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, USA
[2]Department of Biostatistics, University of California, Los Angeles, CA, USA.
[3]Department of Computational Medicine, University of California, Los Angeles, CA, USA
[4]Department of Human Genetics, University of California, Los Angeles, CA, USA
[5]Bioinformatics Interdepartmental Ph.D. Program, University of California, Los Angeles, CA, USA

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 scheme. 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 indicate that among existing metacell partitioning methods, MetaCell and SEACells consistently outperform MetaCell2 and SuperCell, albeit with the trade-off 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) (Picelli et al., 2014; Macosko et al., 2015; Zheng et al., 2017), epigenomics through single-cell assay for transposase-accessible chromatin using sequencing (scATAC-seq) (Buenrostro et al., 2015; Cusanovich et al., 2015), and multiome assays that simultaneously measure RNA-seq and ATAC-seq (Cao et al., 2018; Chen et al., 2019). 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 (Jiang et al., 2022).

The high sparsity presents a substantial challenge for data analysis (Lähnemann et al., 2020), with common strategies to mitigate it including *imputation* and *metacell aggregation*. Imputation addresses sparsity by predicting missing feature measurements, where *features* represent genes or chromatin regions, using cells and/or features with similar measurement profiles. Numerous imputation methods have been developed for single-cell data, including scImpute (Li and Li, 2018), SAVER (Huang et al., 2018), MAGIC (Van Dijk et al., 2018), and DCA (Eraslan et al., 2019), as well as deep generative models (Lopez et al., 2018; Ashuach et al., 2022). 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 (Andrews and Hemberg, 2018) and may suffer from oversmoothing (Huang et al., 2018; Baran et al., 2019), which creates artificial similarities among cells. As an alternative to imputation, the metacell approach groups cells representing the same cell state into a metacell and uses the metacell's measurement profile, typically obtained by averaging the single-cell measurement profiles, for subsequent analysis (Baran et al., 2019). 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 (Cuomo et al., 2021; Xue et al., 2023). 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. For instance, metacell aggregation has demonstrated beneficial for gene co-expression analysis (Morabito et al., 2023). While co-expression analysis using pseudobulk samples (Hartl et al., 2021; Pearl et al., 2019) fails to capture cell-type-specific signals, single-cell analysis is often hindered by the low sensitivity and high technical noise of scRNA-seq data (Chu et al., 2024). The metacell approach offers a valuable middle ground by enhancing co-expression signals within specific cell types.

Despite the increasing use of the metacell concept in high-profile single-cell studies—such as investigating cell differentiation states (Giladi et al., 2018), characterizing tissue compartments and diverse cell populations (Sebé-Pedrós et al., 2018; Bornstein et al., 2018; Li et al., 2019), patient stratification for individualized immunotherapy design (Pfister et al., 2021), and temporal analysis of cell transcriptomes (Liu et al., 2022; Kirschenbaum et al., 2024)—there is still no rigorously definition of metacell or a universally accepted strategy for constructing metacells. This lack of consensus can result in inconsistencies across studies utilizing the metacell concept, undermining the reliability of 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 (Baran et al., 2019), MetaCell2 (Ben-Kiki et al., 2022), SuperCell (Bilous et al., 2022), and SEACells (Persad et al., 2023). MetaCell employs a k-nearest neighbor (kNN) 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 (Bilous et al., 2024). 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 analyses.

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, termed *technical variation*, rather than biological differences, termed *biological variation*. 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 show 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 to distinguish biological from non-biological zeros, detect differentially expressed genes, and reveal temporal trajectories of cellular immune responses (Fig 2).

2 Results

2.1 Overview of the mcRigor method

The mcRigor method is designed to improve the rigor of metacell partitioning and the reliability of downstream analyses by distinguishing between *trustworthy* and *dubious* metacells, optimizing the hyperparameters of any metacell partitioning method, and facilitating comparisons across different metacell partitions. mcRigor is built on the *definition* of a metacell as a homogeneous group of single cells that share the same biological state, characterized by the same true profile of features (e.g., genes or chromatin regions), such that the cells' observed profiles can be considered resampled from the same original cell. The essence of this definition, detailed in the Methods Section, is that within a metacell, any variability among cells should be solely attributable to the technical variation. Aggregating such cells into a metacell helps reduce technical noise while preserving true biological signals, ensuring that the metacell's averaged profile faithfully represents the underlying biological state of these cells.

mcRigor has two main functionalities: detecting dubious metacells and optimizing metacell partitioning. For its dubious metacell detection functionality, mcRigor takes a single-cell dataset and an initial metacell partition as input, producing a refined set of trustworthy metacells as output. Specifically, mcRigor detects dubious metacells in the input partition using a statistical approach that assesses the internal homogeneity of each metacell (Fig 1a). The approach is grounded in the premise that within a trustworthy metacell, which is internally homogeneous, feature correlations are driven exclusively by technical noise and should therefore be minimal. At the core of this approach, we define a feature-correlation-based statistic called the *divergence score* (mcDiv) for each metacell, which measures the deviation of the within-metacell feature correlation matrix from the identity matrix that indicates no feature correlation. Note that mcDiv includes a normalization factor derived from within-feature per-

mutation, which simulates the baseline scenario where the same features become uncorrelated. A larger mcDiv value indicates greater heterogeneity within the metacell, suggesting it is more dubious.

To set mcDiv thresholds for identifying dubious metacells, mcRigor constructs a *null divergence score* ($\text{mcDiv}^{\text{null}}$) for each metacell through within-cell permutation, where feature values are shuffled independently for each cell, preserving cell library sizes (i.e., the sum of feature values per cell) while disrupting any biological correlations among features. As varying cell library sizes can induce spurious feature correlations, it is essential to preserve cell library sizes in defining $\text{mcDiv}^{\text{null}}$. Note that $\text{mcDiv}^{\text{null}}$ also includes a normalization factor derived from within-feature permutation applied to the features already permuted in the previous within-cell permutation step. Therefore, the calculation of $\text{mcDiv}^{\text{null}}$ involves two permutation steps: within-cell permutation followed by within-feature permutation, a novel procedure we refer to as the *double permutation*. Double permutation is necessary because mcDiv and $\text{mcDiv}^{\text{null}}$ involve different features and thus require different normalization factors. Pooling the $\text{mcDiv}^{\text{null}}$ values from all metacells, mcRigor establishes thresholds to distinguish dubious metacells from trustworthy ones. These thresholds are conditional on the *metacell size*, i.e., the number of single cells within the metacell, recognizing that metacells with smaller sizes may exhibit greater variability in mcDiv values. Specifically, to determine if each metacell is dubious, mcRigor employs a sliding window approach to compute a local threshold for the metacell, defined as the 95th percentile of $\text{mcDiv}^{\text{null}}$ values among similarly sized metacells. Metacells whose mcDiv values exceed the metacell-size-specific thresholds are flagged as dubious, indicating likely heterogeneity within the cells they contain.

For its metacell partition optimization functionality, mcRigor inputs a single-cell dataset along with metacell partitioning methods and their candidate hyperparameter settings, returning the optimized hyperparameter setting for each method and assessing the quality of each method's optimal metacell partition. That is, mcRigor optimizes metacell partitioning for a specific dataset by simultaneously evaluating various candidate method-hyperparameter configurations (Fig 2a), with each configuration representing a metacell method combined with its hyperparameter setting. The most critical hyperparameter is the *granularity level*, γ, which represents the average number of single cells per metacell, since γ is required by all existing metacell methods. For each metacell partition obtained from a method-hyperparameter configuration, mcRigor calculates an *evaluation score* that balances the dubious rate (the proportion of cells in dubious metacells) and the sparsity level in the aggregated data (measured by the proportion of zeros in the metacell expression profiles), recognizing that these two factors often trade off against each other (metacells containing more cells are more likely to be dubious but have less sparse profiles). The configuration that maximizes this evaluation score is considered optimal, striking a balance between preserving biological signals and minimizing technical noise. Notably, mcRigor is designed to be flexible and applicable across different metacell methods and single-cell data modalities.

2.2 Validation of mcRigor using barcode multiplet data

Our statistical definition of metacell and the mcRigor method are both built upon the assumption that the observed variation 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 (Lareau et al., 2020). 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), 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 (Lareau et al., 2020). 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.

In (Lareau et al., 2019), Laureau et al. 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 peripheral blood mononuclear cells (PBMCs) (10x Genomics, 2019). In total, 16 barcode multiples (caused by heterogeneous beads) were identified, 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 Supplementary Fig 1a). 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.

