mcRigor: a statistical method to enhance the rigor of metacell partitioning in single-cell data analysis

Pan Liu[1] and Jingyi Jessica Li[1,2,3,4,5]

[1] Department of Statistics and Data Science, University of California, Los Angeles, CA 90095-1554
[2] Interdepartmental Program of Bioinformatics, University of California, Los Angeles, CA 90095-7246
[3] Department of Human Genetics, University of California, Los Angeles, CA 90095-7088
[4] Department of Computational Medicine, University of California, Los Angeles, CA 90095-1766
[5] Department of Biostatistics, University of California, Los Angeles, CA 90095-1772

Abstract. In single-cell 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, which are composed of heterogeneous single cells, and optimize the hyperparameter of a metacell partitioning method. The core of mcRigor is a feature-correlation-based statistic that measures the heterogeneity of a metacell, with its null distribution derived from a double permutation mechanism. As an optimizer for existing metacell partitioning methods, mcRigor has been shown to improve the reliability of discoveries in single-cell RNA-seq and multiome (RNA+ATAC) data analyses, such as uncovering differential gene co-expression modules, enhancer-gene associations, and gene temporal expression. Moreover, mcRigor enables benchmarking and selection of the most suitable metacell partitioning method with optimized hyperparameters tailored to specific datasets, ensuring reliable downstream analysis. Our results showed that MetaCell and SEACells consistently outperformed MetaCell2 and SuperCell, though their superior performance comes at the expense of longer runtimes.

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

1 Introduction

Single-cell sequencing technologies have catalyzed a paradigm shift in genomics by uncovering cellular heterogeneity with unprecedented resolution across multiple modalities, including transcriptomics via single-cell RNA sequencing (scRNA-seq) [48, 42, 61], epigenomics through single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) [8, 17], and multiome assays that simultaneously measure RNA-seq and ATAC-seq [11, 12]. The majority of these technologies are high-throughput and droplet-based, capable of profiling millions of cells, but they are often compromised by high sparsity in the sequencing read counts due to low per-cell sequencing depth and imperfections in the reverse transcription and amplification steps [29].

The high sparsity presents a substantial challenge for data analysis [31], with common strategies to mitigate it including *imputation* and *metacell aggregation*. Imputation addresses sparsity by predicting missing feature expression values, where *features* represent genes or chromatin regions, using cells and/or features with similar expression profiles. Numerous imputation methods have been developed for single-cell data, including scImpute [35], SAVER [27], MAGIC [56], and DCA [19], as well as deep generative models [38, 2]. Imputation methods have the advantage of retaining the full set of single cells; however, they can sometimes induce false positives in the downstream differential gene expression detection [1] and may suffer from oversmoothing [27, 3], which creates artificial proximities between cells. As an alternative to imputation, the metacell approach groups cells representing the same cell state into a metacell and uses the metacell's expression profile, typically obtained by averaging the single-cell expression profiles, for subsequent analysis [3]. The metacell approach is expected to reduce noise and thereby accentuate biological signals that are often obscured in sparse datasets. The metacell concept differs from the *pseudobulk* approach, though both involve cell aggregation. Specifically, a pseudobulk is created by merging all cells within a predefined cell population—typically a cell type—into a single profile, while a metacell aggregates a much smaller, homogeneous group of cells, allowing for multiple metacells within a single cell type. The pseudobulk approach reduces data sparsity and enables the use of computational methods designed for bulk data [16, 60]. However, by merging all cells of a cell type into one pseudobulk, this approach removes all within-cell-type variation. In contrast, metacells aim to preserve this variation, maintaining the resolution advantage of single-cell data that allows cell-type-specific analysis.

Despite the increasing use of the metacell concept in high-profile single-cell studies—such as investigating cell differentiation states [23], characterizing tissue compartments and diverse cell populations [50, 7, 34], patient stratification for individualized immunotherapy design [47], and temporal analysis of cell transcriptomes [36, 30]—there is still no rigorously defined concept of a metacell or a universally accepted strategy for constructing metacells. This lack of consensus can lead to inconsistencies in studies employing the metacell concept, potentially resulting in arbitrariness and suboptimal analysis outcomes.

In addition to the various approaches used for partitioning single cells into metacells in in-house data analyses, several general methods have been developed for this purpose. The most popular ones include MetaCell [3], MetaCell2 [4], SuperCell [6], and SEACells [46]. MetaCell employs a k-nearest neighbor graph of cells, uses graph resampling and clustering to update the graph, and finally identifies metacells as small clusters. It also includes additional steps to detect and exclude outlier cells that are not incorporated into any metacells. Developed by the same authors as MetaCell, MetaCell2 is designed for faster performance through divide-and-conquer. SuperCell applies a walktrap clustering method to a PCA-derived kNN graph of cells. SEACells uses a kernel to define a cell-cell similarity matrix, treating these similarities as cell embeddings for archetypal analysis, with the resulting archetypes used to identify metacells. However, these methods can produce different metacell partitions, which are also influenced by the hyperparameters they employ [5]. This lack of consensus leaves users uncertain about which metacell partition to use and to what extent the resulting metacell profiles preserve biological signals. Therefore, a formal definition and an evaluation standard for metacells are needed to guarantee principled metacell aggregation and ensure unbiasedness in downstream analytical processes.

To fill this gap, we propose a statistical definition of metacell and accordingly develop mcRigor, a novel statistical method to enhance the rigor of metacell partitioning in single-cell data analysis. Theoretically, a metacell is defined as a homogeneous group of single-cell profiles that could be viewed as resamples from the same original cell, with any variation within a metacell attributed solely to technical measurement errors rather than biological differences. Built upon this definition, mcRigor can identify *dubious metacells*

that are heterogeneous and violate this definition, while also optimizing the metacell partitioning strategy to ensure reliable metacell construction. Our results demonstrate that mcRigor successfully identifies and removes dubious metacells, revealing the COVID-related co-expression of adaptive immune response genes, which is enriched in COVID-19 patients compared to healthy individuals (Fig 1d). We also demonstrate that mcRigor enhances gene regulatory analysis by revealing enhancer-gene associations that are obscured in single-cell multiome data or by the presence of dubious metacells, while excluding spurious associations biased by these metacells (Fig 1e). Moreover, mcRigor balances the trade-off between data sparsity and signal distortion, identifying optimal metacell partitions for distinguishing biological from non-biological zeros, detecting differentially expressed genes, and revealing temporal trajectories of cellular immune responses (Fig 3).

2 Methods

2.1 A statistical definition of metacells

As first introduced by Baran et al. (2019) [3], a *metacell* is a homogeneous group of single-cell profiles that could be considered resamples from the same original cell. This means that within a metacell, the single cells should share the same biological state and exhibit the same expected feature expression levels (e.g., genes or chromatin regions), referred to as *biological signals*. Any differences in their observed counts (of sequencing reads or unique molecular identifiers) should therefore be attributed to measurement imperfections, known as *technical variations*. This definition ensures that averaging the profiles within a metacell reduces technical variation while preserving biological signals.

We formalize this metacell definition statistically by following the *observation model* for single-cell sequencing data described in Sarkar et al. (2021) [49]. In this model, observed count variations can be decomposed into two components: the *biological variation*, which reflects differences in biological signals across cells, and the *technical variation*, introduced during the measurement process that obscure these biological signals. Hence, the model is hierarchical, consisting of an *expression model*, describing biological signal distribution (i.e., the biological variation), and a *measurement model*, describing observed count distribution given these signals (i.e., the technical variation).

Mathematically, we consider a total of n cells sequenced to measure the abundance of p features, such as genes or chromatin regions. Regarding the biological signals, we let u_{ij} denote the number of molecules presented in cell i from feature j, $u_{i+} = \sum_{j=1}^{p} u_{ij}$ denote the total number of molecules in cell i, and $\lambda_{ij} = u_{ij}/u_{i+}$ denote the relative abundance of feature j in cell i. It is commonly assumed that $\boldsymbol{\Lambda} = [\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n]^\top = [\lambda_{ij}] \in \mathbb{R}^{n\times p}$ encapsulates all the biological signals that can be estimated from sequencing data, as absolute abundance information is not preserved in sequencing. Regarding the observed counts, we write $\mathbf{Y} = [y_{ij}] \in \mathbb{Z}_{\geq 0}^{n\times p}$, where y_{ij} denotes the observed count in cell i of feature j, and y_{i+} denotes the total observed count (often referred to as the *cell library size*) in cell i. First, the expression model describes the distribution of $\boldsymbol{\lambda}_i$, which typically depends on cell i's covariates, such as the cell type. Second, given $\boldsymbol{\lambda}_i$, the measurement model describes the distribution of y_{ij}, in a form such as

$$y_{ij} | \boldsymbol{\lambda}_i \overset{\text{ind}}{\sim} \text{Poisson}(c_i \lambda_{ij}), \tag{1}$$

where $c_i = \mathbb{E}[y_{i+}|\boldsymbol{\lambda}_i]$, reflecting the technical variation among cells.

Under this hierarchical observation model and following the definition by Baran et al. (2019) [3], we introduce a **statistical definition**: *a metacell is a group of cells that share the same* $\boldsymbol{\lambda}$. We show in the supplementary file that, under this definition, aggregating cells into a metacell through averaging can reduce the technical variance in (1) without introducing bias. In contrast, if cells with different $\boldsymbol{\lambda}$ are aggregated, the estimation of $\boldsymbol{\lambda}$ will most likely be biased, leading to distorted inference of the biological signals. Therefore, rigorous identification of metacells is crucial to guarantee the reliability of downstream analysis. Based on our statistical definition, we term metacells that satisfy this definition as **trustworthy metacells**, and those that do not as **dubious metacells**. In the next section, we will introduce an algorithm called **mcRigor**, which (1) detects dubious metacells from those constructed by any metacell algorithm, and (2) optimizes both the hyperparameters of any algorithm and the algorithm choice among multiple algorithms for a given single-cell dataset.

