Nuance Matters: Quantification of Mutual Exclusive Gene Expressions in Single-Cell and Spatial Transcriptomics

Jinpu Cai[1,2], Cheng Wang[2], Luqi Yang[2], Luting Zhou[1], Xinzhu Jiang[1], Xiaorui Liu[3], Yibo Zhang[3], Bingyi Li[4], Ziqi Rong[5], Haoyang Liu[2], Weixi Luo[6], Hui Cheng[1], Shyam Prabhakar[7], Liang Chen[3⋆] Qiuyu Lian[8,9⋆⋆], and Hongyi Xin[2,10⋆⋆⋆]

[1] Global College, Shanghai Jiao Tong University, Shanghai, China
[2] Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai, China
[3] Fuwai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
[4] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[5] Bioinformatics Interdepartmental Program, University of California, Los Angeles, CA, USA
[6] School Of Computer Science, Shanghai Jiao Tong University, Shanghai, China
[7] Genome Institute of Singapore, A*STAR, Singapore, Singapore
[8] Gurdon Institute, Tennis Court Road, University of Cambridge, Cambridge, UK
[9] Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, University of Cambridge, Cambridge, UK
[10] School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai, China

Abstract. Mutually exclusive gene expression represents a fundamental mode of inter-gene regulation, wherein gene pairs exhibit strict alternative expression across non-overlapping cell populations. When extended to gene groups, collective mutual exclusivity reveals systematic transcriptional rewiring that accompanies cell differentiation and environmental adaptation. Existing computational approaches, including anti-correlation analysis and co-expression testing, fail to account for the continuous, non-binary nature of single-cell and spatial transcriptomics data and are heavily confounded by prevalent dropout events. These methods are inherently biased toward high-expression genes in large cell populations, limiting their applicability to rare cell types and lowly expressed transcripts. To address these limitations, we present MULE, an unbiased, parameter-sparse framework that organizes collective mutually exclusive genes into hierarchical gene modules. MULE employs a continuous mutual exclusivity metric that quantifies the ratio of aggregated co-expression strength to individual expression strengths, capturing the conserved topological characteristics of mutually exclusive gene pairs. Permutation-based statistical testing validates putative mutually exclusive genes while distinguishing genuine signal from spurious patterns arising from dropout events. Benchmarking across synthetic and real datasets demonstrates that MULE is accurate across expression regimes, from high-expression markers to high-dropout genes, with consistent performance in major cell types, subtypes, and rare populations. Integration of MULE-identified gene modules with representation learning improves cell-type separation in latent space and enables accurate temporal resolution of non-overlapping transcriptional states. Applying MULE across single-cell and spatial transcriptomics and comparing mutual exclusivity patterns reveals spatial segregation and co-localization of functionally distinct populations. By explicitly modeling mutually exclusive gene programs, MULE provides a principled framework for detecting subtle, biologically relevant expression differences between closely related cell populations, advancing the interpretation of complex transcriptomic data.

⋆ Corresponding author. chenliang_2012@126.com
⋆⋆ Corresponding author. ql333@cam.ac.uk
⋆⋆⋆ Corresponding author. hongyi.xin@sjtu.edu.cn

Introduction

Mutually exclusive expression is one of the fundamental inter-gene expression patterns. It is characterized by the singular selective expression among parallel gene programs[1]. Mutually exclusive expression reflects the intricate and coordinated gene expression regulation apparatus that balances activation and repression, which ensures precise cellular specialization[2]. From a developmental biology perspective, cell fate commitment involves the simultaneous activation of lineage-specific marker genes and silencing of alternative gene programs. Terminal fully differentiated cell types, thus often exhibit strong mutual exclusive expressions between their respective lineage markers[3]. From a differentiation dynamic perspective, mutual exclusivity denotes the temporal isolation of gene expressions in a trajectory, where the expression between two genes share no time overlap throughout the process[4]. Non-trivial gene groups that share collective mutual exclusivity within a trajectory suggest that there is organized gene expression rewiring between discrete putative stationary sub-states within a seemingly continuous transition[5] (Fig1a).

Accurately identifying and characterizing collective mutual exclusive expressions in scRNA-seq data allows us to investigate subtle gene expression rewiring between nuanced cell sub-populations independent from clustering[6]. When multiple genes exhibit evident collective mutually exclusive expressions, that is: multiple genes that are partitioned into distinct gene sets, where there exists strong mutual exclusivity between any gene pair across the gene set and there is no mutually exclusive expressions between genes within a gene set, then there is strong indication that the data either consists multiple terminal cell types of distinct lineages expressing their respective lineage markers, or there is systematic rewiring taking place between cell-state transitions. When there is nested mutual exclusive expressions, then either there is a nested major-to-sub-type multi-layered cell-type hierarchy, or there are sub-state transitions within major cell states in a trajectory[7]. An illustration of both cases are depicted in Fig 1a. Thus, collective mutual exclusive expressions among genes are indications for the possibility of establishing discrete stratifications within a cell population, either in the form partitioning cells into discrete cell types, or in the form of establishing temporal discrete cell states within a trajectory[8].

There have been a number of attempts at extracting markers, including the most extreme exclusively-expressed ones, independent from clustering. Among them, the most notable ones include SEMITONE[9], scHayStack[10], M3Drop[11], DubStepR[12], Anti-Correlation[13], MarkerPen[14] and MarcoPolo[15]. The above clustering-free marker identification methods designate a gene as a marker gene either 1) if the gene is non-uniformly randomly distributed in a high-quality latent manifold, which is presumed to faithfully reflects both prominent and subtle differences between major-type and sub-type cell populations; or 2) if a gene or a gene pair has expression or co-expression patterns that statistically deviate from a null expression pattern, which is modeled as the default expression pattern of a (or a pair of) non-marker housekeeping gene(s). SEMITONE and scHayStack are two representative methods of the first class, where reference cells or grid points are sampled from the gene expression manifold and the cell distribution of a putative gene is compared against the reference cell/grid placement distribution for alignment assessment. M3Drop, Anti-Correlation, DubStepR, MarkerPen are methods of the second class. They differ by their choice of statistics and their null hypothesis proposals. M3Drop designates a gene as a maker if its dropout rate is much higher than a presumed housekeeping gene of the same average expression strength; Anti-Correlation tests if the anti-correlation between two genes are significantly higher than two presumed housekeeping genes; DubStepR checks the dynamic range between the maximum and the minimum Pearson correlations of a gene and selects the top ranking genes at each expression strength level; MarkerPen expands a pre-defined seed marker gene set by devising a prior-knowledge-incorporated sparse principle component analysis. There are also methods that are a combination of both classes, such as MarcoPolo, which simultaneously tests the dispersion, the correlation and the non-uniform-random distribution in the manifold of genes and gene pairs.

The above clustering-free marker identification methods have improved downstream analysis, with varying degree of success, but their effectiveness at exposing nuanced biological differences remains limited. For nuanced sub-type and sub-state markers, despite the evident collective mutually exclusive expression patterns in between certain gene groups, due to their relatively weak expression strength and the their high-dropout percentage, many of them have insignificant variation or correlation statistics, leading to their small presence in the latent manifold. Consequently, they typically do not emerge as the top ranking markers in the marker selection output. In addition, most of these algorithms aggregates all putative markers into a single bunch, omitting the valuable inter-gene relationships. As such, the candidate markers can only be interpreted with a post-hoc analysis through clustering and differential

expression analysis, which essentially reverts these marker identification algorithms to feature selection pre-processing[16].

