mcRigor: A Statistical Method to Enhance the Rigor of Metacell Partitioning in Single-Cell RNA-seq and ATAC-seq Data Analysis

Pan Liu and Jingyi Jessica Li(✉)

Department of Statistics and Data Science, University of California, Los Angeles, CA, USA
jli@stat.ucla.edu

Abstract. In single-cell sequencing data analysis, addressing sparsity often involves aggregating the profiles of homogeneous single cells into metacells. However, existing metacell partitioning methods lack checks on the homogeneity assumption and may aggregate heterogeneous single cells, potentially biasing downstream analysis and leading to spurious discoveries. To fill this gap, we introduce mcRigor, a statistical method to detect dubious metacells composed of heterogeneous cells and optimize the choice of metacell partitioning methods and hyperparameters.

Keywords: Metacell partitioning · Single-cell RNA-seq · Single-cell ATAC-seq · Sparsity · Permutation

1 Introduction

The high sparsity of single-cell sequencing data, caused by low per-cell sequencing depth and technical sensitivity limitations, poses a significant challenge for data analysis. Metacell partitioning is a key strategy to mitigate sparsity as an alternative to imputation. Unlike imputation, which predicts missing values and risks introducing artifacts, metacell partitioning aggregates cells representing the same cell state into a metacell through averaging, using these metacells for downstream analysis [1]. Metacell partitioning is expected to reduce noise and thereby accentuate biological signals often obscured in sparse single-cell datasets. The metacell concept has been widely adopted in high-profile single-cell studies, and several metacell partitioning methods have been developed, including MetaCell [1], MetaCell2 [2], SuperCell [3], and SEACells [4]. However, the single-cell field lacks a rigorous definition of metacells or a universally accepted metacell partitioning strategy. Algorithm and hyperparameter choices introduce variability in metacell partitions across different methods, creating uncertainty about which partition best preserves biological signals and potentially compromising the reliability of metacell-based data analysis. To address this, we propose a statistical definition of metacells and develop mcRigor, a novel method to enhance the rigor of metacell partitioning.

S. Sankararaman (Ed.): RECOMB 2025, LNBI 15647, pp. 381–385, 2025.
https://doi.org/10.1007/978-3-031-90252-9_44

2 Methods

2.1 A Statistical Definition of Metacells

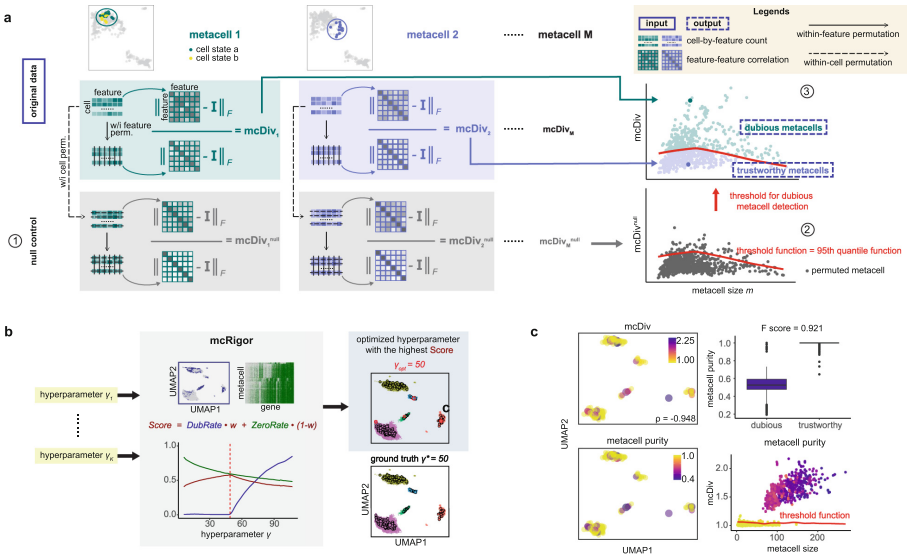

Fig. 1. a, Schematic of mcRigor for detecting dubious metacells. **b,** Schematic of mcRigor for optimizing metacell partitioning. **c,** mcRigor effectively assesses metacell heterogeneity and detects dubious metacells within the metacell partition by the MetaCell method on semi-synthetic data. Left: UMAP plots of metacells, colored by mcDiv values (top) and metacell purity (ground truth, bottom). Right: mcRigor identifies dubious and trustworthy metacells, aligning well with impure and pure metacells (top, with the F-score as the harmonic mean of precision and recall); thresholding mcDiv based on metacell size effectively identifies impure metacells (bottom).