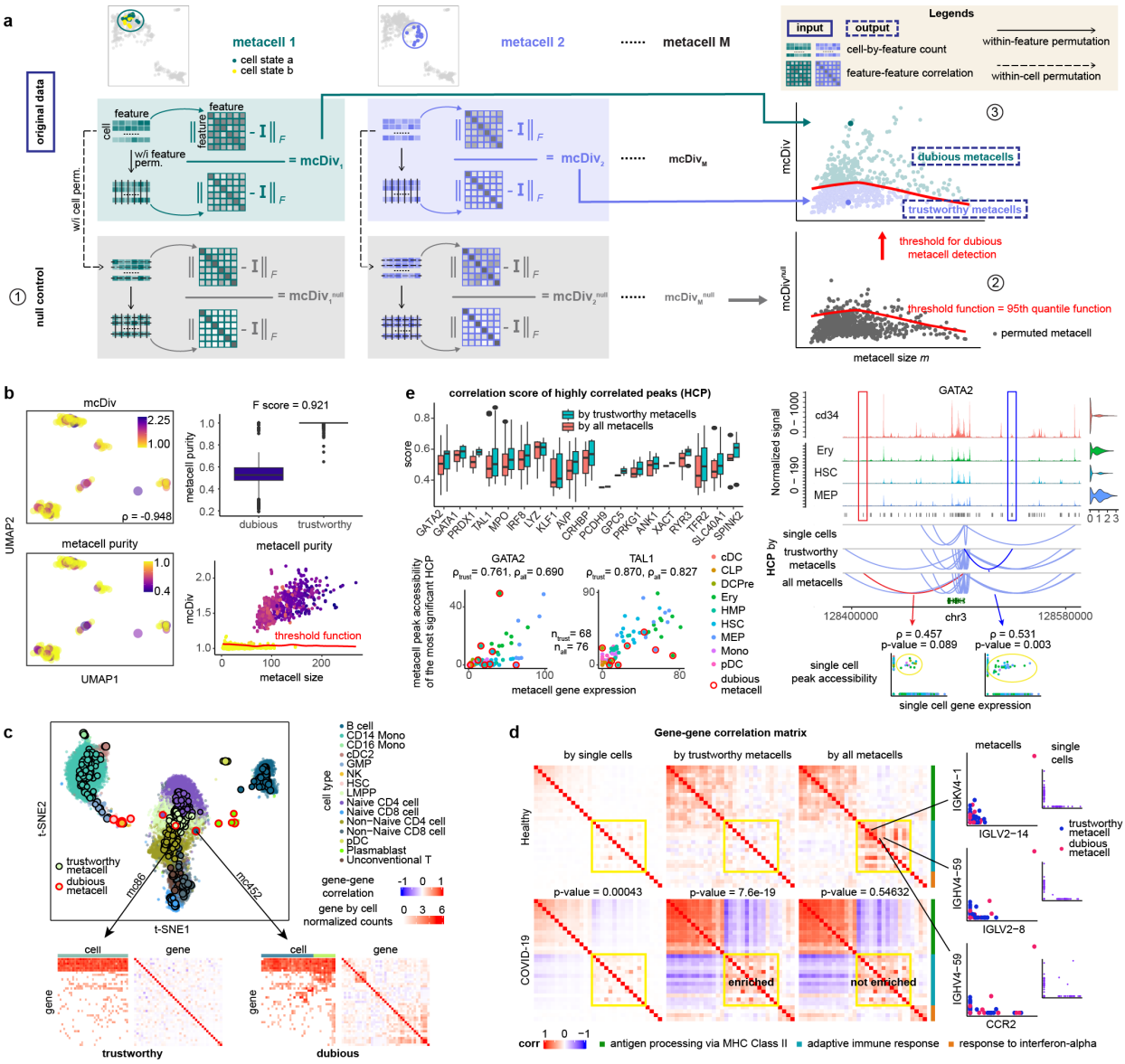

Fig. 1. mcRigor detects dubious metacells and rectifies downstream analysis for both scRNA and scATAC data. a, Schematics of the mcRigor method pipeline for dubious metacell detection. **b,** mcRigor performs well on the MetaCell partitions for the semi-synthetic data. Left: UMAP plots of the metacells colored by mcDiv values versus metacell purity (ground truth). Right: mcRigor distinguishes between dubious and trustworthy metacells with high accuracy. **c,** Dubious metacells identified by mcRigor have spurious embeddings and internal heterogeneity, while trustworthy metacells identified are internally homogeneous. Bottom: Gene by cell count and gene by gene correlation heatmaps. **d,** mcRigor reveals a gene module (highlighted by yellow squares) with enriched co-expression in COVID-19 samples compared to healthy controls for the SuperCell partition at $\gamma = 20$. Left: Gene-gene correlation matrices for three important gene modules under COVID-19 and healthy control. The p-values are obtained through Wilcoxon rank sum test. Right: Scatter plots of three gene pairs showing correlation artifacts caused by dubious metacells. **e,** mcRigor empowers gene regulatory inference (left) and produces reliable discoveries (right). Left: Box plots and and scatter plots showing that removing dubious metacells yield higher correlation between genes and their HCP's. Right: Removing dubious metacells reveals an additional HCP for *GATA2* while filtering out a weakly supported one

2.2 The mcRigor algorithm

Detection of dubious metacells The mcRigor algorithm begins by distinguishing between trustworthy and dubious metacells given a metacell partitioning. The premise is that, assuming cells share the same library

size, any pair of features should display a negligible correlation among the cells within a trustworthy metacell. This lack of correlation follows our definition of a trustworthy metacell: given the same $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^\top$, the variation among cells is purely technical, as represented by the measurement model (1). In the supplementary file, we prove that under the measurement model, features are nearly pairwise uncorrelated, with the minimal correlation decreasing as p increases. Guided by this premise, mcRigor detects dubious metacells by examining the $p \times p$ feature correlation matrix of each metacell, via a statistical test with the null hypothesis:

$$H_0: \; y_{ij} | \, \boldsymbol{\lambda} \overset{\text{ind}}{\sim} \text{Poisson}(c_i \lambda_j) \, , \; \forall \text{ cell } i \in \text{metacell} \, , \; j = 1, \ldots, p \, , \tag{2}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^\top$ denotes the corresponding metacell feature expression level that is shared by all cells within the metacell. It is important to note that this null hypothesis is equivalent to

$$H_0: \; (y_{i1}, \ldots, y_{ip}) | \, \boldsymbol{\lambda}, y_{i+} \overset{\text{ind}}{\sim} \text{Mult}(y_{i+}, \lambda_1, \ldots, \lambda_p) \, , \; \forall \text{ cell } i \in \text{metacell} \, , \tag{3}$$

which specifies the measurement model for p features jointly as a multinomial distribution with the cell library size as the total and $\boldsymbol{\lambda}$ as the p-dimensional probabilities.

Specifically, given an $n \times p$ cell-by-feature data matrix (after `Seurat` V5 default normalization—the "LogNormalize" approach—and `Seurat` V5 default feature selection that selects the top 2000 highly variable features) and a metacell partition that defines M metacells, mcRigor detects dubious metacells via the following four steps (Fig 1a):

Step 1 (metacell divergence scores). mcRigor computes a divergence score for each metacell separately; the divergence score is essentially a test statistic based on the null hypothesis. For the kth metacell, $k = 1, \ldots, M$, we define its *metacell size* as m_k, the number of single cells it includes. Then mcRigor uses the original $m_k \times p$ data matrix to calculate the metacell feature correlation matrix using Pearson correlation, denoted by $\boldsymbol{\Sigma}_k \in [-1, 1]^{p \times p}$. The deviation of $\boldsymbol{\Sigma}_k$ from the identity matrix $\mathbf{I} \in \mathbb{R}^{p \times p}$ is then calculated as the Frobenius norm $\|\boldsymbol{\Sigma}_k - \mathbf{I}\|_F$. To obtain the baseline deviation value under no feature correlation, mcRigor permutes the $m_k \times p$ data matrix by within-feature (column-wise) permutation, i.e., independently shuffling the values of m_k single cells for each feature. This results in a within-feature permuted data matrix from which mcRigor calculates the corresponding feature correlation matrix $\widetilde{\boldsymbol{\Sigma}}_k$. The *metacell divergence score*, called "mcDiv," is then defined as

$$\text{mcDiv}_k = \frac{\|\boldsymbol{\Sigma}_k - \mathbf{I}\|_F}{\|\widetilde{\boldsymbol{\Sigma}}_k - \mathbf{I}\|_F}, \tag{4}$$

where dividing $\|\boldsymbol{\Sigma}_k - \mathbf{I}\|_F$ by $\|\widetilde{\boldsymbol{\Sigma}}_k - \mathbf{I}\|_F$ can be viewed as normalization.

Step 2 (null divergence scores). mcRigor constructs a null divergence score for each metacell in a data-driven manner. These null scores serve as negative controls for thresholding the divergence scores of metacells, determining the trustworthiness of metacells. Specifically, for the kth metacell, mcRigor applies the same procedure as in Step 1 to a within-cell permuted data matrix, obtained by independently shuffling the values of p features for each of the m_k single cells. Through within-cell (row-wise) permutation, the resulting permuted cells retain the original cells' library sizes. Let $\boldsymbol{\Pi}_k$ denote the correlation matrix of this within-cell permuted data matrix. Following the same procedure as in Step 1, a within-feature permutation is applied to the within-cell permutated data matrix, resulting in $\widetilde{\boldsymbol{\Pi}}_k$, the correlation matrix of the double-permuted data matrix (first within-cell, then within-feature permutation). The *null divergence score*, called "$\text{mcDiv}^{\text{null}}$," derived from the kth metacell is defined as

$$\text{mcDiv}_k^{\text{null}} = \frac{\|\boldsymbol{\Pi}_k - \mathbf{I}\|_F}{\|\widetilde{\boldsymbol{\Pi}}_k - \mathbf{I}\|_F}. \tag{5}$$

The double permutation (first within-cell, then within-feature permutation) is necessary for the null score calculation, as the null hypothesis (3) is conditional on the cell library sizes, and within-cell permutation preserves these library sizes. Additionally, we cannot simply compare the numerators $\|\boldsymbol{\Sigma}_k - \mathbf{I}\|_F$ and $\|\boldsymbol{\Pi}_k - \mathbf{I}\|_F$ while ignoring the denominator in the metacell divergence score definition. This is because $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Pi}_k$ are based on different sets of features (since within-cell permutation does not preserve features) and

are therefore not directly comparable. Furthermore, the denominator $\|\widetilde{\boldsymbol{\Sigma}}_k - \mathbf{I}\|_F$, which is derived from within-feature permutation, is not a valid null for the numerator $\|\boldsymbol{\Sigma}_k - \mathbf{I}\|_F$, as the cell library sizes are not preserved in the calculation of $\widetilde{\boldsymbol{\Sigma}}_k$. We will present results to show that within-feature permutation is inappropriate in Note 1.