2.3 Assessment of mcRigor's accuracy in detecting dubious metacells

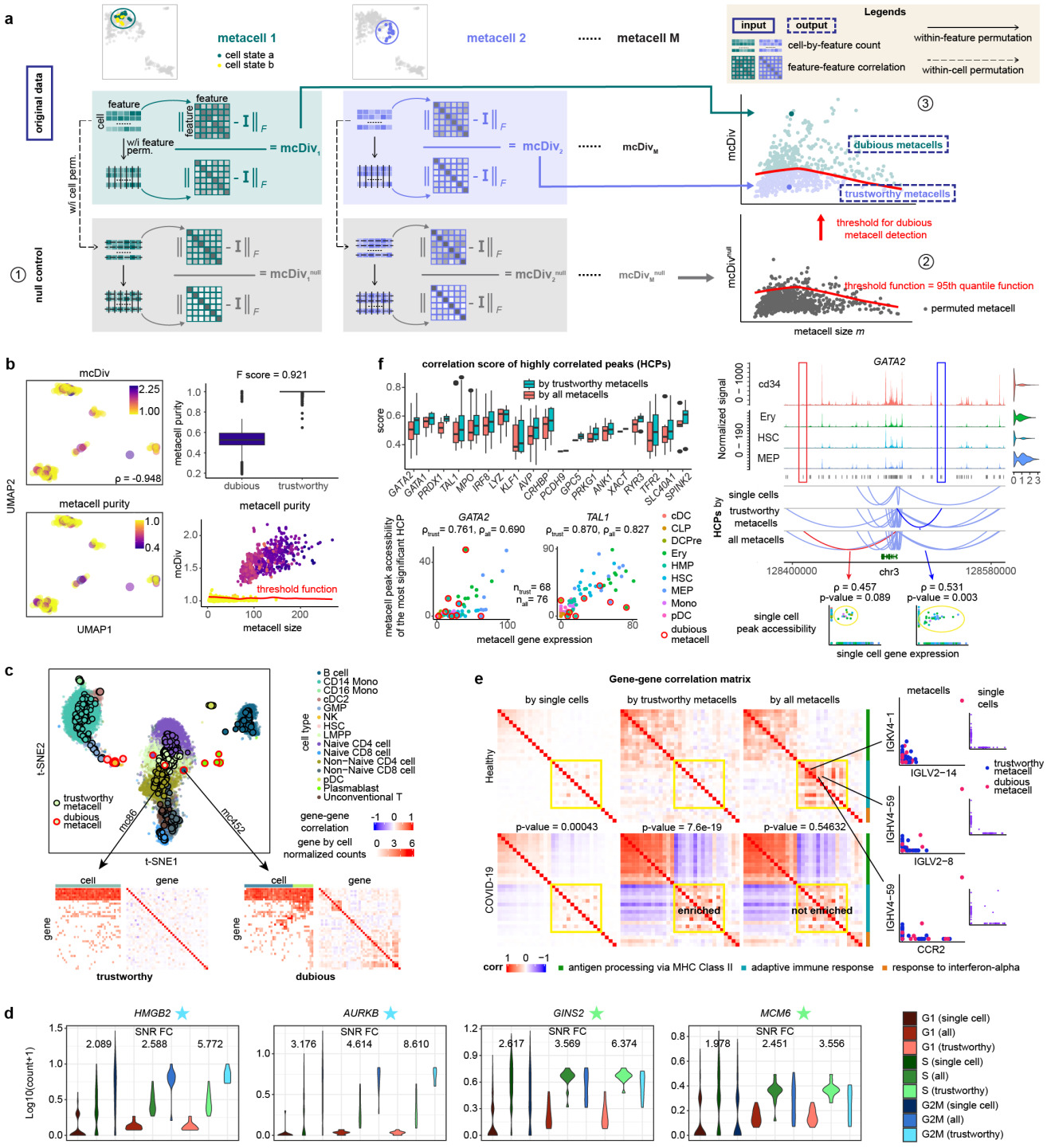
Since barcode multiplets can only approximate trustworthy metacells but not dubious ones, we next assessed mcRigor's accuracy in detecting dubious metacells. We began with a simulation where the trustworthiness of metacells was known. Using the scDesign3 simulator (Song et al., 2024), we generated a 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 (Stuart et al., 2019) (the `bmcite` dataset from the R package `SeuratData` (Hao et al., 2024)). 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 (Supplementary Fig 1b–c). By separating biological and technical variations, we were able to generate 50 ground-truth metacells, each corresponding to a distinct biological state, and then simulate single cells within each metacell to reflect technical variation, with the true granularity level set to $\gamma^* = 50$. 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 were considered truly trustworthy, while all others were 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 better than SuperCell, yielding more 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. At the true

granularity level $\gamma = 50$, MetaCell, SEACells, and SuperCell produced 0.4% 10.1%, and 28.4% dubious metacells, with dubious rates of 0.003, 0.161, and 0.453, 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 were strongly correlated with the ground-truth metacell purity values (Fig 1b, Supplementary Fig 1d–e, Supplementary Fig 2a). 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 (Supplementary Fig 1d–e), highlighting its applicability across various metacell partitioning methods.

Next, we tested mcRigor on the original bone marrow mononuclear cell CITE-seq dataset (the `bmcite` dataset from the R package `SeuratData`). As with the semi-synthetic data, we applied MetaCell, SEACells, and SuperCell, each at varying granularity levels, to this CITE-seq datast. To estimate the purity of metacells in this real dataset, 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; Mono/DC were divided into CD14 Mono, CD16 Mono, cDC2, and pDC. From the 455 metacells generated by MetaCell with granularity level $\gamma = 30$, mcRigor detected 12 dubious metacells. For visualization, we projected the metacells onto the single-cell t-SNE embedding optimized by scDEED (Xia et al., 2024). Interestingly, the dubious metacells either appeared as small clusters or lay at spurious positions (Fig 1c). For example, the centroid of metacell *mc452* fell within the cluster of LMPP and CD4 cells, while it mainly consisted of non-naive CD8 cells. This indicated that metacell *mc452* consisted of heterogeneous cells that were widely dispersed in the t-SNE embedding. The cell-by-gene expression matrix of mc452 supported this observation, revealing that the cells within it can be further divided into at least two clusters based on significant differences in gene expression (bottom right of Fig. 1c). In contrast, trustworthy metacells such as *mc86* exhibited a high level of internal homogeneity and near-zero gene-gene correlations (bottom left of Fig 1c). Other dubious metacells were mostly aggregates of Plasmablasts and HSCs, 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.



Caption on the next page.

Fig 1: mcRigor detects dubious metacells and rectifies downstream analysis for both scRNA-seq and multiome (RNA+ATAC) data. a, Schematic of the mcRigor method for dubious metacell detection. **b,** mcRigor effectively assesses metacell heterogeneity and detects dubious metacells within the MetaCell method's partitioning on semi-synthetic data. Left: UMAP plots showing partitioned metacells, colored by mcDiv values compared to metacell purity (ground truth). Right: mcRigor distinguishes between dubious and trustworthy metacells with high accuracy. **c,** Dubious metacells identified by mcRigor exhibit internal heterogeneity and may occasionally appear as outliers, while trustworthy metacells remain internally homogeneous. Bottom: Heatmaps of gene-by-cell counts and gene-by-gene correlations for a trustworthy and a dubious metacell. **d,** mcRigor enhances cell-cycle marker gene expression within cell lines. The violin plots display the $\log_{10}$(count+1) expression levels of four cell-cycle marker genes across single cells, all metacells ("all"), and trustworthy metacells ("trustworthy"). SNR FC represents the fold change in signal-to-noise ratio for the phase associated with each marker gene (indicated by a star) relative to the other two phases. **e,** mcRigor reveals enriched co-expression of an adaptive immune response gene module (highlighted in yellow) in COVID-19 samples compared to healthy controls, based on SuperCell partitions at $\gamma = 20$. Left heatmaps: Gene-gene correlation matrices for three key gene modules under COVID-19 and healthy control conditions, based on three data types: single cells, trustworthy metacells identified by mcRigor, and all metacells. Each p-value for the correlation comparisons in the adaptive immune response gene module between the two conditions for each data type was obtained using a one-sided Wilcoxon rank-sum test. Right scatter plots: Three gene pairs showing artifact correlations caused by dubious metacells, with no apparent correlations in single-cell data. **f,** Applying mcRigor to the original metacell partition from the SEACells paper empowers gene regulatory inference (left) and produces reliable discoveries (right). Left: Box plots and scatter plots showing that removing dubious metacells increases the correlations between genes and their highly correlated peaks (HCPs). Right: Removing dubious metacells uncovers a validated HCP for *GATA2* and filters out a weakly supported one, corroborated by single-cell data.