Consider a total of n cells sequenced to measure the abundance of p features. The observed count matrix is denoted by $\mathbf{Y} = [y_{ij}] \in \mathbb{Z}_{\geq 0}^{n \times p}$, with y_{ij} as the count of feature j in cell i. The unobserved (true) relative abundance matrix is denoted by $\mathbf{\Lambda} = [\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n]^\top = [\lambda_{ij}] \in [0,1]^{n \times p}$, where λ_{ij} is the relative abundance of feature j in cell i, with $\sum_{j=1}^{p} \lambda_{ij} = 1$. We formalize the definition of metacells statistically, following the two-layer observation model for single-cell sequencing data [5]. The first layer, the *expression model*, describes the distribution of $\boldsymbol{\lambda}_i = (\lambda_{i1}, \ldots, \lambda_{ip})^\top$, which captures the biological variation among cells and typically depends on cell i's covariates, such as cell type. The second layer, the *measurement model*, describes the distribution of y_{ij} given $\boldsymbol{\lambda}_i$:

$$
\begin{aligned}
& y_{ij} | \, \boldsymbol{\lambda}_i \overset{\text{ind}}{\sim} \text{Poisson}(c_i \lambda_{ij}) \,, \\
\text{which implies} \quad & (y_{i1}, \ldots, y_{ip}) | \, \boldsymbol{\lambda}_i, y_{i+} \sim \text{Mult}(y_{i+}, \lambda_{i1}, \ldots, \lambda_{ip}) \,,
\end{aligned} \tag{1}
$$

where $c_i = \mathbb{E}[y_{i+}|\boldsymbol{\lambda}_i]$, with $y_{i+} = \sum_{j=1}^{p} y_{ij}$ representing the library size of cell i; the measurement model describes the technical variation among cells. Under this two-layer observation model, we define a metacell as a group of cells sharing the same $\boldsymbol{\lambda}$. We term metacells that satisfy this definition as *trustworthy metacells*, and those that do not as *dubious metacells*.

2.2 The mcRigor Algorithm

Detection of Dubious Metacells. The mcRigor algorithm begins by detecting dubious metacells given a metacell partitioning, via a per-metacell statistical test with $H_0 : (y_{i1}, \ldots, y_{ip}) | \boldsymbol{\lambda}, y_{i+} \overset{\text{ind}}{\sim} \text{Mult}(y_{i+}, \lambda_1, \ldots, \lambda_p)$, for all cell i in a metacell, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^\top$ is shared by all cells within the metacell. Specifically, mcRigor detects dubious metacells via four steps (Fig 1a):

Step 1 (Metacell Divergence Scores). mcRigor computes a *metacell divergence score* (mcDiv) for each of M metacells separately. For the kth metacell of size m_k (i.e., containing m_k single cells), $k = 1, \ldots, M$, mcRigor calculates the feature correlation matrix $\boldsymbol{\Sigma}_k$ and its deviation from the identity matrix $\mathbf{I}$ using the Frobenius norm $\|\boldsymbol{\Sigma}_k - \mathbf{I}\|_{\text{F}}$. To establish the baseline deviation under no feature correlation, mcRigor applies a within-feature permutation, i.e., independently shuffling the values of m_k single cells for each feature, and then calculates the feature correlation matrix $\widetilde{\boldsymbol{\Sigma}}_k$ from the permuted data. Then, mcDiv is defined as: $\text{mcDiv}_k = \|\boldsymbol{\Sigma}_k - \mathbf{I}\|_{\text{F}} / \|\widetilde{\boldsymbol{\Sigma}}_k - \mathbf{I}\|_{\text{F}}$.

Step 2 (Null Divergence Scores). mcRigor constructs a *null divergence score* ($\text{mcDiv}^{\text{null}}$) for each metacell in a data-driven manner. For the kth metacell, mcRigor generates a within-cell permuted data matrix by independently shuffling the values of p features for each of the m_k single cells, retaining the original cells' library sizes, and calculates its feature correlation matrix, $\boldsymbol{\Pi}_k$. The same procedure as in Step 1 is then applied to calculate $\widetilde{\boldsymbol{\Pi}}_k$, the feature correlation matrix of the double-permuted data matrix (first within-cell, then within-feature permutation). Then, $\text{mcDiv}^{\text{null}}$ is defined as $\text{mcDiv}_k^{\text{null}} = \|\boldsymbol{\Pi}_k - \mathbf{I}\|_{\text{F}} / \|\widetilde{\boldsymbol{\Pi}}_k - \mathbf{I}\|_{\text{F}}$.