From the M null divergence scores, $\mathrm{mcDiv}_1^{\mathrm{null}}, \ldots, \mathrm{mcDiv}_M^{\mathrm{null}}$, mcRigor learns the thresholds for distinguishing between dubious and trustworthy metacells based on their mcDiv scores in the following step.

Step 3 (divergence score thresholds). An intuitive approach might suggest classifying the metacells whose mcDiv scores exceed the 95% quantile of $\mathrm{mcDiv}_1^{\mathrm{null}}, \ldots, \mathrm{mcDiv}_M^{\mathrm{null}}$ as the dubious metacells and those below the quantile as trustworthy metacells. However, the null divergence scores defined in (5) exhibit varying distributions depending on the metacell size (Fig S2b). Hence, mcRigor uses metacell-size-specific mcDiv thresholds instead of a single 95% quantile value to determine the trustworthiness of a metacell. Specifically, the threshold is metacell-specific and defined as a function of the metacell size:

$$\theta(m_k) = \mathrm{q}_{0.95}\left(\left\{\mathrm{mcDiv}_{k'}^{\mathrm{null}} : m_{k'} \in [m_k - h, m_k + h],\, k' = 1, \ldots, M\right\}\right), \tag{6}$$

where $\mathrm{q}_{0.95}(\cdot)$ computes the 95% quantile and h stands for the metacell size bandwidth (default $h = 10$). That is, mcRigor compute the threshold for the kth metacell as the 95% quantile of the null diveregence scores from metacells whose sizes fall within the bandwidth of m_k.

Step 4 (dubious metacell detection). Upon completion of the previous three steps, mcRigor categorizes all M metacells as either dubious or trustworthy. Specifically, for the kth metacell, if $\mathrm{mcDiv}_k > \theta(m_k)$, it is classified as dubious; otherwise, it is considered trustworthy.

Note 1. Using the within-feature permutation instead of the double permutation to find dubious metacell detection thresholds, i.e., the kth metacell is considered dubious if $\|\boldsymbol{\Sigma}_k - \mathbf{I}\|_F > \mathrm{q}_{0.95}\big(\big\{\|\widetilde{\boldsymbol{\Sigma}}_{k'} - \mathbf{I}\|_F : m_{k'} \in [m_k - h, m_k + h],\, k' = 1, \ldots, M\big\}\big)$, sets the threshold too low, leading to the false flagging of trustworthy metacells as dubious. This occurs because $\|\boldsymbol{\Sigma}_k - \mathbf{I}\|_F$ may simply reflect variations in cell library sizes, which are not preserved in the within-feature permutation, resulting in lower values for $\|\widetilde{\boldsymbol{\Sigma}}_k - \mathbf{I}\|_F$. In a simulation study, we demonstrate that more than 35% of genuinely trustworthy metacells are incorrectly classified as dubious, leading to poor overall classification performance (F-score < 0.4) in distinguishing between trustworthy and dubious metacells (Fig S1). This is in sharp contrast to the performance of mcRigor based on the double permutation—accurately identifying over 98% of genuinely trustworthy metacells as trustworthy, with less than 2% incorrectly classified as dubious (Fig S1). Additionally, mcRigor's F-score for distinguishing between dubious and trustworthy metacells consistently exceeds 0.9, highlighting its high detection accuracy.

Note 2. It is important to note that there is extensive literature on high-dimensional covariance testing [13, 9, 10], which might appear relevant for detecting dubious metacells by testing whether the feature covariance matrix within a metacell significantly deviates from a diagonal matrix. However, those tests are not applicable for two main reasons. First, the metacell size is too small (commonly below 100) for the test statistics to approximately follow their asymptotic null distributions. Second, the null hypotheses in these tests do not account for cell library sizes.

Optimization of metacell partitioning: method and hyperparameter choices Employing the dubious metacell detection procedure outlined above, mcRigor further helps optimize metacell partitioning by identifying the best-performing method method and the optimal hyperparameter for a given single-cell dataset. This work focuses on optimizing the *granularity level*, γ, which represents the average number of cells per metacell and is required by all metacell methods (some use *target metacell size*, which can be converted to γ). While other parameters, such as the number of neighbors (k) for kNN graph and the number of principal components used, could also be optimized by mcRigor, we omit their discussion here, as they have been reported to have less impact on partitioning and are not required by all metacell methods [5].

The granularity level hyperparameter controls the extent of aggregation from single cells to metacells [5]. If γ is too large, it combines heterogeneous cells into metacells, distorting downstream analyses. Conversely, a very small γ results in metacells similar to single cells, failing to resolve sparsity. Therefore, optimizing metacell partitioning is, at its core, about balancing between data sparsity and signal distortion, the latter

being addressed by mcRigor's first functionality: detection of dubious metacells. An ideal metacell partitioning should aggregate as many cells as possible, while ensuring that only homogeneous cells are combined. To achieve this, mcRigor quantifies sparsity and distortion for each partition, allowing a quantitative tradeoff.

Specifically, for a given metacell method and hyperparameter, mcRigor evaluates two competing factors that respectively reflect signal distortion and data sparsity. First, mcRigor reports the proportion of single cells that constitute dubious metacells, denoted as the *dubious rate*: $DubRate \in [0, 1]$. Second, to account for the remaining sparsity level, mcRigor considers another metric, $ZeroRate \in [0, 1]$, which is the proportion of zeros in the $M \times p$ metacell expression matrix after single cells are aggregated into M metacells. Combining *DubRate* and *ZeroRate*, mcRigor defines an *evaluation score* for a method-hyperparameter configuration:

$$Score = 1 - w \cdot DubRate - (1 - w) \cdot ZeroRate \in [0, 1]\,, \tag{7}$$

where $w \in (0, 1)$ is the weight assigned to *DubRate*. Our recommendation is the default $w = 0.5$, a robust choice in our simulations, but users can adjust w based on their desired emphasis on the two competing factors. Based on (7), a higher evaluation score indicates a better method-hyperparameter configuration.

The evaluation score allows for universal comparison across varying hyperparameter values and different metacell methods. For a particular dataset, mcRigor identifies the method-hyperparameter configuration that maximizes the evaluation score from a candidate set of metacell methods and hyperparameter values. By default, four metacell methods are considered: SEACells, MetaCell, MetaCell2, and SuperCell, with γ ranging from 2 to 100.

3 Results

3.1 Validation of mcRigor using barcode multiplet data

Our statistical definition of a metacell and the mcRigor method are both built upon the assumption that the observed variation in feature counts among cells consists of two components: biological variation and technical variation. A natural question arises about the validity of this assumption, particularly regarding the presence of technical variation. We addressed this question by investigating technical variation using barcode multiplets identified in droplet-based single-cell assays.

A *barcode multiplet* refers to a set of cell-like observations in which each observation is assigned a unique cell barcode but actually originates from the same physical cell. This occurs when a single cell is tagged with multiple distinct barcodes during the droplet-based sequencing process, leading to multiple barcodes being incorrectly interpreted as separate cells [33]. Note that this term "barcode multiplet" should not be confused with the term "multiplet" as used in the doublet detection literature, where multiplets stand for multiple cells enclosed by a single droplet. A typical type of barcode multiplets, referred to as "barcode multiplets caused by heterogeneous beads" by Lareau et al. (2019) [32], form when multiple beads are encapsulated within a single droplet containing one cell. Consequently, the mRNA fragments captured by these beads are all sampled from the same pool of mRNA fragments from that cell [33]. Therefore, it is reasonable to assume that the variation observed within a barcode multiplet approximates the technical variation present among single cells in the same biological state.

Lareau et al. (2019) [32] designed a computational method, called bead-based ATAC processing (bap), to identify barcode multiples from a public droplet-based scATAC-seq dataset of 5,000 PBMCs [22]. In total, they identified 16 barcode multiples (caused by heterogeneous beads), each consisting of three to six cell-like observations. We found the observations within each barcode multiplet to be dispersed, suggesting the existence of substantial technical variation (Multiplets 1–9 in Fig 2a). Furthermore, we tested mcRigor on this dataset by treating each barcode multiplet as a trustworthy metacell, as the cell-like observations in each barcode multiplet represent the same biological state. Remarkably, mcRigor successfully identified all 16 barcode multiplets as trustworthy metacells.

3.2 Verification of mcRigor's accuracy in detecting dubious metacells

Since barcode multiplets can only approximate trustworthy metacells and not dubious ones, we next assessed mcRigor's accuracy in detecting dubious metacells. We began with a simulation where the trustworthiness

of metacells was known. Using the scDesign3 simulator [51], we generated a semi-synthetic dataset with cells of known biological states. This was done by modeling gene expression variation and separating it into biological and technical variations based on our model assumptions, using a reference dataset of bone marrow mononuclear cells measured by CITE-seq [52] (the `bmcite` dataset from the R package `SeuratData` [26]). Details of the semi-synthetic dataset generation process are provided in the Methods Section. Our semi-synthetic dataset contains five major cell types—B cells, T cells, progenitor cells, natural killer (NK) cells, and monocytes or dendritic cells (Mono/DC)—and closely resembles the reference dataset (Fig 2). By separating biological and technical variations, we were able to generate 50 ground-truth metacells (with the true granularity level $\gamma^* = 50$), each corresponding to a distinct biological state, and then simulate single cells within each metacell to reflect technical variation. Using the ground-truth metacells, we performed two tasks: (1) comparison of metacell partitioning methods by evaluating whether their constructed metacells were trustworthy or dubious, and (2) verification of mcRigor's ability to distinguish between trustworthy and dubious metacells. Both tasks were facilitated by the realistic nature of the semi-synthetic data and the availability of ground truth indicating metacell trustworthiness.