2.4 mcRigor's trustworthy metacells reveal cell-cycle phases within cell lines

As a middle ground between the single-cell and pseudobulk approaches, the metacell approach offers the advantage of revealing intra-cell-type heterogeneity, capturing biological variation among cells within the same cell type, including differences among cell subtypes or states. However, this heterogeneity can be obscured by the presence of dubious metacells, which may erroneously aggregate cells with different biological states, making these states indistinguishable. Therefore, to reveal intra-cell type heterogeneity accurately, it is essential to identify and exclude dubious metacells.

We evaluated mcRigor's ability to detect dubious metacells by testing its effectiveness in revealing cell cycle phases across five cell lines (A549, H2228, HCC827, HEK293T, and Jurkat) using scRNA-seq data. This evaluation is based on the rationale that each cell line represents a single cell type, with a major source of intra-cell-type heterogeneity arising from differences in cell-cycle phases (G1, S, and G2M). We applied MetaCell and SEACells with $\gamma = 20$ to each of the five cell lines separately to generate metacell partitions. To annotate single cells and metacells with cell-cycle phases, we used canonical gene markers to assign each single cell to a specific cell-cycle phase (Tirosh et al., 2016). Each metacell was then assigned a cell-cycle phase based on the majority phase of its constituent single cells, and the phase purity of each metacell was calculated as the highest fraction of single cells belonging to the same phase. Finally, we applied mcRigor to identify dubious metacells within each metacell partition.

We observed that mcRigor successfully distinguished between metacells with high and low phase purity. Specifically, for both metacell methods and across all five cell lines, the trustworthy metacells identified by mcRigor exhibited significantly higher phase purity than the dubious metacells (mean p-value = 0.03740 for one-sided Wilcoxon rank-sum tests across 10 comparisons (2 methods $\times$ 5 cell lines), Supplementary Fig 3). Furthermore, in the trustworthy metacells, cell-cycle marker genes exhibited less sparse expression within their corresponding phases and displayed a more distinct expression difference between the corresponding phase and the other two phases (Fig. 1d). For instance, the G2M phase marker gene *HMGB2* exhibited a fold change in the signal-to-noise (SNR) ratio (calculated as the mean expression divided by the standard deviation) of 5.773 in G2M relative to G1 and S when using

trustworthy metacells. In comparison, the fold change in SNR ratio was only 2.089 and 2.588 when using single cells and all metacells, respectively (Fig. 1d). Similar patterns were observed for other cell-cycle marker genes (Fig 1d, Supplementary Fig 3b), underscoring mcRigor's ability to reveal intra-cell-type heterogeneity.

2.5 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 quantify the relatedness between genes, which is typically measured by the pairwise correlation 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, biasing correlation estimates to an unpredictable degree (Su et al., 2023; Lu and Keleş, 2023). Metacell aggregation offers a solution to better estimate gene correlations by reducing technical noise while retaining intra-cell-type heterogeneity (Morabito et al., 2023). 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 prove 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 PBMCs of seven hospitalized COVID-19 patients and six healthy controls (Wilk et al., 2020). In this analysis, mcRigor successfully rectified co-expression estimates and uncovered gene modules differentially co-expressed between the COVID-19 and healthy cohorts. Given that co-expression patterns are often specific to cell type and condition Su et al. (2023); Morabito et al. (2023), we applied SuperCell with $\gamma = 20$ separately to the 3,028 B cells from COVID-19 samples and the 1,994 B cells from control samples. From the two metacell partitions generated by SuperCell (one per condition), mcRigor detected 22 dubious metacells among the 152 metacells identified in the COVID-19 group and 26 dubious metacells out of the 99 metacells in the control group.

We found that excluding 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 (Fig 1e), representing differential biological functions of B cells, including the antigen processing via MHC Class II gene module (p-value $= 3.7e-31$ for one-sided Wilcoxon rank-sum test), the adaptive immune response gene module (p-value $= 7.6e-19$), and the response to interferon-alpha gene module (p-value $= 0.00328$). These enrichment signals were notably strengthened compared to those in the raw single-cell data, where the p-values for these three gene modules are $2.0e-13$, 0.00043, and 0.08441, respectively. We also observed that these enrichment findings are consistent with those reported by the CS-CORE method (Su et al., 2023).

In contrast, correlation estimation that included dubious metacells (using all metacells) yielded the counterintuitive result that the adaptive immune response gene module was not enriched in COVID-19 patients (p-value $= 0.54632$). We found that this misleading result was caused by an artifact: the strong co-expression of the adaptive immune response gene module under the healthy condition. Notably, this artifact was absent at single-cell resolution and was induced solely by the presence of dubious metacells (Fig 1e). Applying mcRigor to the partitions generated by SEACells, MetaCell, and MetaCell2 similarly removed the artifact correlations introduced by dubious metacells and enhanced enrichment signals (Supplementary Fig 4), demonstrating the applicability and robustness of mcRigor across various metacell methods.

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

Single-cell multiome data or integrated scATAC-seq and scRNA-seq data (Ma et al., 2020; Granja et al., 2021; Trevino et al., 2021), which provide both gene expression and chromatin accessibility modalities in the same cells (either through direct measurement or computational integration), offer a lens to study the association between these two modalities. This approach reveals relationships between genes and regulatory elements (e.g., enhancers) with finer resolution, such as cell-type specificity, that is unattainable 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 the chromatin accessibility modality, 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 (Persad et al., 2023; Mao et al., 2022).

Persad et al. (2023) demonstrated the effectiveness of metacell aggregation for empowering gene regulatory inference by applying their metacell partitioning method, SEACells, to a single-cell multiome dataset, which consists of 6,881 hematopoietic stem and progenitor cells (HSPCs) from healthy bone marrow sorted for the pan-HSPC marker CD34. Compared to single-cell-based analysis, using metacells (the so called "SEACells" 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 (where 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 metacell partition generated by SEACells in the original study (Persad et al., 2023), mcRigor identified 7 dubious metacells out of the 85 metacells (Supplementary Fig 5a). 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 1f). 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, increased from 0.6904 to 0.7606 (bottom left of Fig 1f).

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, although using all metacells and using trustworthy metacells both identified 16 highly correlated peaks (HCPs) for the gene *GATA2*, the specific peaks identified differed between the two approaches. 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 (Fishilevich et al., 2017; LOC117038772 GeneCards, 2016). The peak's correlation with *GATA2* expression was also evident at the single-cell level (bottom right of Fig 1f). In contrast, the peak not detected when using trustworthy metacells, *chr3-128409363-128409863*, showed a 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 1f, Supplementary Fig 5b). These findings confirm that mcRigor enhances regulatory analysis and, importantly, helps pinpoint likely false positives.

2.7 Assessment of mcRigor's ability to 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 a greater reduction (Bilous et al., 2024). If γ is too large, heterogeneous single cells may be aggregated 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 (Fig 2a). 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* (with the default weight set as 0.5), where a higher *Score* indicates a better metacell partition. The γ value that maximizes this score is considered to balance the preservation of biological signals and the improvement of data sparsity.

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 2a). Additionally, the optimized metacell partition closely matched the ground-truth metacells defined in the simulation, containing only 4 dubious metacells (Fig 2a).

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^* (Supplementary Fig 6). 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 (Supplementary Fig 6).

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 (Supplementary Fig 6). Thus, the optimal configuration selected by mcRigor is MetaCell with $\gamma = 50$, which best matched the ground truth. The higher maximal *Score* values for MetaCell and SEACells compared to SuperCell and MetaCell2 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 MetaCell and SEACells 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.

