JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 00, Number 00, 2025

Pp. 000–000
DOI: 10.1177/15578666251383561

mcRigor: A Statistical Software Package for Evaluating and Optimizing Metacell Partitioning in Single-Cell Data Analysis

PAN LIU and JINGYI JESSICA LI*

ABSTRACT

Metacell partitioning is a common preprocessing step in single-cell data analysis, used to reduce sparsity by aggregating similar cells. However, existing metacell partitioning algorithms may inadvertently group heterogeneous cells, potentially biasing downstream analyses. The resulting metacell partitions can vary substantially with different hyperparameter settings, leaving users uncertain about which result to trust. The `mcRigor` R package offers a statistical method for evaluating and optimizing metacell partitioning in single-cell data analysis. This article provides instructions for installing and using `mcRigor` to support more rigorous and interpretable metacell-based workflows.

Keywords: metacell partitioning, single-cell sequencing, hyperparameter optimization.

1. BACKGROUND

Single-cell sequencing data are inherently sparse due to limited capture efficiency and sequencing depth, motivating the use of metacells—aggregated representations of similar cells—to enhance biological signals (Bilous et al., 2024). Popular metacell partitioning algorithms—such as MetaCell (Baran et al., 2019), MetaCell2 (Ben-Kiki et al., 2022), SuperCell (Bilous et al., 2022), and SEACells (Persad et al., 2023)—use graph- or kernel-based strategies to group cells. However, these algorithms do not assess whether each resulting metacell is truly homogeneous, and their outputs can vary substantially with different hyperparameter choices, leaving users uncertain about reliability.

Building on Liu and Li (2025, 2025), the key innovation of `mcRigor` is a statistical framework for evaluating the reliability of metacell partitions. Central to this framework is the *metacell divergence score* (`mcDiv`), a feature-correlation-based statistic that summarizes how strongly gene-gene signals co-vary across the cells within a metacell. If a metacell is truly homogeneous (cells share the same underlying biological state), then the variability across cells is dominated by near-independent sampling noise, so cross-cell gene-gene correlations should be negligible. By contrast, a heterogeneous metacell contains cell subpopulations shaped by biological factors (e.g., cell-cycle state) that induce coordinated shifts in sets of genes, yielding systematic, nonzero gene-gene correlations. `mcDiv` aggregates departures from near-independence in gene-gene correlations across cells within a metacell. To calibrate the score, `mcRigor` employs a double-permutation reference that preserves key one-dimensional margins (e.g., per-gene levels and per-cell library sizes) while disrupting

Department of Statistics and Data Science, University of California, Los Angeles, California, USA.
*Present Address: Biostatistics Program, Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA 98109, USA.

gene-gene dependence, yielding a data-driven threshold that adapts to the metacell size (i.e., the number of constituent single cells) for flagging *dubious* versus *trustworthy* metacells. This principled test fills a critical gap, as existing metacell partitioning algorithms construct metacells but do not provide statistical evaluation of their validity.

A second contribution of `mcRigor` is its strategy for optimizing metacell partitioning across algorithms and hyperparameter settings. Because many algorithms require users to specify a granularity level (e.g., the ratio of cells to metacells), their results can be sensitive to this choice. `mcRigor` systematically evaluates candidate partitions produced by multiple metacell partitioning algorithms across ranges of granularity levels, balances the trade-off between metacell homogeneity and data sparsity, and selects the optimal algorithm-hyperparameter configuration for a given dataset. Together, these innovations allow `mcRigor` to flag unreliable metacells and guide users toward selecting trustworthy metacell partitions, thereby improving the rigor and reproducibility of metacell-based single-cell analysis.

This article serves as a concise guide to using the `mcRigor` R package. For troubleshooting assistance or feedback, please visit the GitHub page (https://github.com/JSB-UCLA/mcRigor), where additional documentation is available.

2. INSTALLATION

To install the `mcRigor` R package from GitHub, users need the `devtools` R package. Running the following commands in R installs the package:

```
if (!require("devtools", quietly = TRUE))
    install.packages("devtools")
devtools::install_github("JSB-UCLA/mcRigor")
```

After the installation, the package can be imported by running `library(mcRigor)`. Note that the `Seurat` package is a dependency of `mcRigor`, so importing `mcRigor` automatically imports `Seurat` as well.

3. FUNCTIONALITY 1: DETECTING DUBIOUS METACELLS

The first functionality of `mcRigor` is to assess the quality of a given metacell partition by identifying dubious metacells within, which are metacells that contain heterogeneous single cells and may distort downstream analyses. The functionality is realized by function `mcRigor_DETECT()` in the package. The `mcRigor_DETECT()` function requires two main inputs: (1) the raw scRNA-seq data and (2) a given metacell partition generated by either existing metacell partitioning algorithms or ad hoc approaches. We describe the detailed formats of these two inputs below.