We applied three metacell partitioning methods—MetaCell, SEACells, and SuperCell—each at varying granularity levels to this semi-synthetic dataset, obtaining a metacell partition for each method and granularity level configuration. For evaluation, within a metacell partition, we defined the *purity* of a metacell as the highest fraction of cells originating from the same ground-truth metacell. Based on purity, the ground truth for metacell trustworthiness was defined as follows: metacells with a purity of 1 are considered truly trustworthy, while all others are considered truly dubious. Using this ground truth, we compared the three methods at each granularity level by evaluating the resulting metacell partitions based on two criteria: (1) the proportion of dubious metacells relative to the total number of metacells, and (2) the *dubious rate*, defined as the proportion of single cells assigned to dubious metacells out of all single cells. The results show that both MetaCell and SEACells performed well, yielding reliable metacell partitions. MetaCell demonstrated a lower proportion of dubious metacells and a lower dubious rate compared to SEACells, but this came at the expense of assigning fewer single cells to metacells. Specifically, at the granularity level $\gamma = 60$, which was above the true granularity level $\gamma^* = 50$, MetaCell, SEACells, and SuperCell produced 20.5% 24.6%, and 26.9% dubious metacells, with dubious rates of 0.289, 0.308, and 0.464, respectively. When the granularity level γ was set to 30, below the true granularity level, MetaCell, SEACells, and SuperCell produced 0.2% 0.2%, and 25.3% dubious metacells, with dubious rates of 0.003, 0.005, and 0.380, respectively.

With the ground-truth metacell purity values and trustworthiness, we then applied mcRigor to the metacell partitions produced by MetaCell, SEACells, and SuperCell, to test if mcRigor can accurately detect the dubious metacells from each partition. For metacell partitions generated by different methods, mcRigor computed per-metacell mcDiv scores that are strongly correlated with the true level of metacell purity (Fig 1b, Fig 2d–e). By applying mcDiv thresholds derived from $\text{mcDiv}^{\text{null}}$ values generated through double permutation, mcRigor effectively distinguished dubious metacells (purity < 1) from trustworthy metacells (purity $= 1$) for the MetaCell partitions, achieving an F score (the harmonic mean of precision and recall for classifying between dubious and trustworthy metacells) of 0.921 (Fig 1b). For the metacell partitions generated by SEACells and SuperCell, mcRigor demonstrated similar effectiveness in detecting dubious metacells (Fig 2d–e), highlighting its applicability across various metacell partitioning methods.

Next, we tested mcRigor on the bone marrow mononuclear cell CITE-seq dataset (the `bmcite` dataset from the R package `SeuratData`). As for the semi-synthetic data, we applied MetaCell, SEACells, and SuperCell, each at varying granularity levels, to this CITE-seq datast. To gain more information on the purity of metacells, we considered fine-grained cell annotations, where T cells were further divided into subtypes including CD4 Memory, CD4 Naive, CD8 Effector, CD8 Memory, CD8 Naive, gdT, MAIT, and Treg. B cells were divided into Memory B, Naive B, and Plasmablast. Progenitor cells were divided into GMP, HSC, and LMPP, while Mono/DC were divided into CD14 Mono, CD16 Mono, cDC2, and pDC. Of the 455 metacells generated by MetaCell with granularity level $\gamma = 30$, mcRigor detected 12 dubious metacells. We overlaid the metacells onto the single-cell t-SNE embedding optimized by scDEED [58] (Fig 1c). Interestingly, the dubious metacells either appeared as small clusters or lay at spurious positions (Fig 1c). For example, the centroid of metacell *mc452* falls within the cluster of LMPP and CD4 cells, while it mainly consists of non-naive CD8 cells. This implies metacell *mc452* is composed of single cells distant from each other in the t-SNE embedding, which are highly likely to represent different biological states. The single-cell by gene count matrix of *mc452* confirms this, as the cells involved can be further divided into

at least two clusters based on notable differences in gene expression (bottom right of Fig 1c). In contrast, trustworthy metacells such as *mc86* exhibit a high level of internal homogeneity and near-zero gene-gene correlation (bottom left of Fig 1c). Other dubious metacells are mostly aggregated from Plasmablast and HSC, which are immature cells characterized by frequent self-renewal and multi-lineage differentiation. These results affirm that mcRigor effectively identifies dubious metacells that warrant further investigation.

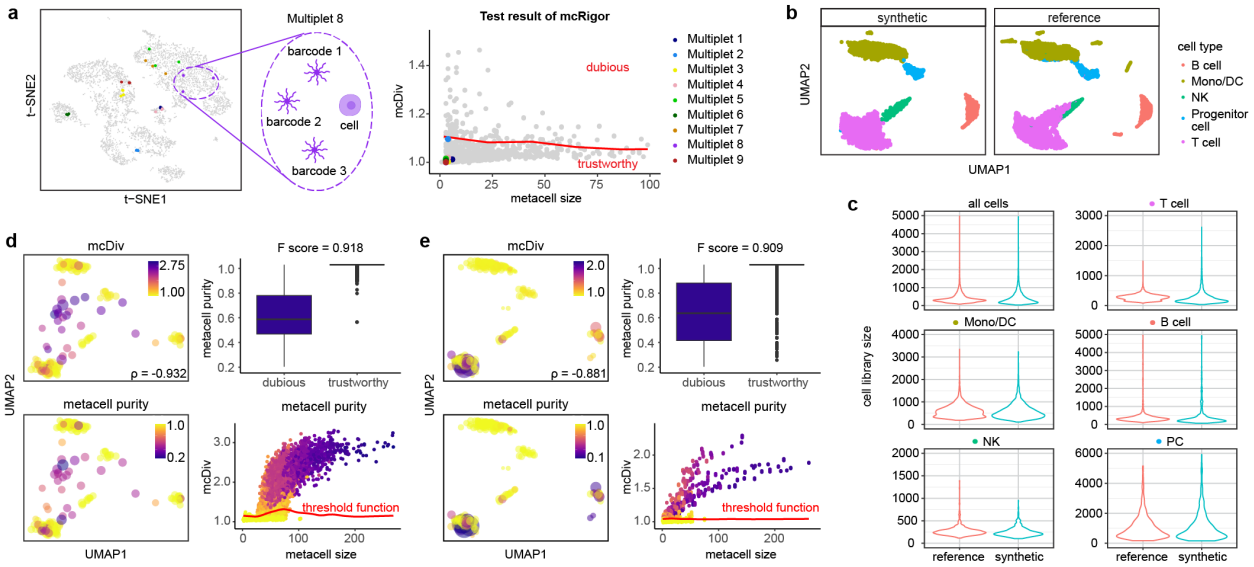

Fig. 2. Justification of the mcRigor method. a, Barcode multiplets justify the statistical definition of metacell (left) and the reliability of mcRigor (right). **b,** UMAP plots of the synthetics data and the reference real data. **c,** Violin plots of cell library sizes for the semi-synthetic data versus the reference data. **d,** mcRigor performs well on the SEACells partitions for the semi-synthetic data. Left: UMAP plots of the metacells colored by mcDiv values versus metacell purity. Top right: mcRigor distinguishes between dubious and trustworthy metacells with high accuracy. Bottom left: The scatter plot of mcDiv versus metacell size with obtained dubious metacell detection threshold. **e,** mcRigor performs well on the SuperCell partitions for the semi-synthetic data.

3.3 mcRigor uncovers differential gene co-expression between healthy and COVID-19 patients

Gene-gene co-expression analysis is widely performed on scRNA-seq data to categorize genes under varying biological conditions by quantifying the relatedness between genes. This relatedness is typically measured pairwise by calculating the correlations of gene expression levels across cells of each cell type. However, technical noise, particularly sparsity in single-cell sequencing data, impedes the inference of gene correlations, as the noise can attenuate or obscure the desired correlation estimates to an unpredictable degree [53, 40]. Metacell aggregation offers a solution to better estimate gene correlations by reducing technical noise while retaining cellular heterogeneity [45]. Nonetheless, this approach achieves its intended effectiveness only when the metacells are trustworthy, which means a metacell only includes cells from the same biological state. As we proved in the Methods Section, including dubious metacells in correlation estimation often leads to spurious co-expression findings. It is therefore crucial to apply mcRigor to eliminate dubious metacells from any downstream analysis, ensuring the reliability of discovering co-expressed gene pairs.

We applied mcRigor to an scRNA-seq dataset from human peripheral blood mononuclear cells (PBMC) of seven hospitalized COVID-19 patients and six healthy controls [57]. In this analysis, mcRigor successfully rectified co-expression estimates and uncovered biological pathways differentially regulated between the COVID-19 and healthy cohorts. Noticing that co-expression patterns are typically specific to both cell type and condition [53, 45], we applied SuperCell with $\gamma = 20$ separately on the 3028 B cells from COVID-19

samples and 1994 B cells from control samples. From the tow metacell partitions given by SuperCell (one per condition), mcRigor detected 26 dubious metacells among the 99 metacells identified for the healthy group and 22 dubious metacells among the 152 metacells for the diseased group.

We found that discarding dubious metacells prior to correlation estimates removed biases and highlighted differentially co-expressed gene modules that may play a role in the COVID-19 disease mechanism. For example, correlation estimation using only trustworthy metacells revealed three co-expression gene modules enriched in the COVID-19 cohort, representing differential biological functions of B cells, including the antigen processing via MHC Class II module (p-value= $3.7e-31$ for Wilcoxon test), the adaptive immune response module (p-value= $7.6e-19$), and the response to interferon-alpha alpha (p-value= 0.00328). These enrichment signals were notably strengthened compared to those in the raw single cell data, where the Wilcoxon test p-values for these three modules are $2.0e-13$, 0.00043, 0.08441, respectively. We also noted that these enrichment findings align with those given by the CS-CORE method [53].

In contrast, estimation with dubious metacells included (using all metacells) yielded the counterintuitive result that the adaptive immune response module was not enriched in COVID-19 patients (p-value= 0.54632). We found that this misleading result arose from an artifact: the strong co-expression of the adaptive immune response module under the healthy condition. Notably, this artifact was absent at single-cell resolution and induced solely by the dubious metacells (Fig 1d). Applying mcRigor to the SEACells, MetaCell, and MetaCell2 partitions similarly removed the artifact correlation introduced by dubious metacells and strengthened enrichment signals (Fig S3), demonstrating the applicability and robustness of mcRigor across different metacell methods.

3.4 mcRigor empowers and rectifies gene regulatory inference on single-cell multiome data

Single-cell multiome or integrated scATAC-seq and scRNA-seq data [41, 24, 55], which provide both gene expression and chromatin accessibility modalities in the same cells (through direct measurement or compuational integration), offer a lens to study the association between the two modalities, revealing the relationships between genes and regulatory elements (e.g., enhancers) at a finer resolution (e.g., being cell-type-specific) than is possible with bulk multi-omics data. However, associating genes with their regulatory elements can be challenging due to the high sparsity and noise in single-cell sequencing data, especially in scATAC-seq, so the reliability of inferred regulatory associations becomes questionable. Hence, aggregating homogeneous single cells into metacells has been implemented to help reduce data sparsity and improve gene regulatory inference [46, 43].

