonic analysis*, vol. 21, no. 1, pp. 53–94, 2006.
18. C. J. Dsilva, R. Talmon, N. Rabin, R. R. Coifman, and I. G. Kevrekidis, "Nonlinear intrinsic variables and state reconstruction in multiscale simulations," *The Journal of chemical physics*, vol. 139, no. 18, 2013.
19. X. Zhang, Y. Lan, J. Xu, F. Quan, E. Zhao, C. Deng, T. Luo, L. Xu, G. Liao, M. Yan, *et al.*, "Cellmarker: a manually curated resource of cell markers in human and mouse," *Nucleic acids research*, vol. 47, no. D1, pp. D721–D728, 2019.
20. B. Ranjan, W. Sun, J. Park, K. Mishra, F. Schmidt, R. Xie, F. Alipour, V. Singhal, I. Joanito, M. A. Honardoost, *et al.*, "Dubstepr is a scalable correlation-based feature selection method for accurately clustering single-cell data," *Nature Communications*, vol. 12, no. 1, p. 5849, 2021.
21. M. Gillespie, B. Jassal, R. Stephan, M. Milacic, K. Rothfels, A. Senff-Ribeiro, J. Griss, C. Sevilla, L. Matthews, C. Gong, *et al.*, "The reactome pathway knowledgebase 2022," *Nucleic acids research*, vol. 50, no. D1, pp. D687–D692, 2022.
22. R. Kumari, U. Roy, S. Desai, N. M. Nilavar, A. Van Nieuwenhuijze, A. Paranjape, G. Radha, P. Bawa, M. Srivastava, M. Nambiar, *et al.*, "Microrna mir-29c regulates rag1 expression and modulates v (d) j recombination during b cell development," *Cell Reports*, vol. 36, no. 2, 2021.

23. M. Jensen, G. Tan, S. Forman, A. M. Wu, and A. Raubitschek, "Cd20 is a molecular target for scfvfc: zeta receptor redirected t cells: implications for cellular immunotherapy of cd20+ malignancy," *Biology of blood and marrow transplantation*, vol. 4, no. 2, pp. 75–83, 1998.
24. T. F. Tedder, J. C. Poe, and K. M. Haas, "Cd22: a multifunctional receptor that regulates b lymphocyte survival and signal transduction," *Advances in immunology*, vol. 88, pp. 1–50, 2005.
25. J. D. Mountz, M. Gao, D. M. Ponder, S. Liu, C.-W. Sun, F. Alduraibi, K. Sullivan, B. Pat, L. J. Dell'Italia, and H.-C. Hsu, "Il-4 receptor blockade is a global repressor of naïve b cell development and responses in a dupilumab-treated patient," *Clinical Immunology*, vol. 244, p. 109130, 2022.
26. H. Xu, M. S. Williams, and L. M. Spain, "Patterns of expression, membrane localization, and effects of ectopic expression suggest a function for ms4a4b, a cd20 homolog in th1 t cells," *Blood*, vol. 107, no. 6, pp. 2400–2408, 2006.
27. J. Sáez de Guinoa, L. Barrio, M. Mellado, and Y. R. Carrasco, "Cxcl13/cxcr5 signaling enhances bcr-triggered b-cell activation by shaping cell dynamics," *Blood, The Journal of the American Society of Hematology*, vol. 118, no. 6, pp. 1560–1569, 2011.
28. K. Hunold, D. Weber-Steffens, V. L. Runza, J. C. Jensenius, and D. N. Männel, "Functional analysis of mouse ficolin-b and detection in neutrophils," *Immunobiology*, vol. 217, no. 10, pp. 982–985, 2012.
29. L. Ding, M. M. Hayes, A. Photenhauer, K. A. Eaton, Q. Li, R. Ocadiz-Ruiz, J. L. Merchant, *et al.*, "Schlafen 4–expressing myeloid-derived suppressor cells are induced during murine gastric metaplasia," *The Journal of clinical investigation*, vol. 126, no. 8, pp. 2867–2880, 2016.
30. P. Delobel, B. Ginter, E. Rubio, K. Balabanian, and G. Lazennec, "Cxcr2 intrinsically drives the maturation and function of neutrophils in mice," *Frontiers in Immunology*, vol. 13, p. 1005551, 2022.
31. K. J. Goodall, A. Nguyen, A. Matsumoto, J. R. McMullen, S. B. Eckle, P. Bertolino, L. C. Sullivan, and D. M. Andrews, "Multiple receptors converge on h2-q10 to regulate nk and $\gamma\delta$t-cell development," *Immunology and Cell Biology*, vol. 97, no. 3, pp. 326–339, 2019.
32. K. H. Kwack, N. A. Lamb, J. E. Bard, E. D. Kramer, L. Zhang, S. I. Abrams, and K. L. Kirkwood, "Discovering myeloid cell heterogeneity in mandibular bone–cell by cell analysis," *Frontiers in Physiology*, vol. 12, p. 731549, 2021.
33. S. Ping, X. Qiu, M. E. Gonzalez-Toledo, X. Liu, and L.-R. Zhao, "Stem cell factor in combination with granulocyte colony-stimulating factor protects the brain from capillary thrombosis-induced ischemic neuron loss in a mouse model of cadasil," *Frontiers in Cell and Developmental Biology*, vol. 8, p. 627733, 2021.
34. Z. Chen, X. Ren, R. Ren, Y. Wang, and J. Shang, "The combination of g-csf and amd3100 mobilizes bone marrow-derived stem cells to protect against cisplatin-induced acute kidney injury in mice," *Stem Cell Research & Therapy*, vol. 12, pp. 1–13, 2021.
35. R. Sinkhorn and P. Knopp, "Concerning nonnegative matrices and doubly stochastic matrices," *Pacific Journal of Mathematics*, vol. 21, no. 2, pp. 343–348, 1967.
36. G. C. Linderman, J. Zhao, M. Roulis, P. Bielecki, R. A. Flavell, B. Nadler, and Y. Kluger, "Zero-preserving imputation of single-cell rna-seq data," *Nature communications*, vol. 13, no. 1, p. 192, 2022.
37. S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *City*, vol. 1, no. 2, p. 1, 2007.
38. P. Langfelder and S. Horvath, "Wgcna: an r package for weighted correlation network analysis," *BMC bioinformatics*, vol. 9, no. 1, pp. 1–13, 2008.
39. F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion?," *Journal of classification*, vol. 31, pp. 274–295, 2014.
40. P. Langfelder, B. Zhang, and S. Horvath, "Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r," *Bioinformatics*, vol. 24, no. 5, pp. 719–720, 2008.
41. Tabula Muris Consortium, "Data Processing." `https://github.com/czbiohub-sf/tabula-muris/blob/master/00_data_ingest/All_FACS_Notebook.Rmd`, 2018.