While conceptually well-defined and simple at first glance, mutual exclusivity is in fact not a concept natively compatible with single cell transcriptomics. By nature, mutual exclusivity is associated with binary random events and describes the pattern that two binary random events never co-occur. Single cell transcriptomic data, on the contrary, represent continuous molecular counts. This per-cell mRNA count data is riddled with low RNA recovery rates[17], doublets[18] and ambient RNA molecules[19]. As such, to measure mutual exclusivity in a canonical sense, gene-specific thresholds must be introduced to convert molecular counts to binary expression status for individual genes, priory to testing for mutual exclusivity. This methodology, unfortunately, requires extensive parameter tuning, which reduces the usability, the consistency and the confidence in mutually exclusive expression identification.

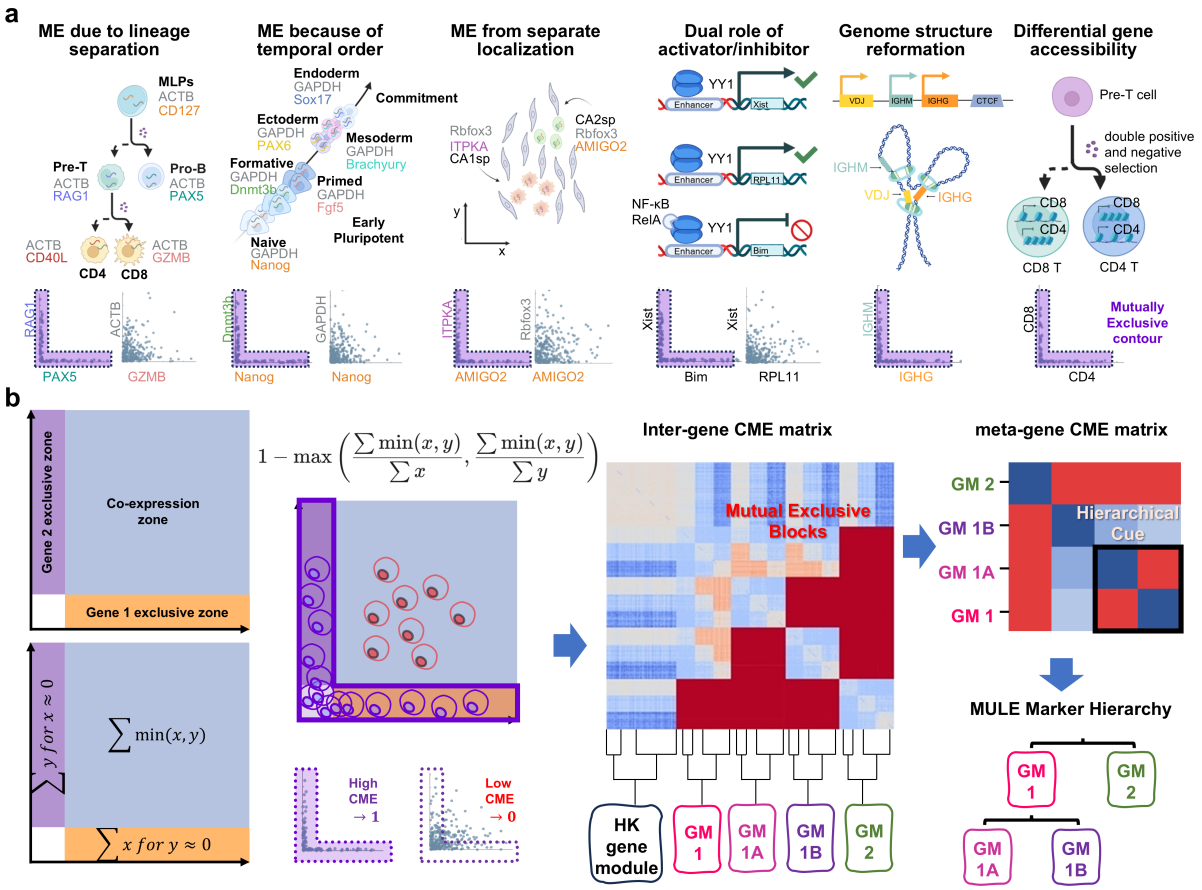

Fig. 1: a) Mutual exclusivity in biology. b) MULE overview. MULE first defines the CME score as a measure of mutual exclusivity between gene pairs. It then computes a symmetric mutual exclusivity matrix across all genes, groups genes into gene modules, and subsequently reconstructs the hierarchical marker hierarchy.

To address the above challenges, here we propose MULE, an end-to-end truly clustering-free MUtualy Exclusive paralleL gene program identification algorithm (Fig1b). MULE not only identifies markers of both prominent and nuanced cell types or states, but also categorizes markers that co-define the same populations into individual gene sets, and organizes the gene sets into a major-to-subtype hierarchy. Our contribution in this paper is three fold: First, we developed a novel, computational-friendly non-parametric index, continuous mutually exclusive (CME) index for mutual exclusive expression identification in single cell transcriptomic data. CME is designed to identify mutual exclusivity between individual genes specifically for the sparse, high-dropout, high dynamic range single cell transcriptomics molecular count data. Second, MULE includes a CME validation regime through permutation testing, where phony mutual exclusivity arose from dropouts are being identified and removed. In this fashion, MULE prevents premature designation of sporadic mutual exclusivity stemmed from random chance.

Third, MULE does not simply extracts all gene pairs with high CME scores and consolidate them into a single gene batch, rather, MULE decomposes and organizes genes into a mutually exclusive hierarchy based on a global inter-gene CME score spectrum. Thus, MULE achieves end-to-end mutually exclusive gene program identification without incurring clustering cells in a preconceived transcriptomic manifold.

We benchmarked CME with a series of simulated and real scRNA-seq and spatial transcriptomic (ST) datasets. Experiments demonstrate that MULE is highly effective at accurately identifying collective mutual exclusivity between gene programs. With the identified mutually exclusive genes, we further show that MULE facilitates the partitioning of cell sub-populations with nuanced differences, separating temporal non-overlapping cell states in refined developmental trajectories, and quantifying co-localization and repulsion between cells expressing respective biological functions.