Persad et al. (2023) [46] demonstrated the effectiveness of metacell aggregation for empowering gene regulatory inference by applying their metacell partitioning method, SEACells, to a single-cell multiome dataset consisting of 6,800 hematopoietic stem and progenitor cells (HSPCs) from healthy bone marrow sorted for the pan-HSPC marker CD34. Compared to single-cell-level analysis, using SEACells (the metacells generated by the SEACells method) significantly reduced data sparsity in both modalities, particularly for chromatin accessibility, and revealed gene-peak associations that were obscured at the single-cell resolution (peaks represent open chromatin regions identified from the chromatin accessibility data).

On top of this analysis, we found that applying mcRigor to filter out dubious metacells improved gene regulatory inference in two key ways: (1) enhancing the identification of reliable gene-peak associations and (2) removing associations that were likely spurious. From the SEACells metacell partition in the original study [46], mcRigor identified 7 dubious metacells out of the 85 SEACells (Fig S4a). The effectiveness of mcRigor in improvement (1) is supported by the observation that consistently identified gene-peak pairs (adjusted p-value < 0.05), displayed significantly higher gene-peak association scores after dubious metacells were excluded (Top left of Fig 1e). For instance, the association score between the key erythroid lineage regulator *TAL1* and its most correlated peak increased from 0.8266 to 0.8703 when dubious metacells were removed. Similarly, the association score for another crucial erythroid factor, *GATA2*, with its highest correlated peak, rose from 0.6904 to 0.7606 (bottom left of Fig 1e).

Furthermore, we demonstrated that mcRigor could recover gene-peak associations supported by the literature while filtering out unlikely associations (improvements (1) and (2)). By using the trustworthy metacells identified by mcRigor, we identified 5,551 highly correlated gene-peak pairs, compared to the 5,536 pairs identified using all metacells. Although the increase in the number of gene-peak pairs was small, mcRigor refined the associations by removing weakly supported pairs and adding those with stronger data support.

For instance, while both all metacells and trustworthy metacells identified 16 correlated peaks for the gene *GATA2*, the specific peaks differed. When dubious metacells were excluded, an additional peak was identified while one previously identified peak was removed. The newly identified peak, *chr3-128532902-128533402*, which overlaps with LOC117038772 (*chr3-128532862-128533362*), is an enhancer for GATA2, as supported by previous reports [20, 37]. The peak's correlation with GATA2 expression was also evident at the single-cell level (bottom right of Fig 1e). In contrast, the peak not detected when using trustworthy metacells, *chr3-128409363-128409863*, showed minimal correlation with GATA2 expression at single-cell resolution, and its accessibility was extremely low across all cell types, providing insufficient evidence for association (Fig 1e, Fig S4b). These findings confirm that mcRigor enhances regulatory analysis and, importantly, helps pinpoint likely false positives.

3.5 Simulation verification of mcRigor's ability optimize metacell partitioning

A key hyperparameter in metacell partitioning is the granularity level γ, defined as the ratio of the number of single cells to the number of metacells (i.e., the average number of single cells per metacell). This hyperparameter controls the extent of size reduction from single cells to metacells, with a larger γ indicating greater reduction [5]. If γ is too large, it aggregates heterogeneous single cells into metacells, distorting downstream analyses. Conversely, if γ is too small, the metacells closely resemble single cells, failing to address the sparsity issue. Therefore, optimizing metacell partitioning is fundamentally about balancing signal distortion and data sparsity.

To quantify this tradeoff and guide hyperparameter selection, mcRigor introduces two metrics: *DubRate* and *ZeroRate*, which measure the level of distortion introduced and the remaining level of sparsity, respectively. Intuitively, as γ increases, *DubRate* rises while *ZeroRate* falls. mcRigor then calculates an evaluation score, termed *Score*, for each γ, defined as one minus a weighted sum of *DubRate* and *ZeroRate*, where a higher *Score* indicates a better metacell partition. The γ value that maximizes this score is considered to balance the improvement of data sparsity and the preservation of biological signals.

We tested the effectiveness of *Score* as an optimization criterion on the semi-synthetic dataset used previously and described in detail in the Methods section, with the true granularity level as $\gamma^* = 50$. MetaCell was applied as the metacell partitioning method, with γ values ranging from 2 to 100, generating one metacell partition per γ value. mcRigor was then used to evaluate the *DubRate*, calculate the *ZeroRate*, and determine the final *Score* for each partition. We observed that *DubRate* exhibited an elbow point at the true granularity level, and the highest *Score* was achieved precisely at $\gamma = \gamma^*$ (Fig 3a). Additionally, the optimized metacell partition closely matched the ground-truth metacells defined in the simulation, containing only 4 dubious metacells (Fig 3a).

We similarly applied SEACells to identify metacell partitions and used mcRigor for granularity level optimization. In this case, *DubRate* displayed a less pronounced elbow around $\gamma = 50$ and mcRigor's optimized granularity level was 42, also close to the true γ^* (Fig S5). However, for SuperCell and MetaCell2, the optimal γ's selected by mcRigor ($\gamma = 4$ for both) are far away from γ^*. This is because most metacells built by these two methods were dubious, resulting in high *DubRate* values even at small granularity levels (Fig S5).

Since mcRigor's optimization criterion, *Score* (ranging from 0 to 1), is comparable across metacell partitions regardless of the partitioning method used, mcRigor can simultaneously determine the optimal hyperparameter γ for each method and guide method selection for a specific single-cell dataset. For the semi-synthetic dataset, the maximal *Score* (i.e., the *Score* at the optimal γ for each method) attained by MetaCell, SEACells, SuperCell, and MetaCell2 were 0.692, 0.642, 0.537, and 0.528, respectively (Fig S5). Thus, the optimal configuration selected by mcRigor is MetaCell with $\gamma = 50$, which best matches the ground truth (Fig 3a, Fig S5). The lower maximal *Score* values for SuperCell and MetaCell2 compared to MetaCell and SEACells imply that MetaCell and SEACells produced better metacell partitions, with fewer dubious metacells and a higher reduction of sparsity. This finding is consistent with our observations on real single-cell data, where mcRigor often recommends SEACells and MetaCell over SuperCell and MetaCell2, supporting mcRigor's reliability for real data analysis. Notably, mcRigor is task-agnostic and does not rely on prior knowledge, enabling its application to metacell-based data analysis and making it an efficient and unbiased benchmarking tool for all metacell partitioning methods.

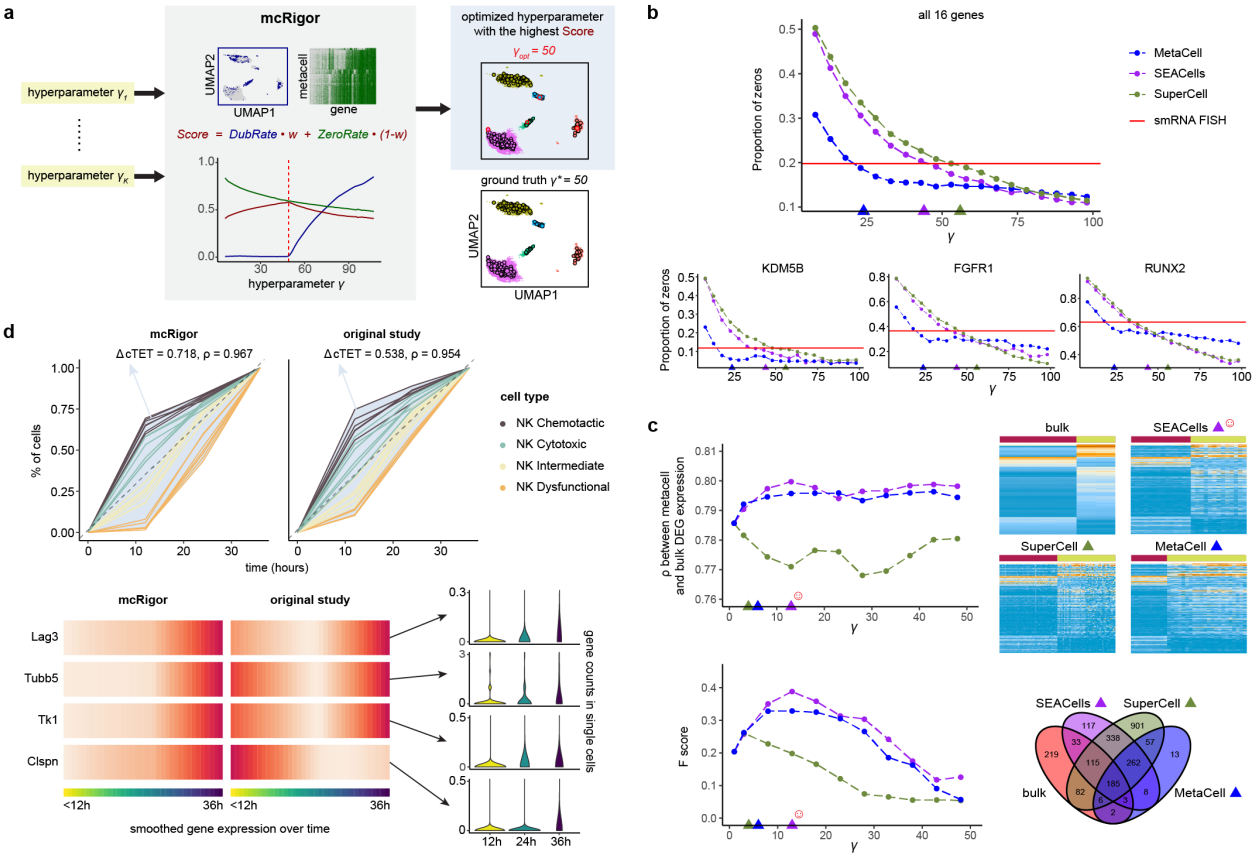

Fig. 3. mcRigor optimizes metacell hyperparameter and method selection for various single cell analysis tasks. a, Schematics of the mcRigor method pipeline for optimizing metacell partitioning using the simulation results on MetaCell method as an example. **b,**. Line plots showing the zero proportions in different metacell profiles. The zero proportion in smRNA FISH is drawn as the golden standard. **c,** mcRigor optimizes metacell configuration for differential gene expression analysis. Top: A line plot and heatmaps comparing the expression of bulk DE genes across different metacell partitions with the expression in bulk data. Bottom: A line plot and a Venn diagram comparing the DE gene lists derived from different metacell partitions and from the bulk data. The red smiling face marks the optimal metacell partition selected across all method-hyperparameter configurations. In **b** and **c**, the colored triangles represent the optimal granularity level selected by mcRigor. **d,** mcRigor's optimized metacell partition better reveals temporal immune cell trajectories. Top: Line plots showing the metacell cTET values derived from the mcRigor partitioning versus the original partitioning. ΔcTET is proportional to the light blue area size. Bottom: Smoothed gene expression of four marker genes derived from the mcRigor partitioning versus the original partitioning.