2.8 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 Jiang et al. (2022). Ideally, biological zeros should be retained as they provide

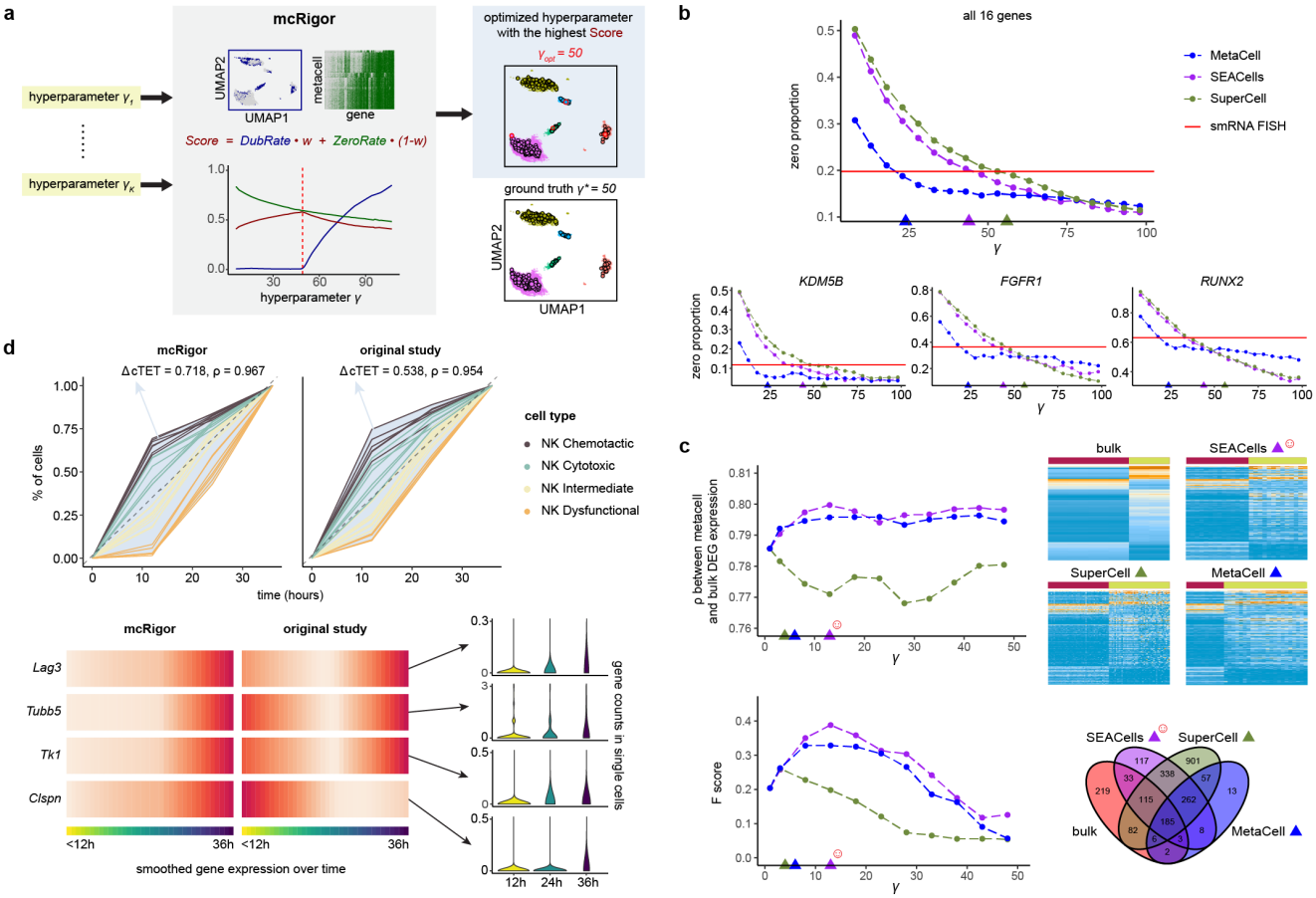


Fig 2: mcRigor optimizes metacell method and hyperparameter selection for various single-cell data analyses. a, Schematic of the mcRigor method for optimizing metacell partitioning, using *Score* as the optimization criterion to balance *DubRate* and *ZeroRate*, illustrated with the optimization of MetaCell partitions on semi-synthetic data as an example. **b,**. Line plots showing the zero proportions in metacell partitions generated by three methods (MetaCell, SEACells, and SuperCell) across varying granularity levels (γ). The optimized metacell partitions (triangles) closely align with the zero proportion observed in smRNA FISH data (red line). **c,** mcRigor optimizes metacell method and hyperparameter selection for differential gene expression analysis. Top: A line plot and heatmaps comparing the expression of the top 200 bulk DE genes across various metacell partitions to their expression in bulk data. The line plot depicts the concordance between the bulk and metacell profiles. Bottom: A line plot showing F-scores and a Venn diagram comparing the DE genes identified from various metacell partitions with those detected from bulk data. The optimal metacell partition (SEACells with $\gamma = 13$, marked by the red smiling face) achieving the highest concordance (Pearson correlation $\rho = 0.800$) and F-score (0.400). In **b** and **c**, the colored triangles indicate the optimal γ values selected by mcRigor for the three methods. **d,** mcRigor's optimized metacell partition better reveals temporal immune cell trajectories compared to the original metacell partition from the Zman-seq study (Kirschenbaum et al., 2024). Top: Line plots comparing the metacells' continuous tumor exposure time (cTET) values calculated from mcRigor's optimized partition to those from the original partition. The value of ΔcTET, proportional to the size of the light blue area, indicates the distinction of tumor transitional stages. Bottom: Smoothed gene expression profiles of four marker genes in mcRigor's optimized partition compared to the original partition. Following optimization, the genes' expression patterns align more closely with biological expectations.

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 (Torre et al., 2018). The rationale for this approach is that smRNA FISH is widely considered the gold standard for single-cell gene expression measurement (Torre et al., 2018; Grün et al., 2014), 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 (Supplementary Fig 7a–b). 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 2b, Supplementary Fig 7c). Compared to single-cell data, the expression distributions of each gene derived from the optimal metacell partitions (one per metacell method) closely approximated the corresponding distribution observed in the smRNA FISH data (Supplementary Fig 7d). These result confirm that mcRigor can help distinguish biological zeros from non-biological zeros by optimizing metacell partitioning.

2.9 mcRigor optimizes metacell partitioning for differential gene expression analysis

Differential gene expression (DGE) analysis on scRNA-seq data often suffers from data sparsity, leading to reduced statistical power. 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 differentially expressed (DE) genes 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 (Li and Li, 2018). 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) (Chu et al., 2016).

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 (Supplementary Fig 8). As the first validation step, we used the DESeq2 method (Love et al., 2014) to identify 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 2c). 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) or the metacell partition (average of metacells).

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 standard, 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 $\gamma = 13$, Supplementary Fig 8) 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 2c). These findings indicate that the DE genes identified using the optimal metacell partition had the best agreement with the bulk DE genes, thereby indirectly affirming mcRigor's optimization of metacell partitioning, including the selection of the metacell method and its hyperparameter value.

2.10 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 (Kirschenbaum et al., 2024). 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 (top right of Fig 2d), 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 (Supplementary Fig 9a–b). 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 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 2d). 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 (Supplementary Fig 9c). 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 2d). For instance, *Lag3*, a gene associated with the suppression of antitumor functions (Gandhi et al., 2006; Matsuzaki et al., 2010) and known to act as a receptor for the ligand LSECtin (Xu et al., 2014), 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 (Dean et al., 2024). 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 (Chen et al., 2023; Jia et al., 2022). 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 (Supplementary Fig 9d), including the upregulated module (*Pmepa1*, *Car2*, *Ctla2a*, *Xcl1*, *Itga1*, *Gzmc*, etc.) and the downregulated module (*S1pr5*, *Cx3cr1*, *Klrg1*, *Cma1*, *Ccl3*, *Gzmb*, etc.).

3 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. However, mcRigor has some limitations and open questions that warrant further exploration. 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.

An additional note is that doublets (multiplets) in the single-cell droplet-based sequencing data should be removed before implementing metacell partitioning and mcRigor. Here, a doublet (multiplet) refers to two (more) cells being encapsulated within the same droplet during the sequencing process, resulting in them being tagged with the same barcode and mistakenly identified as a single cell (Xi and Li, 2021a). These doublets (multiplets) should not be used to construct metacells, as the cells within a doublet (or multiplet) are highly unlikely to originate from the same biological state. Therefore, we recommend applying a doublet-detection method (Xi and Li, 2021b) before any metacell partitioning steps.

Regarding the trustworthiness of metacells, an intuitive belief is that larger metacells are more likely to be dubious. While we observe that metacells exhibit substantial variability in size at the same granularity level (Fig 1c, Supplementary Fig 5a, Supplementary Fig 9b), no clear relationship exists between metacell size and trustworthiness (as determined by mcRigor). This finding underscores that dubious metacells cannot be reliably identified based solely on size. Moreover, metacell size distributions differ significantly across cell types and conditions (Supplementary Fig.11, Supplementary Fig.12). Notably,

smaller metacells are more frequent in unstable biological states, such as in COVID-19 PBMCs compared to healthy controls (Supplementary Fig. 11a). For another example, in the `bmcite` dataset, progenitor cells are generally grouped into smaller metacells, whereas T cells are aggregated into larger metacells (Supplementary Fig.12a). Hence, if we have a good understanding of the relative stability of biological conditions, metacell size distributions can serve as a quick sanity check for the quality of the metacell partitioning before applying mcRigor.