Method

Mutually Exclusive Expression Pattern Identification and Validation

Continuous mutually exclusive score. We define mutual exclusive gene expression as a pair of genes that has low co-expression but robust individual expressions. Formally, let $\mathbf{C}$ represent the entire cell population and $\mathbf{G}$ denote the set of genes that have non-trivial expression in $\mathbf{C}$. Assume there is a pair of genes $g_i, g_j \in \mathbf{G}$, where the expression patterns of g_i, g_j are binary: there exist sub-sets of cells have significantly above-background expression of g_i and g_j. We denote the cell populations that highly express g_i and g_j as $\mathbf{C}_{g_i}$ and $\mathbf{C}_{g_j}$, respectively. Note that $\mathbf{C}_{g_i}$ and $\mathbf{C}_{g_j}$ are biological concepts that represent cells with specific expressions of g_i and g_j, which is not necessarily synonymous to cells with high mRNA read counts. Due to technical limitations such as dropouts, a cell $c \in \mathbf{C}_{g_i}$ that expresses g_i could have none of the mRNA recovered because of sequencing randomness. We designate g_i and g_j as a mutually exclusive gene pair if neither $|\mathbf{C}_{g_i}|$ or $|\mathbf{C}_{g_j}|$ is trivial, and $|\mathbf{C}_{g_i} \cap \mathbf{C}_{g_j}|$ is negligent in comparison to $|\mathbf{C}_{g_i} \cup \mathbf{C}_{g_j}|$, which could be quantified as having a close to zero Jaccard Index score.

While $\mathbf{C}_{g_i}$ and $\mathbf{C}_{g_j}$ memberships are implicit and its accurate designation is complicated by factors such as dropouts and the unimodal-shaped distribution in transcriptomic data, the topological patterns of mutually exclusive gene pairs in their 2-dimensional panels are apparent and consistent. If two genes are truly mutually exclusive, then on a 2-dimensional scatter plot, cells are distributed in a L shape where there are disproportionally more cells having high expressions in either gene, than cells that exhibit high expressions in both genes. This pattern is consistent irrespective of dropout intensity variations, background expression strength changes, or $\mathbf{C}_{g_i}$ to $\mathbf{C}_{g_j}$ population size ratio differences.

The continuous mutually exclusive (CME) score is designed as a surrogate for the disproportionatity between co-expression and exclusive expression. Let $\boldsymbol{g_i}$ and $\boldsymbol{g_j}$ denote the median-normalized expression values of g_i and g_j across all cells in $\mathbf{C}$, and $g_{\cdot,k}$ denote the UMI count of gene $g_\cdot$ for cell $c_k \in \mathbf{C}$. CME is defined as:

$$CME_{i,j} = 1 - max(\frac{\sum_k min(g_{i,k}, g_{j,k})}{\sum_k g_{i,l}}, \frac{\sum_k min(g_{i,k}, g_{j,k})}{\sum_k g_{j,l}}). \tag{1}$$

In Equation 1, $\sum_k min(g_{i,k}, g_{j,k})$ measures the collective co-expression strength, while $\sum_k g_{\cdot,k}$ measures the total expression strengths of the respective genes. For a mutually exclusive gene pair, $CME_{i,j}$ approaches towards 1 as $\sum_k min(g_{i,k}, g_{j,k}) \to 0$ while $\sum_k g_{\cdot,k} \neq 0$ (genes that has no expression across all cells is dropped in quality control). For a non-mutually exclusive gene pair, $CME_{i,j}$ is expected to vary between $[0, 0.5]$, where $CME_{i,j} = 0.5$ indicates independence between g_i and g_j and $CME_{i,j} = 0$ indicates a perfectly linear relationship between g_i and g_j.

Spurious mutual exclusivity detection and correction. Occasionally, phony mutual exclusivity could arise from non-mutually exclusive gene pairs with high dropouts. That is: there is extensive overlap between $\mathbf{C}_{g_i}$ and $\mathbf{C}_{g_j}$ but we have $CME_{i,j} \to 1$. To see why phony mutual exclusivity occurs, we refer to the example in Appendix A. In this example, $\mathbf{C}_{g_i}$ and $\mathbf{C}_{g_j}$ are exactly the same population, however, due to severe dropouts in both genes, there are very few cells that co-express both g_i and g_j, leading to an undesirable high CME score.

To identify and subsequently remove spurious mutual exclusivity, MULE employs a hypothesis testing scheme. The full mathematical framework of the statistical test is included in Appendix B. Briefly, if two genes are independent, their CME score should follow a normal distribution with its mean and variance dependent on the non-zero population sizes. For gene pairs with high-dropouts, their CME distributions have means close to 1 but extremely small standard deviations. Spurious mutual exclusivity between high-dropout genes, therefore, could be identified and rectified by checking their p-values. Authentic

mutually exclusive gene pairs are not affected as their CME values are too large, despite insignificant in numerical differences but statistically profound, to follow the null distribution.

In practice, MULE employs a random permutation test, to approximate the null CME distribution. The decision to approximate the exact test stems from two considerations: 1) Fitting the null distribution model is computational intensive. 2) The analytical null distributions do not fully capture the subtle inter-dependence introduced through CP10K normalization, where UMI counts in cells are normalized to 10K. By default, MULE randomly shuffles the gene expressions gene-wise across all cells for 50 times, adjusts the UMI counts through CP10K normalization (cell-wise normalization) after each shuffle, and recomputes the CME scores between gene pairs. The CME scores after shuffling produces an empirical null distribution. Only the gene pairs whose CME score is close to 1, while maintaining a top ranking in the null CME score array is designated as the high-confidence mutually exclusive gene pairs.

Co-exclusive Gene Module Construction through Clustering

Assume there is a hypothetical biological partitioning of a cell population $\mathbf{C}$ into discrete and non-overlapping sub-populations $\mathbf{C}_\alpha$ and $\mathbf{C}_\beta$. Assume each sub-population is associated with a set of marker genes $\mathbf{G}_\alpha$ and $\mathbf{G}_\beta$ respectively that are specifically expressed in $\mathbf{C}_\alpha$ and $\mathbf{C}_\beta$. Here specific expression is again defined as exhibiting binary expression patterns of low and high expression, which includes but is not limited to zero and non-zero relationship. MULE aims to not only identify member genes in $\mathbf{G}_\alpha$ and $\mathbf{G}_\beta$, but also organize them into respective sets where there is strong mutual exclusivity between genes across the set, but no mutual exclusivity within each set.

MULE organizes genes by the consensus in their mutual exclusive relationship against other putative low-dropout marker genes. While the spurious mutual exclusivity correction scheme is capable of correcting some of the phony mutual exclusivities between high-dropout genes, its capability has limitations: between two co-expressing genes g_i and g_j of extremely high dropouts, it is not entirely uncommon that $\sum_k min(g_{i,k}, g_{j,k}) = 0$. For these extreme dropout genes, the permutation test alone is incapable of discerning authentic mutual exclusivity from phony mutual exclusivity.

While true mutual exclusivity between extreme dropout genes cannot be accurately described with just the expression information between these two genes, their relationship can be indirectly established by analyzing their higher-order relationship to other genes. If a pair of high CME genes g_i, g_j are markers of two biologically distinct populations $\mathbf{C}_\alpha$ and $\mathbf{C}_\beta$, and $\mathbf{C}_\alpha$ and $\mathbf{C}_\beta$ have non-trivial marker gene sets $\mathbf{G}_\alpha$ and $\mathbf{G}_\beta$ with $g_i \in \mathbf{G}_\alpha$ and $g_j \in \mathbf{G}_\beta$, then we expect g_i to share high CME scores with all member genes in $\mathbf{G}_\beta$, high-dropout or otherwise; while having low CME scores with the low-dropout genes in $\mathbf{G}_\alpha$; g_j and $\mathbf{G}_\alpha$ vice versa.