3.6 mcRigor optimizes metacell partitioning to distinguish biological and non-biological zeros

Single-cell data contain two types of zeros: biological zeros, which indicate absent or extremely low gene expression, and non-biological zeros, which result from gene expression being missed during the sequencing process [29]. Ideally, biological zeros should be retained as they provide valuable insights into the cells' biological states, while non-biological zeros should be filtered out to enhance data quality. Metacell aggregation offers a potential solution to this sparsity issue by averaging the expression profiles of multiple cells assumed to share the same biological state, thereby reducing the number of zeros. However, current metacell partitioning methods provide no guarantee to achieve two critical objectives: (1) preserving biological zeros and (2) removing non-biological zeros. This challenge arises due to the inherent trade-off between these two objectives—when the granularity level increases (i.e., more cells are merged into a metacell), the chance of removing non-biological zeros improves, but the risk of inadvertently eliminating biological zeros also rises.

Despite this challenge, mcRigor helps to balance the two competing objectives, which are about distinguishing biological zeros from non-biological zeros, by optimizing the granularity level for a metacell partitioning method. The optimization approach of mcRigor essentially minimizes the formation of dubious metacells (addressing objective (1)) while maximizing the removal of non-biological zeros (addressing objective (2)). Notably, minimizing the occurrence of dubious metacells is crucial for objective (1), as a dubious metacell contains cells from different states, where a gene may be expressed in one state but not in another. Averaging such mixed states could inadvertently eliminate biological zeros.

We evaluated whether mcRigor can effectively distinguish biological zeros from non-biological ones using an scRNA-seq Drop-seq dataset paired with single-molecule RNA fluorescence in situ hybridization (smRNA FISH) data for 16 genes from a melanoma cell line [54]. The rationale for this approach is that smRNA FISH is widely considered the gold standard for single-cell gene expression measurement [54, 25], making it reasonable to assume that all zeros in smRNA FISH data are biological zeros. We applied MetaCell, SEACells, and SuperCell to the scRNA-seq dataset, varying the granularity levels ($\gamma = 2, \ldots, 100$), and calculated the proportion of zeros for the 16 genes in the resulting metacell-by-gene data matrix for each metacell partition. We then used mcRigor to calculate a *Score* for each partition and identify the optimal γ that yielded the best *Score* for each method (Fig S6a). Notably, the γ value selected by mcRigor for each metacell partitioning method resulted in a proportion of zeros that closely matched the proportion of zeros in the smRNA FISH data (Fig 3b). This result confirms that mcRigor can help distinguish biological zeros from non-biological zeros by optimizing metacell partitioning.

3.7 mcRigor optimizes metacell configuration for differential gene expression analysis

Differential gene expression (DGE) analysis is a critical downstream application of scRNA-seq data, but it often suffers from data sparsity, leading to biased results and reduced statistical power when detecting gene expression differences between cell populations. Aggregating single cells into metacells is one strategy to mitigate the sparsity, but the reliability of metacells is crucial. To evaluate whether mcRigor improves the reliability of metacell-based DGE analysis, we faced the challenge of the absence of ground truth DEGs in scRNA-seq data. Therefore, we adopted an indirect validation approach: comparing the consistency of scRNA-seq DGE results with those from paired bulk RNA-seq data [35]. We applied this strategy to validate mcRigor's effectiveness, assessing whether the optimized metacell partition improved the consistency of DGE results with paired bulk RNA-seq data using a dataset obtained from both bulk and single-cell RNA-seq experiments on human embryonic stem cells (ESC) and definitive endoderm cells (DEC) [15].

We applied MetaCell, SuperCell, and SEACells with varying granularity levels ($\gamma = 2, \ldots, 100$) to generate metacell partitions from the scRNA-seq data. mcRigor then calculated the *Score* value for each partition and optimized γ for each method: $\gamma = 6$ for MetaCell, $\gamma = 4$ for SuperCell, and $\gamma = 13$ for SEACells. Among the three methods, SEACells with $\gamma = 13$ achieved the highest *Score*, making its metacell partition the optimal choice (Fig S7). As the first validation step, we used the DESeq2 method [39] to identify differentially expressed (DE) genes from the bulk RNA-seq data and assessed the concordance between the bulk and metacell profiles based on the expression of the top 200 DE genes. The optimal metacell partition showed the strongest concordance with the bulk data (Pearson correlation $\rho = 0.800$) (top panel of Fig 3c). Concordance here is defined as the Pearson correlation between two expression vectors, each representing the top 200 DE genes in the bulk data (average of bulk samples) and or the metacell partition (average of metacells). For details, see the Methods Section.

Next, we evaluated mcRigor's ability to identify the metacell partition that could best recover the bulk DE genes from the paired scRNA-seq data. We applied DESeq2 to the scRNA-seq data given each metacell partition and recorded the DE genes identified at the false discovery rate (FDR) threshold of 0.05. Treating the bulk DE genes obtained from bulk data (also using a 0.05 FDR threshold) as the ground truth, we then calculated an F-score (the geometric mean of precision and recall) for the metacell DE genes for each metacell partition. The metacell configuration selected by mcRigor (SEACells with $\gamma = 13$, Fig S7) produced the highest F-score (0.400) which is almost twice of the F-score (0.204) for the single-cell DE genes without using metacells (bottom panel of Fig 3c). These findings indicate that the DE genes identified based on the optimal metacell partition had the best agreement with the bulk DE genes, thereby affirming the reliability of the optimization of metacell partitioning (the selection of metacell method and hyperparameter) by mcRigor.

3.8 mcRigor's optimized metacell partition better reveals temporal immune cell trajectories

In oncogenesis, immune cells infiltrating tumors can be co-opted by the immunosuppressive tumor microenvironment (TME), transitioning from a cytotoxic to a dysfunctional state with reduced antitumor activity. To explore the transcriptomic dynamics of immune cells over time, a new technology, Zman-seq, was developed to collect data from 2,431 intratumoral T and natural killer (NK) cells in mouse glioblastoma, along with time stamps documenting each cell's exposure duration within the tumor [30]. In the original study, the 2,431 single cells were aggregated into 37 metacells using the MetaCell method with $\gamma = 66$, and temporal trajectories were tracked among these metacells, which were expected to represent distinct transcriptional states.

Specifically, the original study calculated a continuous tumor exposure time (cTET) value for each metacell. This was defined as one minus the normalized area under the curve (AUC) of the cumulative distribution function (CDF) for the frequencies of single cells within the metacell at three time points: 12h, 24h, and 36h. In other words, the CDF represents a three-value discrete distribution (if all single cells in a metacell are from the 12h time point, the AUC of the CDF is 1, and the cTET is 0). The cTET values were then used as continuous temporal labels for the metacells, where larger values indicated longer tumor exposure times. Accordingly, the following transitional stages of NK cells were expected to show increasing cTET values: NK chemotactic, NK cytotoxic, NK intermediate, and finally NK dysfunctional cells. However, in the original study, the cTET values of the metacells did not clearly distinguish these stages (see the right panel of Fig 2e), suggesting that some metacells might be dubious (containing a mix of NK cells of different biological states).

We applied mcRigor to optimize the metacell method-hyperparameter configuration and found that SEACells with $\gamma = 43$ provided the optimal metacell partition (Fig S8b). Specifically, compared to the original metacell partition (with $DubRate = 0.222$), the optimized partition had a significantly lower $DubRate$ of 0.056. With this optimal metacell partition, the cTET values more effectively distinguished the four transitional stages, as indicated by a greater cTET difference between the earliest- and latest-stage metacells (0.718 vs. 0.538 in the original study). Additionally, the cTET values showed a much stronger correlation with the transitional stages of the metacells (Spearman's rank correlation = (0.967 vs. 0.954 in the original study) (top panel of Fig 3d). See the Methods section for further details on this data analysis.

In the original study, subsequent analysis involved ordering metacells based on their cTET values to identify genes most correlated with tumor exposure time. To evaluate the reliability of the metacell partition optimized by mcRigor, we identified 75 differentially expressed (DE) genes (FDR = 0.05) between the 12h and 36h time points using the original single-cell data and assessed whether these DE genes could be successfully recovered at the metacell level. The rationale is that while single-cell DE analysis often lacks power due to data sparsity, the identified single-cell DE genes should still be detectable from metacells, which are expected to enhance the power of DE analysis. Our results showed that these single-cell DE genes exhibited higher correlations with tumor exposure time (mean correlation 0.633 vs. 0.570 in the original study) and lower p-values in DE analysis between 12h and 36h (mean adjusted p-value (0.035 vs. 0.118 in the original study) when using the metacells from the optimal partition compared to the original partition (Fig S8c). Furthermore, the DE p-values of these genes at the metacell level were more consistent with the single-cell p-values (Spearman's correlation $\rho = 0.517$ vs. 0.452 in the original study). In total, the optimal metacell partition, compared to the original partition, helped identify more genes that are significantly correlated with tumor exposure time (260 vs. 135 in the original study). These findings indicate that by reducing dubious metacells, mcRigor helps preserve biological signals in single-cell data, improving the reliability of downstream analyses, and gain more power in uncovering important genes associated with time.