4 Methods

4.1 A statistical definition of metacells

As first introduced by Baran et al. (2019), a *metacell* is defined as a homogeneous collection of single-cell profiles that could have been resampled from the same original cell. This means that within a metacell, the single cells should be in the same biological state, exhibiting the same expected expression levels of features (e.g., genes or chromatin regions) at equilibrium, referred to as *biological signals*. Any differences in the observed counts (of sequencing reads or unique molecular identifiers) among these single cells should therefore be attributed to measurement imperfections, known as *technical variations*. This definition ensures that aggregating these single cells into a metacell by averaging can reduce technical variations while preserving biological signals.

We formalize this metacell definition statistically by following the *observation model* for single-cell sequencing data described in (Sarkar and Stephens, 2021). In this model, the variations in observed counts 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, including sequencing, that obscures these biological signals. Hence, the observation model is hierarchical, consisting of an *expression model*, which describes the distribution of biological signals across cells (i.e., the biological variation), and a *measurement model*, which describes the distribution of observed counts given these biological 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 $c_i = \mathrm{E}[y_{i+}|\boldsymbol{\lambda}_i]$. The measurement model reflects the technical variation among cells.

Under this hierarchical observation model and following the definition by Baran et al. (2019), we introduce a **statistical definition**: *a metacell is a group of cells that share the same* $\boldsymbol{\lambda}$. We will show in the 'Justification for the statistical definition of metacells' Section 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 are referred to as **dubious**

metacells.

4.2 The mcRigor algorithm

4.2.1 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 'Justification for the statistical definition of metacells' Section, we prove that under the measurement model, features are nearly pairwise uncorrelated, with only minimal correlations induced by the constraint $\sum_{j=1}^{p} \lambda_j = 1$, and the correlations decrease as p increases and become negligible when p is large. Guided by this premise, mcRigor detects dubious metacells by examining the $p \times p$ feature correlation matrix of each metacell, via a statistical test for each metacell 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 that reflects deviation from the null hypothesis H_0. 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 $\mathbf{R}_k \in [-1, 1]^{p \times p}$. The deviation of $\mathbf{R}_k$ from the identity matrix $\mathbf{I} \in \mathbb{R}^{p \times p}$ is then calculated as the Frobenius norm $\|\mathbf{R}_k - \mathbf{I}\|_F$. 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 $\widetilde{\mathbf{R}}_k$. The *metacell divergence score*, called "mcDiv," is then defined as

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

where dividing $\|\mathbf{R}_k - \mathbf{I}\|_F$ by $\|\widetilde{\mathbf{R}}_k - \mathbf{I}\|_F$ 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, so as to determine 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 $\mathbf{\Pi}_k$ denote the correlation matrix of this within-cell permuted data matrix. Following the same procedure as in Step 1, within-feature permutation is applied to the within-cell permutated data matrix, resulting in $\widetilde{\mathbf{\Pi}}_k$, 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

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

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. Notably, we cannot simply compare the numerators $\|\mathbf{R}_k - \mathbf{I}\|_F$ and $\|\mathbf{\Pi}_k - \mathbf{I}\|_F$ while ignoring the denominator in the metacell divergence score definition. This is because $\mathbf{R}_k$ and $\mathbf{\Pi}_k$ are based on different sets of features (since within-cell permutation does not preserve features) and are therefore not directly comparable. Furthermore, the denominator $\|\widetilde{\mathbf{R}}_k - \mathbf{I}\|_F$, which is derived from within-feature permutation, is not a valid null for the numerator $\|\mathbf{R}_k - \mathbf{I}\|_F$, as the cell library sizes are not preserved in the calculation of $\widetilde{\mathbf{R}}_k$. We will present a result to demonstrate that within-feature permutation is inappropriate in Note 1 (below Step 4).

From the M null divergence scores, $\text{mcDiv}_1^{\text{null}}, \ldots, \text{mcDiv}_M^{\text{null}}$, 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 $\text{mcDiv}_1^{\text{null}}, \ldots, \text{mcDiv}_M^{\text{null}}$ 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 (Supplementary Fig 2b). 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:

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

where $\text{q}_{0.95}(\cdot)$ 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 $\text{mcDiv}_k > \theta(m_k)$, it is classified as dubious; otherwise, it is considered trustworthy.

Note 1. If we use the within-feature permutation instead of the double permutation to find dubious metacell detection thresholds, i.e., the kth metacell is considered dubious if $\|\mathbf{R}_k - \mathbf{I}\|_F > \text{q}_{0.95}\left(\left\{\|\widetilde{\mathbf{R}}_{k'} - \mathbf{I}\|_F : m_{k'} \in [m_k - h, m_k + h],\, k' = 1, \ldots, M\right\}\right)$, this approach sets the threshold too low, leading to the false flagging of trustworthy metacells as dubious. This issue arises because the magnitude of $\|\mathbf{R}_k - \mathbf{I}\|_F$ may simply reflect variations in cell library sizes within the kth metacell. However, such variations are not preserved in the within-feature permutation, resulting in a smaller magnitude of $\|\widetilde{\mathbf{R}_k} - \mathbf{I}\|_F$ compared to $\|\mathbf{R}_k - \mathbf{I}\|_F$. In a simulation study, we demonstrate that more than 35% of ground-truth trustworthy metacells—comprising single cells of the same cell state (according to our statistical definition)—are incorrectly classified as dubious, leading to poor overall classification perfor-

mance (F-score < 0.4) in distinguishing between trustworthy and dubious metacells (Supplementary Fig 10). This is in sharp contrast to the performance of mcRigor based on the double permutation: of all the ground-truth trustworthy metacells, mcRigor accurately identifies over 98% as trustworthy, with less than 2% incorrectly classified as dubious (Supplementary Fig 10). 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 (Chen et al., 2010; Cai and Ma, 2013; Cai, 2017), 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.

4.2.2 Optimization of metacell partitioning: method and hyperparameter choices

Employing the dubious metacell detection procedure outlined in the previous subsection, mcRigor further assists users in optimizing metacell partitioning by simultaneously identifying the best-performing metacell method and the optimal hyperparameter for a particular single-cell dataset. In this work, we primarily focus on optimizing the *granularity level* hyperparameter, γ, which means the average number of single cells per metacell and is universally required by all existing metacell partitioning methods (some methods use the hyperparameter *target metacell number* instead, which is equal to $\lfloor n/\gamma \rfloor$, with n as the number of single cells). While other hyperparameters—such as the number of neighbors (k) in kNN graph construction and the number of principal components used—can also be optimized by mcRigor, we omit their discussion here. We choose to optimize γ, as these other parameters have been reported to have a smaller impact on metacell partitioning and are not universally required by metacell methods (Bilous et al., 2024).

As discussed in the Results Section, the optimization of metacell partitioning is, at its core, about balancing between data sparsity and signal distortion, the latter being addressed by mcRigor's first functionality: the detection of dubious metacells. Intuitively, a smaller γ value results in fewer dubious metacells and less signal distortion, as aggregating a smaller number of cells lowers the chance of merging cells of different states. However, the minimal value of $\gamma = 1$, which retains single cells without aggregating them into metacells and therefore avoids signal distortion, is not always desirable, as it leaves the sparsity issue unresolved. Hence, an ideal metacell partitioning should aggregate as many cells as possible, while ensuring that only homogeneous cells are combined. To strike this balance, mcRigor assesses the level of sparsity and signal distortion for each metacell partition, enabling a quantitative tradeoff.

Specifically, for a given metacell method and hyperparameter, mcRigor evaluates two competing factors that respectively reflect signal distortion and data sparsity (Fig 2a). 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 metacell partition:

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

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 this definition (7), a higher evaluation score indicates a better metacell

partition.

The evaluation score allows for universal comparison of metacell partitions across varying hyperparameter values of 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.

4.3 Justification for the statistical definition of metacells

A typical way of defining the metacell-by-feature expression matrix, $\mathbf{Z} = [z_{kj}] \in \mathbb{R}^{M \times p}$, is averaging the assigned single cell profiles, namely defining

$$z_{kj} = \frac{1}{m_k} \sum_{i \in \mathrm{mc}_k} \psi(y_{ij}/y_{i+}) \tag{8}$$

where mc_k denotes the index set of single cells assigned to the kth metacell ($|\mathrm{mc}_k| = m_k$), and $\psi(\cdot)$ is a function specified for a particular normalization method. For ease of understanding, we consider the simple case where $\psi(x) = x$.