Mutual exclusivity quantified by CME between high-dropout and low-dropout genes are relatively more robust. If two genes $g_i, g_j \in \mathbf{G}_\alpha$, with g_i being a high-dropout gene and g_j having non-zero values in a majority of $\mathbf{C}_\alpha$, then the likelihood that non-zero counts of g_i and g_j being distributed to completely non-overlapping sub-populations in $\mathbf{C}_\alpha$, thus gives rise to $\sum_k min(g_{i,k}, g_{j,k}) = 0$, diminishes exponentially as the coverage of g_j in $\mathbf{C}_\alpha$ expands. Accordingly, MULE expects all markers in $\mathbf{G}_\alpha$ to share similar mutual exclusivity profiles against other low-dropout genes: for all genes in $\mathbf{G}_\alpha$, it must share high CME scores against low-dropout genes in other marker gene set $\mathbf{G}_{\neq\alpha}$, and low CME scores against low-dropout genes in $\mathbf{G}_\alpha$.

MULE identifies low-dropout genes by utilizing a non-parametric test of the relationship between the non-zero expression percentage and non-zero median expression. The detailed framework is included in Appendix C. Briefly, a low-dropout gene g_i is defined as a gene with non-zero expression values in the majority of its corresponding cell population $\mathbf{C}_{g_i}$. While the membership of $\mathbf{C}_{g_i}$ is implicit, there is an intrinsic relationship between expression strength and dropout percentage: as the gene expression strength increases, the fraction of zero-value cell population decreases, until it saturates when all cells in $\mathbf{C}_{g_i}$ report non-zero UMI values. Putative low-dropout marker genes are identified as genes that have robust expression in its respective population, indicated by a large non-zero median UMI count, but disproportionally small non-zero population sizes, suggesting that the expression has saturated in its corresponding population.

Hierarchical Ordering of Gene Modules through Mutual Exclusivity Graph Deconvolution

MULE constructs marker gene modules from the cross-module exclusivity (CME) matrix through agglomerative hierarchical clustering, which progressively merges genes or clusters that minimize the increase in within-cluster variance according to Ward's linkage criterion [20]. Once the hierarchical dendrogram is

formed, the optimal number of clusters is determined by locating the knee point of the within-cluster variance curve, achieving a balance between overly coarse and overly fragmented partitions. Each resulting cluster defines a coherent marker gene module. To quantify mutual exclusivity between modules, CME scores are aggregated across all inter-module gene pairs by averaging their pairwise CME values. The statistical significance of module–module exclusivity is then assessed via a permutation test, where gene labels are shuffled while preserving module sizes to compute empirical p-values. Details of this method are provided in Appendix D.

This procedure ensures that the detected module-level exclusivity reflects genuine biological structure rather than chance fluctuations. After obtaining the module–module exclusivity relations, we construct an undirected *mutual exclusivity graph*

$$G = (V, E),$$

where each node $v_i \in V$ corresponds to a gene module M_i, and an edge $(v_i, v_j) \in E$ exists if modules M_i and M_j are significantly mutually exclusive. Because mutual exclusivity is a qualitative property—either two modules exhibit exclusivity or not—we treat G as an *unweighted* graph.

Intuitively, exclusivity only arises between distinct transcriptional programs, such as between different major lineages or functionally opposing states. Modules within the same lineage, in contrast, should not be mutually exclusive but instead exhibit some degree of co-expression or regulatory overlap. To reveal this internal organization, we construct the *complement graph*

$$G^c = (V, E^c), \quad E^c = \{(v_i, v_j) \mid (v_i, v_j) \notin E,\, i \neq j\},$$

where edges now connect *non-exclusive* module pairs. Thus, G^c naturally captures hierarchical relations among modules: modules that share many non-exclusive connections likely belong to the same broader lineage, while those with few connections represent more specialized subtypes.

To infer this hierarchy, we perform a *breadth-first search (BFS)* traversal on G^c. We first identify the root module as the node with the highest degree:

$$v_{\text{root}} = \arg\max_{v_i \in V} \deg(v_i), \tag{2}$$

where $\deg(v_i)$ denotes the number of connected non-exclusive modules. This module serves as the representative of a broad transcriptional lineage that connects to many descendant submodules. The BFS traversal then assigns hierarchical levels according to geodesic distance from the root:

$$\text{level}(v_i) = \text{dist}_{G^c}(v_{\text{root}}, v_i), \tag{3}$$

where dist_{G^c} is the shortest-path distance in the complement graph. Modules with smaller level$(\cdot)$ values correspond to broader, high-level programs, while deeper nodes represent more refined subtypes within these programs.

This topological construction provides a biologically consistent interpretation: mutual exclusivity delineates distinct cell lineages, while the complement graph reconstructs the nested relationships among their constituent subtypes.

Mutual-Exclusivity-Guided Cell Embedding Refinement via Metric Learning

We aim to construct a cell embedding that reflects the natural hierarchical structure of single-cell transcriptomic data, where major cell groups are mutually exclusive and each group further branches into subtypes or states. Mutual exclusivity provides a principled way to push apart cells belonging to distinct groups, and through metric learning, we can further optimize the embedding to faithfully preserve this hierarchical separation. To extract such structure, we first derive a gene tree whose internal nodes represent meta-genes—gene modules defined by coherent co-activation or exclusivity. The expression of a meta-gene in cell i is given by

$$M_{iu} = \frac{1}{|G_u|} \sum_{g \in G_u} X_{ig},$$

and each cell is assigned to the module with maximal activation, $y_i = \arg\max_u M_{iu}$. This produces a hierarchical supervision signal aligned with the mutual-exclusivity structure of the data.

We embed each cell into a structured low-dimensional space that preserves biologically meaningful variation while remaining compatible with hierarchical metric constraints. To achieve this, we learn a linear embedding $z_i = Wx_i$ together with a tied linear decoder $\hat{x}_i = W^\top z_i$. A reconstruction penalty,

$$L_{\text{recon}} = \|x_i - \hat{x}_i\|_2^2,$$

is introduced to maintain interpretable expression structure and prevent excessive geometric distortion caused by metric learning.

The central component is a hierarchical triplet loss [21] whose margin depends on the depth of the lowest common ancestor (LCA) in the gene tree. Pairs diverging high in the tree (distinct major lineages) receive strong repulsion, while pairs diverging low (subtypes within the same lineage) receive weaker repulsion:

$$L_{\text{triplet}} = \max\Big(0,\ \|z_a - z_p\|_2 - \|z_a - z_n\|_2 + \gamma\big(L_{\max} - h(a,n)\big)\Big).$$

To further stabilize subtype structure, we include an intra-class compactness penalty, $L_{\text{compact}} = \|z_a - z_p\|_2$, encouraging samples belonging to the same meta-gene module to cluster more tightly.

The final loss function balances hierarchical separation, subtype compactness, and expression-level fidelity:

$$L = \alpha L_{\text{triplet}} + \beta L_{\text{compact}} + \lambda L_{\text{recon}}.$$

This framework transforms the mutual-exclusivity structure encoded in the gene tree into geometric constraints in the embedding space, yielding a representation that simultaneously preserves major lineage separation and within-lineage subtype organization of underlying gene-expression signals.