The raw scRNA-seq data should be provided as a `Seurat` object and passed to the `obj_singlecell` argument of `mcRigor_DETECT()`. A semi-synthetic scRNA-seq dataset, generated as described in Liu and Li (2025) and saved as an RDS file (`syn.rds`), is included with the `mcRigor` package as an example. We first load this scRNA-seq dataset.

```
sc_dir = system.file('extdata', 'syn.rds', package = 'mcRigor')
obj_singlecell = readRDS(file = sc_dir)
obj_singlecell
#> An object of class Seurat
#> 2000 features across 13400 samples within 1 assay
#> Active assay: RNA (2000 features, 2000 variable features)
#> 3 layers present: counts, data, scale.data
#> 2 dimensional reductions calculated: pca, umap
```

The metacell partition should be provided as a data frame, showing the assignment of single cells to metacells, and passed to the `cell_membership` argument of `mcRigor_DETECT()`. Specifically, the input data frame should contain a single column, with each row representing one single cell. As an example, the `mcRigor` package includes metacell partitions for the semi-synthetic scRNA-seq dataset generated by SEACells (Persad et al., 2023), stored in the file `seacells_cell_membership_rna_syn.csv`. This CSV file provides a series of metacell partitions, each corresponding to a different granularity level. The granularity level, denoted by γ, is a key hyperparameter in metacell construction and is defined as the ratio of the number

of single cells to the number of metacells. In this section, we use the metacell partition corresponding to $\gamma = 50$ and load the associated data frame.

```
membership_dir = system.file('extdata', 'seacells_cell_membership_rna_syn.csv',
                             package = 'mcRigor')
cell_membership_all = read.csv(file = membership_dir, check.names = F,
                               row.names = 1)
cell_membership = cell_membership_all['50']
head(cell_membership, 2)
#>                                50
#> 1_Cell1 mc50-allcells-SEACell-98
#> 2_Cell1 mc50-allcells-SEACell-98
```

We then call the `mcRigor_DETECT()` function to assess whether each metacell is *dubious* or *trustworthy*. The `feature_use` argument controls the number of highly variable genes used in the analysis (default: 2000). The `test_cutoff` parameter specifies the significance level for classifying a metacell as dubious, based on the computed metacell divergence score (`mcDiv`), which quantifies the degree of heterogeneity within a metacell. The `aggregate_method` argument defines how single-cell counts are aggregated within each metacell to produce a representative profile—typically through direct averaging (`aggregate_method = "mean"`) or *log1p*-averaging followed by *expm1* transformation (`aggregate_method = "geom"`).

```
detect_res <- mcRigor_DETECT(obj_singlecell = obj_singlecell,
                             cell_membership = cell_membership,
                             feature_use = 2000,
                             test_cutoff = 0.05,
                             aggregate_method = 'mean')
```

The `Seurat` object of metacells is stored in the `obj_metacell` field of the output `detect_res`. The `mcRigor` detection results are recorded both in the `mc_res` field of `detect_res` and in the metadata of the `Seurat` object under the variable name `mcRigor`.

```
table(detect_res$mc_res)
#>     dubious trustworthy
#>          57         211
head(detect_res$obj_metacell$mcRigor, 2)
#>   mc50-allcells-SEACell-0   mc50-allcells-SEACell-1
#>             "trustworthy"                 "dubious"
```

4. FUNCTIONALITY 2: OPTIMIZING METACELL PARTITIONING

The second functionality of `mcRigor` is to evaluate a series of metacell partitions generated at varying granularity levels and select the optimal level for a given single-cell dataset and metacell partitioning algorithm. This is particularly useful because existing metacell algorithms (Baran et al., 2019; Ben-Kiki et al., 2022; Bilous et al., 2022; Persad et al., 2023) either rely on fixed default granularity levels or require users to manually specify the granularity level.

To begin, we load the same semi-synthetic scRNA-seq dataset along with a series of candidate metacell partitions —outputs of SEACells at varying granularity level (γ) values—and store them in a data frame, `cell_membership_all`. Each column of this data frame should represent the metacell partition for a specific granularity level, with the column name indicating the corresponding γ value, and each row should represent a single cell.

```
sc_dir = system.file('extdata', 'syn.rds', package = 'mcRigor')
obj_singlecell = readRDS(file = sc_dir)
membership_dir = system.file('extdata', 'seacells_cell_membership_rna_syn.csv',
                             package = 'mcRigor')
```

```
cell_membership_all = read.csv(file = membership_dir, check.names = F,
                               row.names = 1)
cell_membership[1:2, 1:2]
#>                               100                       99
#> 1_Cell1 mc100-allcells-SEACell-86 mc99-allcells-SEACell-35
#> 2_Cell1 mc100-allcells-SEACell-26 mc99-allcells-SEACell-35
```

We call the `mcRigor_OPTIMIZE()` function to evaluate each metacell partition stored in `cell_membership_all` and select the optimal partition from among them.

```
optimize_res = mcRigor_OPTIMIZE(obj_singlecell = obj_singlecell,
                                cell_membership = cell_membership_all)
```