If the kth metacell satisfies our statistical definition, the single cells assigned to the kth metacell share the same relative expression levels $\boldsymbol{\lambda}_k = (\lambda_{k1}, \ldots, \lambda_{kp})^\top$. For cell i in the kth metacell, we have discussed in (3) that conditional on cell library size y_{i+} and relative expression level $\boldsymbol{\lambda}_k$, the observed feature counts follow a multinomial distribution:

$$H_0: \ (y_{i1}, \ldots, y_{ip}) | \, \boldsymbol{\lambda}_k, y_{i+} \overset{\text{ind}}{\sim} \mathrm{Mult}(y_{i+}, \lambda_{k1}, \ldots, \lambda_{kp}) \, .$$

By the property of the multinomial distribution, we have

$$\begin{aligned} \mathrm{E}\left[z_{kj} | \boldsymbol{\lambda}_k, \mathbf{y}_{k+}\right] &= \frac{1}{m_k} \sum_{i \in \mathrm{mc}_k} \frac{\mathrm{E}\left[y_{ij} | \boldsymbol{\lambda}_k, \mathbf{y}_{k+}\right]}{y_{i+}} = \lambda_{kj} \, , \\ \mathrm{Var}\left[z_{kj} | \boldsymbol{\lambda}_k, \mathbf{y}_{k+}\right] &= \frac{1}{m_k^2} \sum_{i \in \mathrm{mc}_k} \mathrm{Var}\left[\frac{y_{ij}}{y_{i+}} \middle| \boldsymbol{\lambda}_k, \mathbf{y}_{k+}\right] < \frac{1}{m_k} \cdot \max_{i \in \mathrm{mc}_k} \mathrm{Var}\left[\frac{y_{ij}}{y_{i+}} \middle| \boldsymbol{\lambda}_k, \mathbf{y}_{k+}\right] < \max_{i \in \mathrm{mc}_k} \mathrm{Var}\left[\frac{y_{ij}}{y_{i+}} \middle| \boldsymbol{\lambda}_k, \mathbf{y}_{k+}\right] \, , \end{aligned} \tag{9}$$

where $\mathbf{y}_{k+} = (y_{i+})_{i \in \mathrm{mc}_k}$. This indicates that the obtained metacell profile provides an unbiased estimate of the true expression level with reduced variance.

Moreover, the sample covariance within the kth metacell between any pair of features j and ℓ is given by

$$\widehat{\Sigma}_{k,j\ell} = \frac{1}{m_k - 1} \sum_{i \in \mathrm{mc}_k} \frac{y_{ij}}{y_{i+}} \frac{y_{i\ell}}{y_{i+}} - \frac{1}{m_k(m_k - 1)} \sum_{i \in \mathrm{mc}_k} \frac{y_{ij}}{y_{i+}} \sum_{i \in \mathrm{mc}_k} \frac{y_{i\ell}}{y_{i+}} \, , \tag{10}$$

whose expectation is the conditional covariance between features j and ℓ given $\boldsymbol{\lambda}_k$ and $\mathbf{y}_{k+}$:

$$
\begin{aligned}
&\Sigma_{k,j\ell|\boldsymbol{\lambda}_k,\mathbf{y}_{k+}} = \mathrm{E}\left[\widehat{\Sigma}_{k,j\ell}\middle|\boldsymbol{\lambda}_k,\mathbf{y}_{k+}\right] \\
&= \frac{1}{m_k-1}\sum_{i\in\mathrm{mc}_k}\mathrm{E}\left[\frac{y_{ij}}{y_{i+}}\frac{y_{i\ell}}{y_{i+}}\middle|\boldsymbol{\lambda}_k,\mathbf{y}_{k+}\right] - \frac{1}{m_k(m_k-1)}\mathrm{E}\left[\sum_{i\in\mathrm{mc}_k}\frac{y_{ij}}{y_{i+}}\sum_{i\in\mathrm{mc}_k}\frac{y_{i\ell}}{y_{i+}}\middle|\boldsymbol{\lambda}_k,\mathbf{y}_{k+}\right] \\
&= \frac{1}{m_k-1}\sum_{i\in\mathrm{mc}_k}\mathrm{E}\left[\frac{y_{ij}}{y_{i+}}\middle|\boldsymbol{\lambda}_k,\mathbf{y}_{k+}\right]\mathrm{E}\left[\frac{y_{i\ell}}{y_{i+}}\middle|\boldsymbol{\lambda}_k,\mathbf{y}_{k+}\right] + \frac{1}{m_k-1}\sum_{i\in\mathrm{mc}_k}\mathrm{Cov}\left(\frac{y_{ij}}{y_{i+}},\frac{y_{i\ell}}{y_{i+}}\middle|\boldsymbol{\lambda}_k,\mathbf{y}_{k+}\right) \\
&\quad - \frac{1}{m_k(m_k-1)}\sum_{i\in\mathrm{mc}_k}\mathrm{E}\left[\frac{y_{ij}}{y_{i+}}\middle|\boldsymbol{\lambda}_k,\mathbf{y}_{k+}\right]\sum_{i\in\mathrm{mc}_k}\mathrm{E}\left[\frac{y_{i\ell}}{y_{i+}}\middle|\boldsymbol{\lambda}_k,\mathbf{y}_{k+}\right] - \frac{1}{m_k(m_k-1)}\sum_{i\in\mathrm{mc}_k}\mathrm{Cov}\left(\frac{y_{ij}}{y_{i+}},\frac{y_{i\ell}}{y_{i+}}\middle|\boldsymbol{\lambda}_k,\mathbf{y}_{k+}\right) \\
&= \frac{1}{m_k-1}m_k\lambda_{kj}\lambda_{k\ell} - \frac{1}{m_k-1}\sum_{i\in\mathrm{mc}_k}\frac{\lambda_{kj}\lambda_{k\ell}}{y_{i+}} - \frac{1}{m_k(m_k-1)}m_k\lambda_{kj}m_k\lambda_{k\ell} + \frac{1}{m_k(m_k-1)}\sum_{i\in\mathrm{mc}_k}\frac{\lambda_{kj}\lambda_{k\ell}}{y_{i+}} \\
&= -\frac{\lambda_{kj}\lambda_{k\ell}}{m_k}\sum_{i\in\mathrm{mc}_k}\frac{1}{y_{i+}}\,. \qquad (11)
\end{aligned}
$$

Similarly, the sample variance of feature j within the kth metacell is given by

$$
\widehat{\Sigma}_{k,jj} = \frac{1}{m_k-1}\sum_{i\in\mathrm{mc}_k}\left(\frac{y_{ij}}{y_{i+}}\right)^2 - \frac{1}{m_k(m_k-1)}\left(\sum_{i\in\mathrm{mc}_k}\frac{y_{ij}}{y_{i+}}\right)^2, \qquad (12)
$$

whose expectation is the conditional variance of feature j given $\boldsymbol{\lambda}_k$ and $\mathbf{y}_{k+}$.

$$
\Sigma_{k,jj|\boldsymbol{\lambda}_k,\mathbf{y}_{k+}} = \mathrm{E}\left[\widehat{\Sigma}_{k,jj}\middle|\boldsymbol{\lambda}_k,\mathbf{y}_{k+}\right] = \frac{\lambda_{kj}(1-\lambda_{kj})}{m_k}\sum_{i\in\mathrm{mc}_k}\frac{1}{y_{i+}}\,. \qquad (13)
$$

Thus, correspondingly, the conditional correlation within the kth metacell between features j and ℓ is

$$
\mathbf{R}_{k,j\ell|\boldsymbol{\lambda}_k,\mathbf{y}_{k+}} = \frac{\Sigma_{k,j\ell|\boldsymbol{\lambda}_k,\mathbf{y}_{k+}}}{\sqrt{\Sigma_{k,jj|\boldsymbol{\lambda}_k,\mathbf{y}_{k+}}\Sigma_{k,\ell\ell|\boldsymbol{\lambda}_k,\mathbf{y}_{k+}}}} = -\sqrt{\frac{\lambda_{kj}}{1-\lambda_{kj}}}\sqrt{\frac{\lambda_{k\ell}}{1-\lambda_{k\ell}}}\,, \qquad (14)
$$

whose absolute value is under $\max_{j=1,\ldots,p}\frac{\lambda_{kj}}{1-\lambda_{kj}}$, a negligible value when p is sufficiently large.

4.3.1 Justification for why dubious metacells bias co-expression analysis

If some metacells appear as a mixed collection of single cells in different biological states, the aggregated expressions of those metacells are most likely biased, and inflated feautre correlations may be observed within the metacells. What makes things worse is that such distortion could be carried to the inference of feature co-expression using metacells. Let us consider a simple case for cell-type-specific co-expression analysis where the single cells are of one cell type but belong to N distinct cell states of balanced sizes. Let cs_k denote the single cell index set of the kth cell state and $\boldsymbol{\lambda}_k^* = (\lambda_{k1}^*,\ldots,\lambda_{kp}^*)^\top$ denote the corresponding cell-state feature expression levels.