Additionally, mcRigor's optimized metacell partition corrected the questionable gene temporal patterns inferred from the metacells in the original study (bottom panel of Fig 3d). For instance, *Lag3*, a gene associated with the suppression of antitumor functions [21, 44] and known to act as a receptor for the ligand LSECtin [59], blunting tumor-specific immune responses, was incorrectly inferred to follow a pattern of initial downregulation followed by upregulation in the original study. In contrast, the optimized metacell partition identified a more biologically plausible pattern, showing *Lag3* as steadily upregulated over time, consistent with single-cell observations and previous studies that found *Lag3* increasingly expressed in NK cells as they progress towards a terminal intratumoral state [18]. Similarly, the optimized metacell partition corrected the temporal pattern for *Clspn*, aligning with both single-cell results and the known biology of the gene's human ortholog (gene *CLSPN*), which is typically upregulated in the immune microenvironment of most

cancer types [14, 28]. This was in contrast to the original study, where *Clspn* was incorrectly inferred to be downregulated over time. These findings demonstrate the necessity of using mcRigor to generate reliable metacell partitions, ensuring the accuracy of temporal gene expression analysis.

It is also worth noting that despite the finer granularity ($\gamma = 43$ in the optimized partition vs. $\gamma = 66$ in the original partition), the optimized partition still effectively resolved data sparsity and enhanced biological signals. Specifically, the optimized partition successfully recovered all the important temporally dynamic genes identified in the original study (Fig S8d), including the upregulated module (*Pmepa1*, *Car2*, *Ctla2a*, *Xcl1*, *Itga1*, *Gzmc*, etc.) and the downregulated module (*S1pr5*, *Cx3cr1*, *Klrg1*, *Cma1*, *Ccl3*, *Gzmb*, etc.).

3.9 Execution time of mcRigor

We report the execution time of mcRigor on a Intel Xeon E5-2660 v3 system with 2.6 GHz CPU and 90.64 GB RAM (Table 1). mcRigor's computational cost for detecting dubious metacells is low, consistently under 2 minutes for all datasets. While optimizing metacell configuration takes longer——typically under two hours for a large dataset of around $10,000$ cells——it remains manageable. Moreover, the cost can be further reduced by using a smaller candidate pool of γ values, such as $\gamma = 5, 10, \ldots, 100$ instead of $\gamma = 2, 3, \ldots, 100$, This is a reasonable approach in practice since close γ values often yield similar partitionings [5].

Table 1. The execution time of mcRigor (in minutes) for detecting dubious metacells and optimizing metacell configuration. original: the time for implementing the original metacell methods; mcRigor: the time for applying mcRigor on the derived metacell partition(s).; "—" indicates the method is not applicable to the data.

Dataset	# cell	γ	Detecting dubious metacells								γ	Optimizing metacell partitioning							
			SEACells		MetaCell		MetaCell2		SuperCell			SEACells		MetaCell		MetaCell2		SuperCell	
			original	mcRigor	original	mcRigor	original	mcRigor	original	mcRigor		original	mcRigor	original	mcRigor	original	mcRigor	original	mcRigor
COVID-19 PBMC	3028	50	5.646	0.185	3.840	0.232	0.367	0.216	0.162	0.186	2—100	920.064	8.788	225.412	7.382	17.281	7.734	17.283	8.962
scMulitome	6800	100	9.519	1.319	—	—	—	—	—	—	2—200	>1000	91.183	—	—	—	—	743.041	104.013
semi-synthetic	13400	50	3.001	1.620	2.066	1.543	2.830	1.782	1.712	1.695	2—100	>1000	136.022	220.464	125.906	327.598	116.879	167.522	148.076
Dropseq + smRNA FISH	8498	50	58.608	0.721	4.057	0.673	2.394	0.659	1.956	0.801	2—100	>1000	75.586	479.197	68.833	200.246	73.795	193.688	75.281
scRNA-seq + bulk ESC	350	30	0.568	0.192	0.821	0.173	—	—	0.038	0.153	2—50	26.562	4.070	39.654	3.247	—	—	1.935	3.646
Zman-seq	2431	50	4.749	0.185	1.341	0.167	0.578	0.185	0.219	0.177	2—100	774.616	8.284	154.848	7.926	30.493	7.383	22.242	7.925

4 Discussion

mcRigor is a novel statistical method designed to enhance the rigor of metacell partitioning in single-cell data analysis, ensuring reliable downstream analyses on metacells. By evaluating a given metacell partition, mcRigor identifies dubious metacells through quantifying per-metacell heterogeneity (i.e., the presence of mixed biological states) using a feature-correlation-based statistic. This statistic is assessed against a null distribution generated through a novel double permutation mechanism. Our findings demonstrate that the dubious metacells identified by mcRigor are indeed heterogeneous, and their removal significantly enhances downstream analyses, such as gene co-expression studies and enhancer-gene regulatory inference. This improvement is achieved by revealing biological signals otherwise obscured by data sparsity and correcting conclusions drawn from distorted signals provided by dubious metacells. Leveraging its dubious metacell detection capability, mcRigor further optimizes metacell partitioning by selecting the method-hyperparameter configuration that best balances the trade-off between reducing data sparsity (fewer, larger metacells) and preserving biological signals (more, smaller metacells). This optimized configuration has demonstrated improved performance across various tasks, including distinguishing biological from non-biological zeros, identifying differentially expressed genes, and elucidating temporal immune cell trajectories.

We anticipate that mcRigor will be a valuable computational tool for single-cell researchers, enabling them to generate reliable metacells for addressing key scientific questions with metacell-level data. First, mcRigor detects dubious metacells by constructing null values for the per-cell mcDiv score through double permutation, which is performed only once on the cell-by-feature matrix of each metacell. From this double-permuted matrix, a null value (mcDiv^{null}) is computed for that metacell, and the mcDiv threshold is determined by pooling the null values of metacells of similar sizes. Future research could explore whether

performing multiple rounds of double permutations, thereby generating multiple null values for each metacell, could lead to more robust detection of dubious metacells.

Second, while mcRigor can detect dubious metacells, it does not offer a strategy for reorganizing these metacells into trustworthy metacells; users are currently limited to removing dubious metacells in downstream analysis. A more comprehensive approach could involve recursively partitioning dubious metacells to achieve internal homogeneity, ensuring that all metacells are trustworthy without losing individual cells. Developing such a recursive partitioning algorithm remains an open question for future research. Furthermore, an interesting direction would be to design a new metacell partitioning method that constructs metacells directly from single cells based on mcRigor's evaluation criteria. This method should focus on generating trustworthy metacells from the outset, thereby minimizing the need for post-generation filtering.

Third, mcRigor could be enhanced to better handle multi-modality data (e.g., scRNA-seq and scATAC-seq). Currently, it operates on metacell partitions derived from a single modality, requiring users or the partitioning method to select which modality to prioritize. This single-modality approach may lead to suboptimal metacell partitions when modalities contain complementary biological signals. Therefore, a key future direction is to extend mcRigor to integrate multiple modalities in a data-driven manner. Additionally, our future work will explore metacell partitioning beyond single-cell gene expression and chromatin accessibility data. For example, the spatial niche concept in spatial transcriptomics is related to, but distinct from, the metacell concept, as spatial niches often encompass multiple cell types and represent complex microenvironments rather than homogeneous cell groupings. Defining reliable criteria for spatial niches remains an open question. Expanding mcRigor's capabilities to support additional modalities and data types would increase its utility as a comprehensive tool for addressing data sparsity in diverse genomics data analyses.

5 Supplementary Material

The supplementary material includes the statistical justification for our metacell definition, details on the semi-synthetic data generation process, further analysis methodologies, data and code availability, as well as additional figures.

References

1. Andrews, T.S., Hemberg, M.: False signals induced by single-cell imputation. F1000Research **7**, 1740 (2018)
2. Ashuach, T., Reidenbach, D.A., Gayoso, A., Yosef, N.: Peakvi: A deep generative model for single-cell chromatin accessibility analysis. Cell reports methods **2**(3), 100182 (2022)
3. Baran, Y., Bercovich, A., Sebe-Pedros, A., Lubling, Y., Giladi, A., Chomsky, E., Meir, Z., Hoichman, M., Lifshitz, A., Tanay, A.: Metacell: analysis of single-cell rna-seq data using k-nn graph partitions. Genome Biology **20**, 1–19 (2019)
4. Ben-Kiki, O., Bercovich, A., Lifshitz, A., Tanay, A.: Metacell-2: a divide-and-conquer metacell algorithm for scalable scrna-seq analysis. Genome Biology **23**(1), 100 (2022)
5. Bilous, M., Hérault, L., Gabriel, A.A., Teleman, M., Gfeller, D.: Building and analyzing metacells in single-cell genomics data. Molecular Systems Biology **20**, 744–766 (2024)
6. Bilous, M., Tran, L., Cianciaruso, C., Gabriel, A., Michel, H., Carmona, S.J., Pittet, M.J., Gfeller, D.: Metacells untangle large and complex single-cell transcriptome networks. BMC Bioinformatics **23**(1), 336 (2022)
7. Bornstein, C., Nevo, S., Giladi, A., Kadouri, N., Pouzolles, M., Gerbe, F., David, E., Machado, A., Chuprin, A., Tóth, B., et al.: Single-cell mapping of the thymic stroma identifies il-25-producing tuft epithelial cells. Nature **559**(7715), 622–626 (2018)
8. Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., Greenleaf, W.J.: Single-cell chromatin accessibility reveals principles of regulatory variation. Nature **523**(7561), 486–490 (2015)
9. Cai, T.T., Ma, Z.: Optimal hypothesis testing for high dimensional covariance matrices. Bernoulli **19**(5B), 2359–2388 (2013)
10. Cai, T.T.: Global testing and large-scale multiple testing for high-dimensional covariance structures. Annual Review of Statistics and Its Application **4**(1), 423–446 (2017)
11. Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L., et al.: Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science **361**(6409), 1380–1385 (2018)