The output `optimize_res` contains the optimal granularity level (`best_granularity_level`), the evaluation score of the corresponding optimal metacell partition (`best_score`), and the `Seurat` object of metacells contructed at the optimal granularity level (`opt_metacell`).

```
opt_metacell = optimize_res$opt_metacell
opt_metacell
#> An object of class Seurat
#> 2000 features across 319 samples within 1 assay
#> Active assay: RNA (2000 features, 0 variable features)
#> 2 layers present: counts, data
# Note: "0 variable features" simply means FindVariableFeatures() has not been run on
  opt_metacell yet
```

The evaluation scores for all the provided metacell partitions are also stored in the `score` field of the output

```
optimize_res.
head(optimize_res$scores, 2)
#>    gamma    DubRate    ZeroRate     Score
#>  1     2 0.01106557  0.9201214  0.5344065
#>  2     3 0.01193333  0.9021871  0.5429398
```

Note that the optimal metacell partition may still contain some dubious metacells. The results of `mcRigor`'s dubious metacell detection are stored in the metadata of `opt_metacell` under the name `mcRigor`. Users may choose to exclude these dubious metacells from the optimal partition to avoid any potential bias if the resulting information loss is acceptable.

```
opt_metacell_tuned = subset(opt_metacell, mcRigor == 'trustworthy')
```

The optimized metacells, either `opt_metacell_tuned` or `opt_metacell`, can then be directly used for any downstream analysis, including cell clustering, differential expression analysis, pseudotime inference, and more, in the same manner as analyses performed on the original single-cell data.

5. SOFTWARE AVAILABILITY

The `mcRigor` R package is under the MIT Liscence and available at https://github.com/JSB-UCLA/mcRigor.

ACKNOWLEDGMENTS

The authors thank the members of the Junction of Statistics and Biology lab at UCLA (https://jsb-lab.org/) for their helpful feedback and comments. The full article is available as a preprint on *bioRxiv* (Liu and Li, 2025), and a peer-reviewed, abridged version appeared in the proceedings of the RECOMB conference (Liu and Li, 2025).

AUTHORS' CONTRIBUTIONS

All authors have read and approved the final article. Author contributions (CRediT): Conceptualization: P.L. and J.J.L.; Methodology: P.L. and J.J.L.; Software: P.L.; Formal analysis: P.L.; Investigation: P.L.; Visualization: P.L.; Writing—original draft: P.L.; Writing—review and editing: P.L. and J.J.L.; Supervision: J.J.L.; Project administration: J.J.L.; Funding acquisition: J.J.L.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

FUNDING INFORMATION

This work was supported by the National Science Foundation (DBI-1846216 and DMS-2113754), the National Institutes of Health/NIGMS (R01GM120507 and R35GM140888), and the Chan Zuckerberg Initiative Single-Cell Biology Data Insights Grant (to J.J.L.). This work was supported by the following grants: National Science Foundation DBI-1846216 and DMS2113754, NIH/NIGMS R35GM140888, and Silicon Valley Community Foundation 2022-249355 (Chan-Zuckerberg Initiative Single-Cell Biology Data Insights Grant) (to J.J.L.). Additional support was provided by the National Human Genome Research Institute (NHGRI) through an Opportunity Fund subaward from the Technology Development Coordinating Center (TDCC) U24HG011735. This work was also supported by the Institute for Quantitative and Computational Biosciences (QCBio) at University of California, Los Angeles, through the 2025 QCBio Award (to P.L.).

REFERENCES

Baran Y, Bercovich A, Sebe-Pedros A, et al. Metacell: Analysis of single-cell rna-seq data using k-nn graph partitions. Genome Biol 2019;20(1):206.

Ben-Kiki O, Bercovich A, Lifshitz A, et al. Metacell-2: A divide-and-conquer metacell algorithm for scalable scrna-seq analysis. Genome Biol 2022;23(1):100.

Bilous M, Tran L, Cianciaruso C, et al. Metacells untangle large and complex single-cell transcriptome networks. BMC Bioinformatics 2022;23(1):336.

Bilous M, Hérault L, Gabriel AA, et al. Building and analyzing metacells in single-cell genomics data. Mol Syst Biol 2024;20(7):744–766.

Liu P, Li JJ. mcRigor: A statistical method to enhance the rigor of metacell partitioning in single-cell data analysis. Nat Comms 2025;16:8602.

Liu P, Li JJ. mcRigor: A statistical method to enhance the rigor of metacell partitioning in single-cell rna-seq and atac-seq data analysis. In: Research in Computational Molecular Biology. (Sankararaman S, ed.) Springer Nature: Cham, Switzerland, 2025, pp. 381–385. ISBN 978-3-031-90252-9; doi: 10.1007/978-3-031-90252-9_44

Persad S, Choo Z-N, Dien C, et al. Seacells infers transcriptional and epigenomic cellular states from single-cell genomics data. Nat Biotechnol 2023;41(12):1746–1757.

Address correspondence to:
Dr. Jingyi Jessica Li
Department of Statistics and Data Science
University of California
Los Angeles, CA 90095-1554
USA

E-mail: lijy03@fredhutch.org