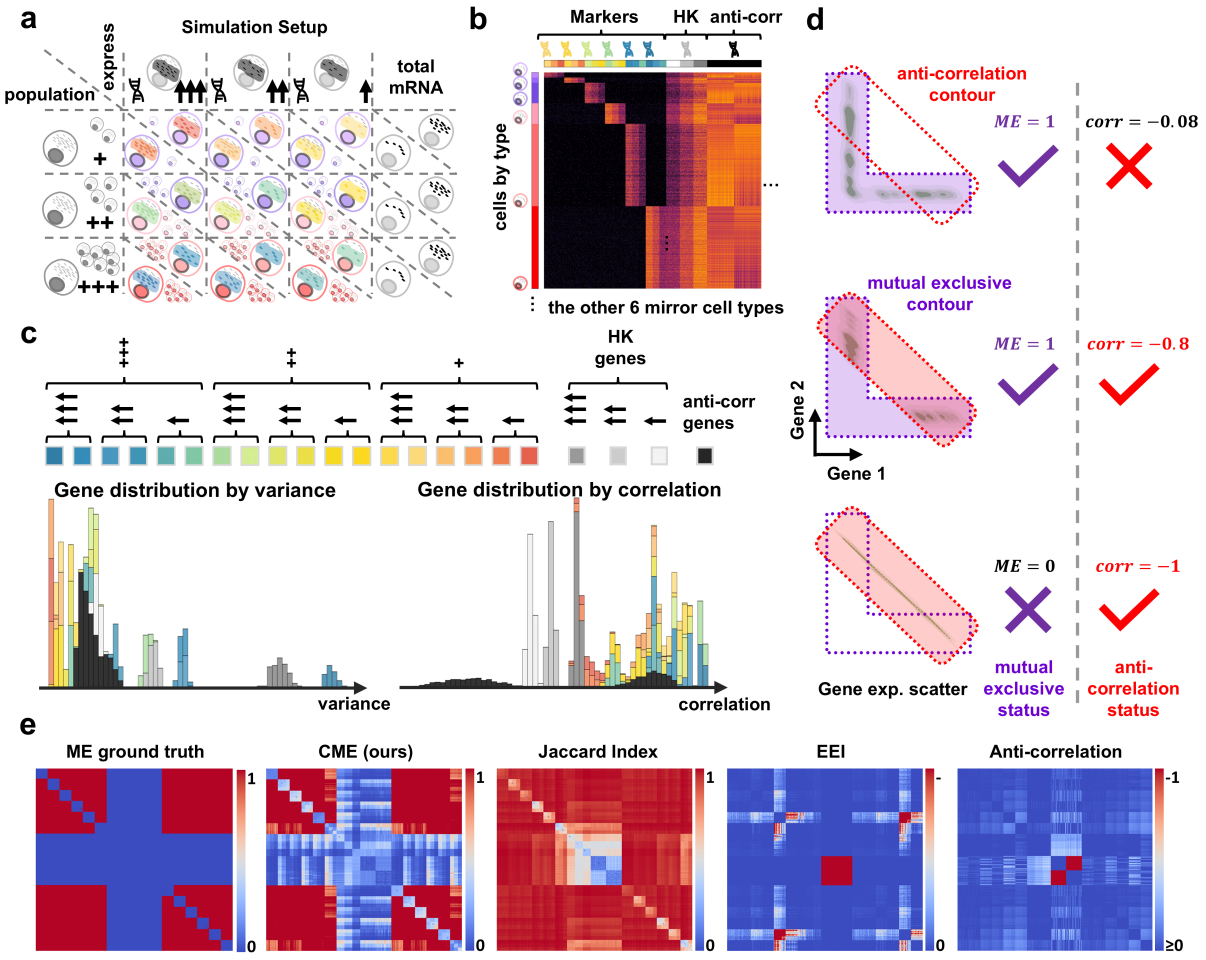

Fig. 2: a) Simulation setup. b) Gene expression heatmap of simulation data. c) The inherent dependency of variance and correlation against expressn strength and population size. d) Marker selection by variance and correlation is biased in nature. e) Benchmark heatmap of gene statistical scores.

Results

MULE consistently identifies collective exclusively expressed gene modules of both prominent and subtle cell populations

Current marker identification algorithms are biased toward highly expressed genes in abundant cell types. This bias arises from the dependence of variance or correlation metrics on gene expression strength, cell abundance, and transcriptome size. Fig 2 illustrates the dependency of variance and correlation on cell and gene configurations. We simulated 12 cell types, each with strong, mild, or weak exclusive markers. We also simulated a set of cell-type-unrelated anti-correlation genes, accompanied with three sets of strong-, mild- and weak-expressing housekeeping genes. The 12 cell types are then partitioned into 6 pairs of variate population size and sequencing depths, with the two cell types in a pair sharing the same configurations. The simulated gene-cell UMI matrix is depicted in Fig 2b.

Fig 2c presents the ranking of variance and Pearson correlation between cell-type exclusive marker genes, housekeeping (HK) genes and anti-correlation genes. For both variance and correlation, we observe the following trend: stronger markers and markers of abundant cell types show higher variance and correlation, while shallow-transcriptome markers have higher variance but lower correlation. Genes of shallow-transcriptome-size cell types exhibit greater gene expression variations than their deep-transcriptome-size counterparts of the same categories because their variations are amplified through CP10K normalization (Appendix E). Surprisingly, highly expressed housekeeping genes also show high variance and correlations, often exceeding exclusive markers. The above results demonstrate the unreliable and the biased nature of using variation or correlation as a primary index for marker selection.

The idea of exploiting mutual exclusivity in gene expression, especially for marker identification, has been explored in prior works, but a proper index for scRNA-seq remains to be established. In their method SINCERA, Scott R. Tyler et al. proposed using anti-correlation as a surrogate for mutual exclusivity. Natsu Nakajima et al.[22] designed a modified binary expression metric, the exclusive expressed index (EEI), that measures if there is extensive two-way exclusive expression (one gene having zero expression and the other gene having non-zero expression) between two genes in a gene pair. Both anti-correlation and EEI, however, do not adapt well to the high-dropout, variance highly instable single cell gene expression data. Neither are approriate metrics for mutual exclusivity in scRNA-seq data: 1) Anti-correlation measures a statistical event is inherently different from mutual exclusivity. mathematically, the two are quite disconnected concepts. Pearson correlation, and its non-parametric counter part Spearman correlation, measures a continuous value or ranking dependency between two random variables. In the case of expression mutual exclusivity, the continuous dependency is absent. 2) EEI is the p-value of a dependency test, but not mutual exclusivity test. EEI tests how much a pair of gene, after converting to binary expression, deviates from a null hypothesis where the two genes are assumed to be completely independent. As seen in Fig 2e, EEI has a broad dynamic range and its value is again associated with the expression strength of the gene and the abundance of the cell type. The CME index of MULE accommodates the high-dropout, ultra-sparse and the variance-instable nature of single cell RNA-seq data and faithfully reflects the mutual exclusivity between markers. By focusing on the minimum ratio between the co-expression strength and the cumulative gene expression strengths, the dependency between CME and the gene/cell configuration is significantly mitigated, as Fig 2 shows. In comparison to anti-correlation and EEI score, CME scores between marker genes are consistently high across the board, irrespective of the gene expression strength and/or the expressing cell type abundance. In addition, the CME scores between markers and housekeeping genes are consistently low, so are the CME scores in-between housekeeping genes. When consolidated, we observe a clear separation between housekeeping genes and marker genes in their CME scores, with the exclusive markers consistently rank top in CME scores. We also observe that the interlinking between marker expression strength, cell type relative abundance and CME scores is much attenuated. These prove that CME is a stable and reliable metric for identifying and extracting mutual exclusive expressions among genes.