Step 3 (Divergence Score Thresholds). From the M null divergence scores, $\text{mcDiv}_1^{\text{null}}, \ldots, \text{mcDiv}_M^{\text{null}}$, mcRigor learns the mcDiv thresholds for distinguishing between dubious and trustworthy metacells. Specifically, the threshold is defined as a function of the metacell size:

$$\theta(m_k) = \text{q}_{0.95}\left(\left\{\text{mcDiv}_{k'}^{\text{null}} : m_{k'} \in [m_k - h, m_k + h],\, k' = 1, \ldots, M\right\}\right), \quad (2)$$

where $\text{q}_{0.95}(\cdot)$ computes the 95% quantile, and h is the metacell size bandwidth.

Step 4 (Dubious Metacell Detection). Upon completion of **Step 1–3**, mcRigor categorizes each of the M metacells as dubious or trustworthy, classifying the kth metacell as dubious if $\text{mcDiv}_k > \theta(m_k)$, otherwise deeming it trustworthy.

Optimization of Metacell Partitioning: Method and Hyperparameter Choices. Employing the above dubious metacell detection procedure, mcRigor further optimizes metacell partitioning by identifying the best method and hyperparameter for a given single-cell dataset. This work focuses on optimizing the *granularity level* hyperparameter, γ, which represents the average number of cells per metacell, balancing sparsity reduction and signal preservation. To quantify this trade-off, mcRigor evaluates each method-hyperparameter configuration using two metrics: *DubRate*, which captures signal distortion as the proportion of cells in dubious metacells, and *ZeroRate*, which measures the remaining sparsity as the proportion of zeros in the metacell expression matrix. mcRigor then defines an *evaluation score*: $Score = 1 - w \times DubRate - (1-w) \times ZeroRate \in [0, 1]$, where the weight $w \in (0, 1)$ has a default value of 0.5. From candidate method-hyperparameter configurations, mcRigor selects the one that maximizes this score (Fig 1b).

3 Results and Conclusion

We evaluated mcRigor on a semi-synthetic dataset with 50 ground-truth metacells with true granularity level $\gamma^* = 50$. MetaCell, SEACells, and SuperCell were each applied at $\gamma = 2, \ldots, 100$, obtaining metacell partitions for each method-hyperparameter configuration. Metacell *purity* was defined as the highest fraction of cells from the same ground-truth metacell, with purity $= 1$ indicating a truly trustworthy metacell and < 1 indicating a dubious metacell. Using ground-truth purity, we applied mcRigor to test its ability to detect dubious metacells. Notably, for MetaCell partitions, mcRigor obtained per-metacell mcDiv scores strongly correlated with metacell purity (Pearson correlation $\rho = 0.948$, Fig 1c) and its thresholding via double permutation accurately distinguished dubious from trustworthy metacells, achieving an F score of 0.921 (Fig 1c). To evaluate mcRigor's ability to optimize metacell partitioning, we computed the *Score* metric for each partition. For MetaCell partitions at varying γ, the highest *Score* was achieved precisely at $\gamma = \gamma^*$, pinpointing a partition closely matching the ground truth with only four dubious metacells (Fig 1b). For SEACells and SuperCell partitions, mcRigor showed similar effectiveness.

To conclude, mcRigor enhances the rigor of metacell partitioning in single-cell data analysis, ensuring reliable downstream analysis on metacells. The R package `mcRigor` is available at https://github.com/JSB-UCLA/mcRigor.

Full Paper: A preprint of the full paper is available on bioRxiv at https://www.biorxiv.org/content/10.1101/2024.10.30.621093v1, where we provide detailed justification of the mcRigor method and multiple real-data analyses showcasing mcRigor's effectiveness in enhancing metacell-based data analysis.

References

1. Baran, Y., et al.: Metacell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biol. **20**, 1–19 (2019)
2. Ben-Kiki, O., Bercovich, A., Lifshitz, A., Tanay, A.: Metacell-2: a divide-and-conquer metacell algorithm for scalable scRNA-seq analysis. Genome Biol. **23**(1), 100 (2022)
3. Bilous, M., et al.: Metacells untangle large and complex single-cell transcriptome networks. BMC Bioinformatics **23**(1), 336 (2022)
4. Persad, S., et al.: Seacells infers transcriptional and epigenomic cellular states from single-cell genomics data. Nat. Biotechnol. **41**(12), 1746–1757 (2023)
5. Sarkar, A., Stephens, M.: Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. Nat. Genet. **53**(6), 770–777 (2021)