12. Chen, S., Lake, B.B., Zhang, K.: High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nature biotechnology **37**(12), 1452–1457 (2019)
13. Chen, S.X., Zhang, L.X., Zhong, P.S.: Tests for high-dimensional covariance matrices. Journal of the American Statistical Association **105**(490), 810–819 (2010)
14. Chen, Y., Wen, H., Li, Y., Han, Y., Tan, J., Guo, C., Cai, C., Liu, P., Peng, Y., Liu, Y., et al.: A multi-omics analysis reveals clspn is associated with prognosis, immune microenvironment and drug resistance in cancers. Biological Procedures Online **25**(1), 16 (2023)
15. Chu, L.F., Leng, N., Zhang, J., Hou, Z., Mamott, D., Vereide, D.T., Choi, J., Kendziorski, C., Stewart, R., Thomson, J.A.: Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Genome Biology **17**, 1–20 (2016)
16. Cuomo, A.S., Alvari, G., Azodi, C.B., eQTLGen Consortium, S.C., McCarthy, D.J., Bonder, M.J.: Optimizing expression quantitative trait locus mapping workflows for single-cell studies. Genome Biology **22**(1), 188 (2021)
17. Cusanovich, D.A., Daza, R., Adey, A., Pliner, H.A., Christiansen, L., Gunderson, K.L., Steemers, F.J., Trapnell, C., Shendure, J.: Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. Science **348**(6237), 910–914 (2015)
18. Dean, I., Lee, C.Y., Tuong, Z.K., Li, Z., Tibbitt, C.A., Willis, C., Gaspal, F.a., Kennedy, B.C., Matei-Rascu, V., Fiancette, R., et al.: Rapid functional impairment of natural killer cells following tumor entry limits anti-tumor immunity. Nature Communications **15**(1), 683 (2024)
19. Eraslan, G., Simon, L.M., Mircea, M., Mueller, N.S., Theis, F.J.: Single-cell rna-seq denoising using a deep count autoencoder. Nature communications **10**(1), 390 (2019)
20. Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., et al.: Genehancer: genome-wide integration of enhancers and target genes in genecards. Database **2017**, bax028 (2017)
21. Gandhi, M.K., Lambley, E., Duraiswamy, J., Dua, U., Smith, C., Elliott, S., Gill, D., Marlton, P., Seymour, J., Khanna, R.: Expression of lag-3 by tumor-infiltrating lymphocytes is coincident with the suppression of latent membrane antigen–specific cd8+ t-cell function in hodgkin lymphoma patients. Blood **108**(7), 2280–2289 (2006)
22. 10x Genomics: (July 2019), 5k PBMCs from a healthy donor, Single Cell ATAC dataset analyzed using Cell Ranger ATAC v1.1.0
23. Giladi, A., Paul, F., Herzog, Y., Lubling, Y., Weiner, A., Yofe, I., Jaitin, D., Cabezas-Wallscheid, N., Dress, R., Ginhoux, F., et al.: Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. Nature Cell Biology **20**(7), 836–846 (2018)
24. Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., Greenleaf, W.J.: Archr is a scalable software package for integrative single-cell chromatin accessibility analysis. Nature genetics **53**(3), 403–411 (2021)
25. Grün, D., Kester, L., Van Oudenaarden, A.: Validation of noise models for single-cell transcriptomics. Nature Methods **11**(6), 637–640 (2014)
26. Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., et al.: Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nature Biotechnology **42**(2), 293–304 (2024)
27. Huang, M., Wang, J., Torre, E., Dueck, H., Shaffer, S., Bonasio, R., Murray, J.I., Raj, A., Li, M., Zhang, N.R.: Saver: gene expression recovery for single-cell rna sequencing. Nature methods **15**(7), 539–542 (2018)
28. Jia, Y., Cheng, X., Liang, W., Lin, S., Li, P., Yan, Z., Zhang, M., Ma, W., Hu, C., Wang, B., et al.: Clspn is a potential biomarker associated with poor prognosis in low-grade gliomas based on a multi-database analysis. Current Research in Translational Medicine **70**(4), 103345 (2022)
29. Jiang, R., Sun, T., Song, D., Li, J.J.: Statistics or biology: the zero-inflation controversy about scrna-seq data. Genome Biology **23**(1), 31 (2022)
30. Kirschenbaum, D., Xie, K., Ingelfinger, F., Katzenelenbogen, Y., Abadie, K., Look, T., Sheban, F., San Phan, T., Li, B., Zwicky, P., et al.: Time-resolved single-cell transcriptomics defines immune trajectories in glioblastoma. Cell **187**(1), 149–165 (2024)
31. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., et al.: Eleven grand challenges in single-cell data science. Genome biology **21**, 1–35 (2020)
32. Lareau, C.A., Duarte, F.M., Chew, J.G., Kartha, V.K., Burkett, Z.D., Kohlway, A.S., Pokholok, D., Aryee, M.J., Steemers, F.J., Lebofsky, R., et al.: Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nature Biotechnology **37**(8), 916–924 (2019)
33. Lareau, C.A., Ma, S., Duarte, F.M., Buenrostro, J.D.: Inference and effects of barcode multiplets in droplet-based single-cell assays. Nature Communications **11**(1), 866 (2020)
34. Li, H., van der Leun, A.M., Yofe, I., Lubling, Y., Gelbard-Solodkin, D., van Akkooi, A.C., van den Braber, M., Rozeman, E.A., Haanen, J.B., Blank, C.U., et al.: Dysfunctional cd8 t cells form a proliferative, dynamically regulated compartment within human melanoma. Cell **176**(4), 775–789 (2019)

35. Li, W.V., Li, J.J.: An accurate and robust imputation method scimpute for single-cell rna-seq data. Nature communications **9**(1), 997 (2018)
36. Liu, B., Hu, X., Feng, K., Gao, R., Xue, Z., Zhang, S., Zhang, Y., Corse, E., Hu, Y., Han, W., et al.: Temporal single-cell tracing reveals clonal revival and expansion of precursor exhausted t cells during anti-pd-1 therapy in lung cancer. Nature cancer **3**(1), 108–121 (2022)
37. LOC117038772 GeneCards: (2016), `https://www.genecards.org/cgi-bin/carddisp.pl?gene=LOC117038772`
38. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., Yosef, N.: Deep generative modeling for single-cell transcriptomics. Nature methods **15**(12), 1053–1058 (2018)
39. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biology **15**, 1–21 (2014)
40. Lu, S., Keleş, S.: Debiased personalized gene coexpression networks for population-scale scrna-seq data. Genome Research **33**(6), 932–947 (2023)
41. Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., et al.: Chromatin potential identified by shared single-cell profiling of rna and chromatin. Cell **183**(4), 1103–1116 (2020)
42. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al.: Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell **161**(5), 1202–1214 (2015)
43. Mao, H., Jia, M., Dou, J.X., Zhang, H., Benos, P.V.: Coem: cross-modal embedding for metacell identification. arXiv preprint arXiv:2207.07734 (2022)
44. Matsuzaki, J., Gnjatic, S., Mhawech-Fauceglia, P., Beck, A., Miller, A., Tsuji, T., Eppolito, C., Qian, F., Lele, S., Shrikant, P., et al.: Tumor-infiltrating ny-eso-1–specific cd8+ t cells are negatively regulated by lag-3 and pd-1 in human ovarian cancer. Proceedings of the National Academy of Sciences **107**(17), 7875–7880 (2010)
45. Morabito, S., Reese, F., Rahimzadeh, N., Miyoshi, E., Swarup, V.: hdwgcna identifies co-expression networks in high-dimensional transcriptomics data. Cell Reports Methods **3**(6), 100498 (2023)
46. Persad, S., Choo, Z.N., Dien, C., Sohail, N., Masilionis, I., Chaligné, R., Nawy, T., Brown, C.C., Sharma, R., Pe'er, I., et al.: Seacells infers transcriptional and epigenomic cellular states from single-cell genomics data. Nature Biotechnology **41**(12), 1746–1757 (2023)
47. Pfister, D., Núñez, N.G., Pinyol, R., Govaere, O., Pinter, M., Szydlowska, M., Gupta, R., Qiu, M., Deczkowska, A., Weiner, A., et al.: Nash limits anti-tumour surveillance in immunotherapy-treated hcc. Nature **592**(7854), 450–456 (2021)
48. Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., Sandberg, R.: Full-length rna-seq from single cells using smart-seq2. Nature protocols **9**(1), 171–181 (2014)
49. Sarkar, A., Stephens, M.: Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. Nature Genetics **53**(6), 770–777 (2021)
50. Sebé-Pedrós, A., Saudemont, B., Chomsky, E., Plessier, F., Mailhé, M.P., Renno, J., Loe-Mie, Y., Lifshitz, A., Mukamel, Z., Schmutz, S., et al.: Cnidarian cell type diversity and regulation revealed by whole-organism single-cell rna-seq. Cell **173**(6), 1520–1534 (2018)
51. Song, D., Wang, Q., Yan, G., Liu, T., Sun, T., Li, J.J.: scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. Nature Biotechnology **42**(2), 247–252 (2024)
52. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., Satija, R.: Comprehensive integration of single-cell data. Cell **177**(7), 1888–1902 (2019)
53. Su, C., Xu, Z., Shan, X., Cai, B., Zhao, H., Zhang, J.: Cell-type-specific co-expression inference from single cell rna-sequencing data. Nature Communications **14**(1), 4846 (2023)
54. Torre, E., Dueck, H., Shaffer, S., Gospocic, J., Gupte, R., Bonasio, R., Kim, J., Murray, J., Raj, A.: Rare cell detection by single-cell rna sequencing as guided by single-molecule rna fish. Cell Systems **6**(2), 171–179 (2018)
55. Trevino, A.E., Müller, F., Andersen, J., Sundaram, L., Kathiria, A., Shcherbina, A., Farh, K., Chang, H.Y., Pașca, A.M., Kundaje, A., et al.: Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. Cell **184**(19), 5053–5069 (2021)
56. Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al.: Recovering gene interactions from single-cell data using data diffusion. Cell **174**(3), 716–729 (2018)
57. Wilk, A.J., Rustagi, A., Zhao, N.Q., Roque, J., Martínez-Colón, G.J., McKechnie, J.L., Ivison, G.T., Ranganath, T., Vergara, R., Hollis, T., et al.: A single-cell atlas of the peripheral immune response in patients with severe covid-19. Nature Medicine **26**(7), 1070–1076 (2020)
58. Xia, L., Lee, C., Li, J.J.: Statistical method scdeed for detecting dubious 2d single-cell embeddings and optimizing t-sne and umap hyperparameters. Nature Communications **15**(1), 1753 (2024)
59. Xu, F., Liu, J., Liu, D., Liu, B., Wang, M., Hu, Z., Du, X., Tang, L., He, F.: Lsectin expressed on melanoma cells promotes tumor progression by inhibiting antitumor t-cell responses. Cancer Research **74**(13), 3418–3428 (2014)

60. Xue, A., Yazar, S., Neavin, D., Powell, J.E.: Pitfalls and opportunities for applying latent variables in single-cell eqtl analyses. Genome Biology **24**(1), 33 (2023)
61. Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al.: Massively parallel digital transcriptional profiling of single cells. Nature Communications **8**(1), 14049 (2017)