MULE not only identifies mutual exclusive expressions between gene pairs, it also correctly organizes genes into respective co-exclusive modules. The gene module partitioning capability is a key feature that sets MULE apart from the rest of the marker identification algorithms. In comparison to SINCERA and DubStepR, MULE does not simply designate a gene a marker because it has more than a certain number of above-threshold CME scores; nor does it examine the maximum range of the CME scores. Instead, MULE partitions genes into respective modules through normalized max-cuts. Appendix E shows the overall accuracy of marker identification in respective to variant marker configurations. From the table we see that MULE correctly identifies and stratifies exclusive markers of both strong and week expression strength, across all cell types irrespective of their cell type abundance and transcriptomic

size. scHayStack, SEMITONE, GiniClust3, SINCERA and DubStepR, on the other hand, only lumps all putative markers into a single marker collection, but also biases the identification towards strong markers of the most abundant cell type with large transcriptome sizes.

MULE identifies reliable and consistent lineal relationships between co-exclusive gene modules in complex multi-granularity cell type or cell state configurations

From an pure analytical perspective, in the context of clustering, it is a natural tendency to accept and embrace the concept that cells could be organized into a flat, discrete categories where their identities are characterized by their exclusive markers. However, it is important to be always mindful that cell types or cell states are artificial constructs that help us establish orders in the cell biology system. From a philosophical perspective, one can always cluster and sub-cluster ad infinitum. Therefore, it is important to adopt a principled approach for transparent and sensible major-to-minor cell partitioning and categorization.

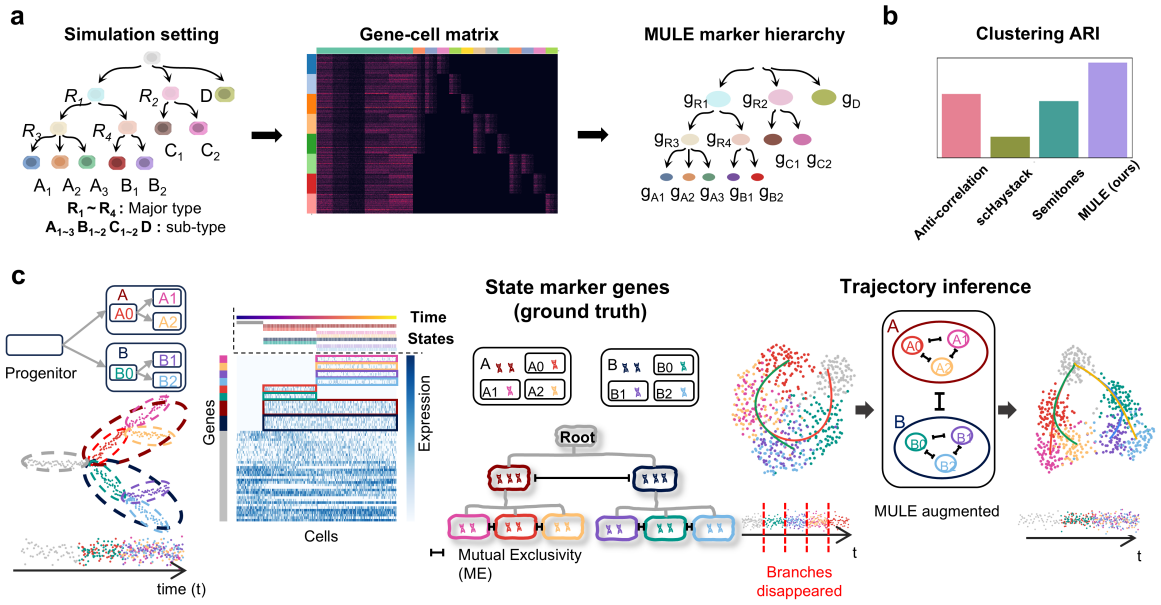

Fig. 3: a) MULE identifies both markers as well as the hierarchical structure of markers. b) MULE facilitates accurate clustering to sub type in BMMC dataset. c) MULE identifies sub-states in trajectory analysis.

Analogous to cell nomenclatures being organized in a multi-tiered hierarchical fashion, we take the perspective that the cell-type exclusive markers, which are essentially cell type surrogates in the gene space, should also reflect the hierarchical nature of the cell type nomenclature. MULE aptly fulfills this objective by exploiting the presence or the absence of mutual exclusivity between co-exclusive gene modules. Fig 3a demonstrates the mutual exclusivity hierarchy extracted from the mutually exclusive expression gene graph of a simulated multi-tiered multi-cell-type scRNA-seq dataset. Fig 3 shows the MULE inferred co-exclusive gene hierarchy. As shown in the figure, MULE accurately captures the hierarchical relationship of the co-exclusive gene modules, where genes are correctly organized into top-tier, mid-tier and low-tier gene modules with parent-to-child relationships that correspond to the ground truth major-to-minor cell type hierarchal nomenclature.

A key advantage of MULE is that MULE can consistently infer co-exclusive gene modules and their hierarchical structure under a wide range of population abundance imbalances and variations. Appendix F demonstrates the co-exclusive identification results of MULE in comparison to the simulation ground truth, under various population imbalances between the major cell types, as well as in between mid-tier cell types and terminal sub-types within major and mid-tier cell types, respectively. It can be observed that MULE consistently reconstructs authentic major-to-minor hierarchies across a wide range of population abundances. In comparison, None of the popular gene selection algorithms supports major-to-minor hierarchical gene organization. In terms of co-exclusive gene detection sensitivity, all three algorithms

include a large fraction of non-exclusive genes, while missing a large fraction of true exclusive gene. Among the identified exclusive genes, all three algorithms are clearly biased towards the identification of major-type markers and strong-expressing markers. This is expected as the sub-cell-type populations are typically smaller in size, which de-prioritize the ranks of their respective markers as we have thoroughly investigated in the previous section.

We then applied MULE to a public BMMC scRNA-seq dataset produced by [23]. In this dataset, with the aid of surface markers that are co-sequenced via CITE-seq technology, we can manually curate 34 sub-types spanning across 4 major cell types. Appendix G shows a truncated gene-list for each co-exclusive gene module and the co-exclusive gene module hierarchy extracted by MULE. The full list of co-exclusive putative markers for each modules is supplied in Appendix G. As for high-resolution clustering enhancement, MULE-augmented clustering-and-visualization result of the BMMC RNA data, in contrast to clustering with genes derived from ranking by variance, putative markers produced by SINCERA(Anti-correlation), and features genes extracted from DubStepR. The ARI scores of the clustering results, using manual cell type curation by surface marker expression as the ground truth, are also consolidated in Appendix G. Overall, MULE augmented clustering produces the best clustering performance in comparison to other gene selection algorithms that most matches the manual cell type curation.