Suppose we want to estimate the covariance between features j and ℓ, $\Omega_{j\ell}$, from the metacell-by-feature expression matrix $\mathbf{Z}$. If metacells are correctly identified, i.e., $M = N$ and $\mathrm{mc}_k = \mathrm{cs}_k$ for

$k = 1, \ldots, M$,

$$\Omega_{j\ell} = \frac{1}{M-1}\sum_{k=1}^{M}\lambda_{kj}^{*}\lambda_{k\ell}^{*} - \frac{1}{M(M-1)}\sum_{k=1}^{M}\lambda_{kj}^{*}\sum_{k=1}^{M}\lambda_{k\ell}^{*}\,, \tag{15}$$

and the estimate is

$$\widehat{\Omega}_{j\ell} = \frac{1}{M-1}\sum_{k=1}^{M}z_{kj}z_{k\ell} - \frac{1}{M(M-1)}\sum_{k=1}^{M}z_{kj}\sum_{k=1}^{M}z_{k\ell}\,. \tag{16}$$

This estimate is asymptotically unbiased since

$$\begin{aligned}
&\mathrm{E}\left[\widehat{\Omega}_{j\ell}\right] = \mathrm{E}\left[\mathrm{E}\left[\widehat{\Omega}_{j\ell}\middle|\,\boldsymbol{\lambda}_1^{*},\ldots,\boldsymbol{\lambda}_M^{*},\mathbf{y}_{1+},\ldots,\mathbf{y}_{M+}\right]\right]\\
&= \mathrm{E}\left[\frac{1}{M-1}\sum_{k=1}^{M}\mathrm{E}\left[z_{kj}|\,\boldsymbol{\lambda}_k^{*},\mathbf{y}_{k+}\right]\mathrm{E}\left[z_{k\ell}|\,\boldsymbol{\lambda}_k^{*},\mathbf{y}_{k+}\right] + \frac{1}{M-1}\sum_{k=1}^{M}\mathrm{Cov}\left(z_{kj},z_{k\ell}|\,\boldsymbol{\lambda}_k^{*},\mathbf{y}_{k+}\right)\right]\\
&\quad - \mathrm{E}\left[\frac{1}{M(M-1)}\sum_{k=1}^{M}\mathrm{E}\left[z_{kj}|\,\boldsymbol{\lambda}_k^{*},\mathbf{y}_{k+}\right]\sum_{k=1}^{M}\mathrm{E}\left[z_{k\ell}|\,\boldsymbol{\lambda}_k^{*},\mathbf{y}_{k+}\right] + \frac{1}{M(M-1)}\sum_{k=1}^{M}\mathrm{Cov}\left(z_{kj},z_{k\ell}|\,\boldsymbol{\lambda}_k^{*},\mathbf{y}_{k+}\right)\right]\\
&\overset{(*)}{=} \mathrm{E}\left[\frac{1}{M-1}\sum_{k=1}^{M}\mathrm{E}\left[z_{kj}|\,\boldsymbol{\lambda}_k^{*},\mathbf{y}_{k+}\right]\mathrm{E}\left[z_{k\ell}|\,\boldsymbol{\lambda}_k^{*},\mathbf{y}_{k+}\right] - \frac{1}{M(M-1)}\sum_{k=1}^{M}\mathrm{E}\left[z_{kj}|\,\boldsymbol{\lambda}_k^{*},\mathbf{y}_{k+}\right]\sum_{k=1}^{M}\mathrm{E}\left[z_{k\ell}|\,\boldsymbol{\lambda}_k^{*},\mathbf{y}_{k+}\right]\right]\\
&\quad + o\left(\frac{1}{Mp^2}\right)\\
&= \mathrm{E}\left[\frac{1}{M-1}\sum_{k=1}^{M}\lambda_{kj}^{*}\lambda_{k\ell}^{*} - \frac{1}{M(M-1)}\sum_{k=1}^{M}\lambda_{kj}^{*}\sum_{k=1}^{M}\lambda_{k\ell}^{*}\right] + o\left(\frac{1}{Mp^2}\right)\\
&= \Omega_{j\ell} + o\left(\frac{1}{Mp^2}\right),
\end{aligned} \tag{17}$$

where $\mathbf{y}_{k+} = (y_{i+})_{i\in\mathrm{mc}_k}$ and the third equality $(*)$ holds under the assumptions $\max_{j=1,\ldots,p}\lambda_{kj} = O(1/p)$, indicating that no single feature dominates, and $M = o(m_k \cdot \min_{i\in\mathrm{mc}_k} y_{i+})$, indicating that the number of metacells is of a lower order compared to the total number of feature counts within a metacell. We consider these two assumptions reasonable based on empirical observations. Then we have

$$\begin{aligned}
&\left|\mathrm{E}\left[\frac{1}{M-1}\sum_{k=1}^{M}\mathrm{Cov}\left(z_{kj},z_{k\ell}|\,\boldsymbol{\lambda}_k^{*},\mathbf{y}_{k+}\right) - \frac{1}{M(M-1)}\sum_{k=1}^{M}\mathrm{Cov}\left(z_{kj},z_{k\ell}|\,\boldsymbol{\lambda}_k^{*},\mathbf{y}_{k+}\right)\right]\right|\\
&= \left|\mathrm{E}\left[\frac{1}{M}\sum_{k=1}^{M}\mathrm{Cov}\left(z_{kj},z_{k\ell}|\,\boldsymbol{\lambda}_k^{*},\mathbf{y}_{k+}\right)\right]\right| = \left|\mathrm{E}\left[\frac{1}{M}\sum_{k=1}^{M}\frac{1}{m_k^2}\sum_{i\in\mathrm{mc}_k}\mathrm{Cov}\left(\frac{y_{ij}}{y_{i+}},\frac{y_{i\ell}}{y_{i+}}\middle|\,\boldsymbol{\lambda}_k^{*},\mathbf{y}_{k+}\right)\right]\right|\\
&= \left|\mathrm{E}\left[-\frac{1}{M}\sum_{k=1}^{M}\frac{1}{m_k^2}\sum_{i\in\mathrm{mc}_k}\frac{\lambda_{kj}\lambda_{k\ell}}{y_{i+}}\right]\right| \le \mathrm{E}\left[\frac{1}{M}\sum_{k=1}^{M}\frac{1}{m_k}\cdot\frac{\max_{j=1,\ldots,p}\lambda_{kj}^2}{\min_{i\in\mathrm{mc}_k} y_{i+}}\right] = o\left(\frac{1}{Mp^2}\right).
\end{aligned} \tag{18}$$

However, this asymptotic unbiasedness may not hold if single cells of different states are grouped into a metacell, creating a dubious metacell. For example, let us consider a scenario where the first two metacells are both equal mixtures of two distinct cell states. Let A_1 represent the index set of single cells belonging to the first cell state and assigned to the first metacell, and B_1 represent the index set of single cells belonging to the second cell state and assigned to the first metacell. Namely, we assume

$\mathrm{mc}_1 = A_1 \cup B_1$ and $\mathrm{mc}_2 = \mathrm{cs}_1 \cup \mathrm{cs}_2 \setminus \mathrm{mc}_1$, where $A_1 \subset \mathrm{cs}_1$, $B_1 \subset \mathrm{cs}_2$, and $|A_1| = |B_1| = m_1/2$, $|\mathrm{cs}_1| = |\mathrm{cs}_2| = (m_1 + m_2)/2$. Then we have

$$\begin{aligned} z_{1j} &= \frac{1}{m_1} \sum_{i \in A_1} \frac{y_{ij}}{y_{i+}} + \frac{1}{m_1} \sum_{i \in B_1} \frac{y_{ij}}{y_{i+}}, \\ z_{2j} &= \frac{1}{m_2} \sum_{i \in \mathrm{cs}_1 \setminus A_1} \frac{y_{ij}}{y_{i+}} + \frac{1}{m_2} \sum_{i \in \mathrm{cs}_2 \setminus B_1} \frac{y_{ij}}{y_{i+}}, \end{aligned} \tag{19}$$

whose conditional expectations are $\mathrm{E}\left[z_{1j} | \boldsymbol{\lambda}_1^*, \boldsymbol{\lambda}_2^*, \mathbf{y}_{1+}\right] = \mathrm{E}\left[z_{2j} | \boldsymbol{\lambda}_1^*, \boldsymbol{\lambda}_2^*, \mathbf{y}_{2+}\right] = \left(\lambda_{1j}^* + \lambda_{2j}^*\right)/2$, and similar derivation as in (17) gives

$$\begin{aligned} \mathrm{E}\left[\widehat{\Omega}_{j\ell}\right] &\overset{(*)}{=} \mathrm{E}\left[\frac{1}{M-1}\left(2 \cdot \frac{\lambda_{1j}^* + \lambda_{2j}^*}{2} \frac{\lambda_{1\ell}^* + \lambda_{2\ell}^*}{2} + \sum_{k=3}^{M} \lambda_{kj}^* \lambda_{k\ell}^*\right) - \frac{1}{M(M-1)} \sum_{k=1}^{M} \lambda_{kj}^* \sum_{k=1}^{M} \lambda_{k\ell}^*\right] + o\left(\frac{1}{Mp^2}\right) \\ &= \Omega_{j\ell} + \frac{1}{2(M-1)} \mathrm{E}\left[\lambda_{1j}^* \lambda_{2\ell}^* + \lambda_{2j}^* \lambda_{1\ell}^* - \lambda_{1j}^* \lambda_{1\ell}^* - \lambda_{2j}^* \lambda_{2\ell}^*\right] + o\left(\frac{1}{Mp^2}\right). \end{aligned} \tag{20}$$