Identifying the hierarchical co-exclusivity relationship among genes is also important in trajectory analysis. From a developmental perspective, mutual exclusivity among genes in a developmental snapshot denotes that two genes share no temporal overlap and it is highly likely that there is a clear temporal order between the two genes as Fig3c shows. Collective mutual exclusive expression among gene modules is a strong indicator for time-synchronized co-expression. This could arise from cells transitioning between discrete, transcriptomically distinct steady states, or cells committing to separate fates through differentiation. For Fig3c case, mutual exclusivity outlines the discrete gene expression sub-states that could span across multiple cell fate trajectories, awaiting to be connected by a pseudo-time trajectory. Similar to the case in clustering, the hierarchical structure among co-exclusive genes provides valuable information and must be preserved. In a multi-pronged cell differentiation trajectory, the hierarchical co-exclusive expression establishes the temporal relationships of differentiation events and precisely in which branch a sub-split takes place. In single-branch trajectories, hierarchical co-exclusivity reveals major states and layered sub-states. Fig 3c and Appendix H demonstrates the MULE-augmented trajectory analysis by Slingshot for a multi-pronged differentiation simulation and a multi-layered-cell-state single trajectory analysis simulation.

The incorporation of mutual exclusivity provides a crucial constraint for optimizing the cell embedding: cells expressing mutually exclusive gene pairs are naturally pushed apart in the embedding space, enhancing separation between different developmental stages or differentiation branches, while cells within co-expressed modules remain closely clustered, preserving continuous state transitions. This mutual-exclusivity-driven geometric constraint produces embeddings that better reflect the temporal and fate relationships inherent in developmental biology, resulting in more robust pseudo-time inference and clearer branch identification. In other words, mutual exclusivity not only captures the antagonistic relationships between transcriptional programs but also imposes a natural geometric framework that encodes both temporal directionality and hierarchical differentiation structure in the embedding space. We also give the trajectory analysis results for a real Pancreas [24] ductal-to-islet-cell differentiation trajectory in Appendix H. The above results demonstrate that correctly identifying mutual exclusivity and accurately portraying hierarchical co-exclusivity among gene modules is important and valuable in trajectory analysis.

MULE highlights the differential localization and co-localization of cell types in spatial niches by examining the persistent and the mitigated mutual exclusivity in spatial transcriptomics

In the context of spatial transcriptomes (ST), mutual exclusivity serves as a proximity indicator between a pair of genes or gene groups. If two groups of genes exhibit sustaining mutual exclusivity in a ST, defined as demonstrating no co-expressions in the same spatial spot, then it means that in addition to no concurrent expressions in the same cell, cells that separately express the respective gene modules are also spatially distant on a tissue slice. In contrast, when two gene modules that are evidently mutually exclusive in scRNA-seq data, but exhibit significantly mitigated mutual exclusivity in ST, then it indicates that there exist non-trivial overlaps between the respective cell populations in the spatial context.

Appendix I demonstrates MULE effectively identifies spatial domains within tissue sections by detecting mutually exclusive gene modules that correspond to distinct anatomical regions. In the myocar-

dial infarction dataset[25], MULE-revealed modules accurately delineate the infarct, border, and remote zones, consistent with known pathological structures. This demonstrates MULE's capability to uncover biologically meaningful tissue domains solely from gene expression patterns, without relying on spatial coordinates.

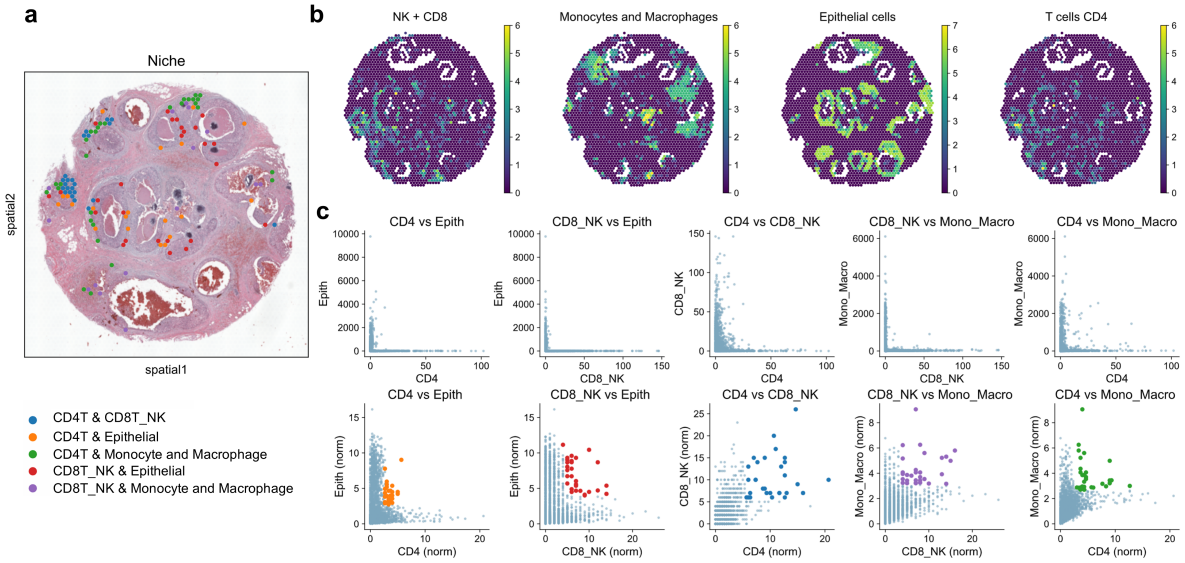

Fig. 4: a) Spatial co-localization of immune cells identified by MULE. b) Spatial expression patterns of different immune cell types visualized using CYTOSPACE[26] in the spatial transcriptomics map. c) Scatter plots showing the mutually exclusive expression patterns among cells, derived from the meta-features of gene modules identified by MULE.

MULE not only detects spatial gene-expression patterns and delineates tissue domains but also identifies regions where distinct cell types co-localize. It does so by pinpointing ST spots with significant co-expression of two gene sets previously defined as mutually exclusive in scRNA-seq data, using a compound expression score. To illustrate this, we applied MULE to a public breast cancer scRNA-seq dataset [27], identifying 11 coarse-grained co-exclusive gene modules corresponding to 11 cell types: six immune subsets, perivascular-like cells (PVL), plasma cells (PC), epithelial (including malignant) cells, endothelial cells, and fibroblasts. We merged NK and CD8 T cell markers into one module based on their shared cytotoxic program, and combined monocyte and macrophage markers by lineage. These meta-markers retained mutual exclusivity in scRNA-seq (Fig4c). Transferring these modules to a breast cancer ST dataset, we observed epithelial meta-marker expression forming tumor cell "encirclements" of lymphocytes (NK, CD8 , and CD4 T cells), mirroring the scRNA-seq exclusivity (Appendix I). We also detected sporadic co-expression of lymphocyte and epithelial markers, indicating tumor lymphocyte infiltration (TIL). Additionally, myeloid and lymphoid meta-markers co-localized near tumor TIL regions, suggesting nascent tertiary lymphoid structure formation in tumor-adjacent stroma, priming immune cells for tumor toxicity.

Conclusion

In conclusion, we expect numerous biological and medical research to benefit from MULE. MULE could be highly effective for research problems where the main focus is to identify unknown makers and to reliably partition cells of nuanced differences into distinctive sub-groups in a complex tissue with tranches of cell type classifications. MULE also help elicit subtle but decisive differences within a seemingly homogeneous cell population and help bring focus to the most biologically significant divergences within the population, especially when such differences are coherently registered in a limited number of key genes that do not rank top in expression strengths or variations. We anticipate that MULE will uncover numerous additional sub-types in healthy or pathological tissues and organisms that are defined by important low-profile markers, thereby advancing our understanding of biology and diseases. MULE is publicly available at https://github.com/Carroll105/MULE.

References

1. Giovanni Ciriello, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research*, 22(2):398–406, 2012.
2. M Sensi, G Nicolini, C Petti, I Bersani, F Lozupone, A Molla, C Vegetti, D Nonaka, R Mortarini, G Parmiani, et al. Mutually exclusive nrasq61r and brafv600e mutations at the single-cell level in the same human melanoma. *Oncogene*, 25(24):3357–3364, 2006.
3. Elisa Rodríguez-Seguel, Nancy Mah, Heike Naumann, Igor M Pongrac, Nuria Cerdá-Esteban, Jean-Fred Fontaine, Yongbo Wang, Wei Chen, Miguel A Andrade-Navarro, and Francesca M Spagnoli. Mutually exclusive signaling signatures define the hepatic and pancreatic progenitor cell lineage divergence. *Genes & development*, 27(17):1932–1946, 2013.
4. Brian K Hall. *Evolutionary developmental biology*. Springer Science & Business Media, 2012.
5. Molly Lewis, Veronica Cristiano, Brenden M Lake, Tammy Kwan, and Michael C Frank. The role of developmental change and linguistic experience in the mutual exclusivity effect. *Cognition*, 198:104191, 2020.
6. Hongkui Zeng. What is a cell type and how to define it? *Cell*, 185(15):2739–2755, 2022.
7. Charles Darwin, John Wyon Burrow, and John Wyon Burrow. *The origin of species by means of natural selection: or, the preservation of favored races in the struggle for life*. AL Burt New York, 2009.
8. Xiang Ge Luo, Jack Kuipers, and Niko Beerenwinkel. Joint inference of exclusivity patterns and recurrent trajectories from tumor mutation trees. *Nature communications*, 14(1):3676, 2023.
9. Anna Hendrika Cornelia Vlot, Setareh Maghsudi, and Uwe Ohler. Cluster-independent marker feature identification from single-cell omics data using semitones. *Nucleic Acids Research*, 50(18):e107–e107, 2022.
10. Alexis Vandenbon and Diego Diez. A clustering-independent method for finding differentially expressed genes in single-cell transcriptome data. *Nature communications*, 11(1):4318, 2020.
11. Tallulah S Andrews and Martin Hemberg. M3drop: dropout-based feature selection for scrnaseq. *Bioinformatics*, 35(16):2865–2867, 2019.
12. Bobby Ranjan, Wenjie Sun, Jinyu Park, Kunal Mishra, Florian Schmidt, Ronald Xie, Fatemeh Alipour, Vipul Singhal, Ignasius Joanito, Mohammad Amin Honardoost, et al. Dubstepr is a scalable correlation-based feature selection method for accurately clustering single-cell data. *Nature Communications*, 12(1):5849, 2021.
13. Scott R Tyler, Daniel Lozano-Ojalvo, Ernesto Guccione, and Eric E Schadt. Anti-correlated feature selection prevents false discovery of subpopulations in scrnaseq. *Nature Communications*, 15(1):699, 2024.
14. Yixuan Qiu, Jiebiao Wang, Jing Lei, and Kathryn Roeder. Identification of cell-type-specific marker genes from co-expression patterns in tissue samples. *Bioinformatics*, 37(19):3228–3234, 2021.
15. Chanwoo Kim, Hanbin Lee, Juhee Jeong, Keehoon Jung, and Buhm Han. Marcopolo: a clustering-free approach to the exploration of differentially expressed genes along with group information in single-cell rna-seq data. *bioRxiv*, pages 2020–11, 2020.
16. Dongyuan Song, Kexin Li, Xinzhou Ge, and Jingyi Jessica Li. Clusterde: a post-clustering differential expression (de) method robust to false-positive inflation caused by double dipping. *Research Square*, pages rs–3, 2023.
17. Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. Saver: gene expression recovery for single-cell rna sequencing. *Nature methods*, 15(7):539–542, 2018.
18. Christopher S McGinnis, Lyndsay M Murrow, and Zev J Gartner. Doubletfinder: doublet detection in single-cell rna sequencing data using artificial nearest neighbors. *Cell systems*, 8(4):329–337, 2019.
19. Matthew D Young and Sam Behjati. Soupx removes ambient rna contamination from droplet-based single-cell rna sequencing data. *Gigascience*, 9(12):giaa151, 2020.
20. Sinan Saraçli, Nurhan Doğan, and İsmet Doğan. Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of inequalities and Applications*, 2013(1):203, 2013.
21. Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European conference on computer vision (ECCV)*, pages 269–285, 2018.
22. Natsu Nakajima, Tomoatsu Hayashi, Katsunori Fujiki, Katsuhiko Shirahige, Tetsu Akiyama, Tatsuya Akutsu, and Ryuichiro Nakato. Codependency and mutual exclusivity for gene community detection from sparse single-cell transcriptome data. *Nucleic acids research*, 49(18):e104–e104, 2021.
23. Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.
24. Aimée Bastidas-Ponce, Sophie Tritschler, Leander Dony, Katharina Scheibner, Marta Tarquis-Medina, Ciro Salinno, Silvia Schirge, Ingo Burtscher, Anika Böttcher, Fabian J Theis, et al. Comprehensive single cell mrna profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development*, 146(12):dev173849, 2019.
25. DM Calcagno, N Taghdiri, VK Ninh, JM Mesfin, A Toomu, R Sehgal, J Lee, Y Liang, JM Duran, E Adler, et al. Single-cell and spatial transcriptomics of the infarcted heart define the dynamic onset of the border zone in response to mechanical destabilization. *Nature cardiovascular research*, 1(11):1039–1055, 2022.
26. Milad R Vahid, Erin L Brown, Chloé B Steen, Wubing Zhang, Hyun Soo Jeon, Minji Kang, Andrew J Gentles, and Aaron M Newman. High-resolution alignment of single-cell and spatial transcriptomes with cytospace. *Nature biotechnology*, 41(11):1543–1548, 2023.

27. Sunny Z Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R Torpy, Nenad Bartonicek, et al. A single-cell and spatially resolved atlas of human breast cancers. *Nature genetics*, 53(9):1334–1347, 2021.