This indicates that the covariance estimate is biased, and the bias $\frac{1}{2(M-1)} \mathrm{E}\left[\lambda_{1j}^* \lambda_{2\ell}^* + \lambda_{2j}^* \lambda_{1\ell}^* - \lambda_{1j}^* \lambda_{1\ell}^* - \lambda_{2j}^* \lambda_{2\ell}^*\right]$ is non-negligible if the number of dubious metacells is in the same order of M (so the bias does not go down to zero as M becomes large). An extreme example is that the covariance estimate would be have expectation 0 when all single cells from various cell states are indiscriminately merged into equal-sized metacells, regardless of the true covariance value.

4.4 Semi-synthetic data generation for simulation

We generated a semi-synthetic dataset with true granularity level $\gamma^* = 50$ in a reference-based manner, making use of the scDesign3 simulator (Song et al., 2024). The reference data is a single-cell RNA-seq dataset from $n = 13{,}408$ bone marrow mononuclear cells of one donor processed using CITE-seq, which can be accessed from the dataset `bmcite` in the R package `SeuratData`, and the top $p = 2{,}000$ highly variable genes are considered as features. We use the variable $x_i \in \{1, 2, 3, 4, 5\}$ to denote the cell type of cell i, which is one-dimensional and consists of five categories, including B cells, T cells, progenitor cells, natural killer cells (NK), and monocytes or dendritic cells (Mono/DC).

4.4.1 Model fitting

Specifically, we followed scDesign3 (Song et al., 2024) to model the distribution of y_{ij}, the count of feature j in cell i, as a Negative Binomial (NB) distribution with the mean parameter, μ_{ij}, and the dispersion (size) parameter, σ_{ij}, conditional on the cell type variable x_i. For each feature j,

$$\begin{cases} y_{ij} | x_i & \overset{\text{ind}}{\sim} \mathrm{NB}(\mu_{ij}, \sigma_{ij}) \\ \log(\mu_{ij}) & = \alpha_{j0} + f_j(x_i) + \log(y_{i+}) \\ \log(\sigma_{ij}) & = \beta_{j0} + g_j(x_i) \end{cases}. \tag{21}$$

Since the NB distribution models both the biological and technical variations, we decompose it into the expression model and the measurement model for feature j:

$$y_{ij} | x_i \overset{\text{ind}}{\sim} \mathrm{NB}(\mu_{ij}, \sigma_{ij}) \quad \Longleftrightarrow \quad \begin{cases} \theta_{ij} | x_i & \overset{\text{ind}}{\sim} \mathrm{Gamma}(\sigma_{ij}, 1/\sigma_{ij}) \\ y_{ij} | \theta_{ij}, x_i & \overset{\text{ind}}{\sim} \mathrm{Poisson}(\mu_{ij}\theta_{ij}) \end{cases}, \tag{22}$$

where θ_{ij} denotes the biological variation around the true expression level μ_{ij} given x_{ij}. Note that $\mu_{ij}\theta_{ij} = c_i\lambda_{ij}$ in (1).

The distribution of $y_{ij}|x_i$ in (21) is fitted by the function `gamlss()` in the R package `gamlss` (version 5.4-22), which provides the estimated parameters $\hat{u}_{ij}$ and $\hat{\sigma}_{ij}$, which are needed for the measurement model and expression model, respectively. For convenience of notation in the following description, we denote the cumulative distribution function (CDF) of feature j's expression model Gamma$(\sigma_{ij}, 1/\sigma_{ij})$ as $F_j(\cdot|x_i)$ and, accordingly, the CDF of the fitted distribution Gamma$(\hat{\sigma}_{ij}, 1/\hat{\sigma}_{ij})$ as $\hat{F}_j(\cdot|x_i)$.

With a slight modification to scDesign3, we fitted a joint expression model of the p features conditional on the cell type variable, denoted by the CDF $F(\cdot|x_i) : \mathbb{R}^p \to [0,1]$, using a Gaussian copula:

$$F(\boldsymbol{\theta}_i|x_i) = \Phi_p\left(\Phi^{-1}\left(F_1(\theta_{i1}|x_i)\right), \cdots, \Phi^{-1}\left(F_p(\theta_{ip}|x_i)\right); \mathbf{R}(x_i)\right),$$

where $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{ip})^\top$, Φ^{-1} denotes the inverse of the CDF of the standard Gaussian distribution, and $\Phi_p(\cdot; \mathbf{R}(x_i)) : \mathbb{R}^p \to [0,1]$ denotes the CDF of a p-dimensional Gaussian distribution with a zero mean vector and a covariance matrix equal to the correlation matrix $\mathbf{R}(x_i)$. To estimate the copula, we adopted the plug-in approach where $\hat{\mathbf{R}}(x_i)$ is calculated as the sample correlation matrix of $\left\{\left(\Phi^{-1}\left(\hat{F}_1(y_{i'1}/\hat{\mu}_{i'1}|x_{i'})\right), \ldots, \Phi^{-1}\left(\hat{F}_p(y_{i'p}/\hat{\mu}_{i'p}|x_{i'})\right)\right)^\top : x_{i'} = x_i\right\}$, which includes all cells of the same type as cell i. The estimated joint CDF is thus

$$\hat{F}(\boldsymbol{\theta}_i|x_i) = \Phi_p\left(\left[\Phi^{-1}\left(\hat{F}_1\left(y_{i1}/\hat{\mu}_{i1}|x_i\right)\right), \cdots, \Phi^{-1}\left(\hat{F}_p\left(y_{ip}/\hat{\mu}_{ip}|x_i\right)\right)\right]^\top; \hat{\mathbf{R}}(x_i)\right). \tag{23}$$

4.4.2 Sampling from the fitted model

Next, we generated the semi-synthetic data by sampling from the fitted expression and measurement models. We started by generating $M = \lfloor n/\gamma^* \rfloor$ synthetic metacells, each representing a different cell state, by sampling without replacement from the n real cells. We kept the M sampled real cells' parameter estimates for the Poisson mean parameter (of the measurement model for feature $j = 1, \ldots, p$) in (22) as $\{\tilde{\mu}_{kj}\}_{k=1}^M \subset \{\hat{\mu}_{ij}\}_{i=1}^n$ and their cell types as $\{\tilde{x}_k\}_{k=1}^M \subset \{x_i\}_{i=1}^n$ for the M synthetic metacells, where $\tilde{x}_k \in \{1,2,3,4,5\}$. Then we independently sampled the biological variation of the kth synthetic metacell, $\tilde{\boldsymbol{\theta}}_k$, from the joint expression model $\hat{F}(\cdot|\tilde{x}_k)$, $k = 1, \ldots, M$ (23). In detail, to sample the kth synthetic metacell, given its cell-type covariate $\tilde{x}_k$, we independently sampled a p-dimensional vector from the estimated p-dimensional Gaussian distribution:

$$(v_{k1}, \ldots, v_{kp})^\top \overset{\text{ind}}{\sim} \Phi_p\left(\cdot; \hat{\mathbf{R}}(\tilde{x}_k)\right), \quad k = 1, \ldots, M.$$

Then based on the fitted expression model $\hat{F}_j(\cdot|\tilde{x}_k)$ of feature $j = 1, \ldots, p$, we calculated the expression level of feature j in the kth synthetic metacell:

$$\tilde{\theta}_{kj} = \hat{F}_j^{-1}\left(\Phi(v_{kj})|\, \tilde{x}_k\right), \quad j = 1, \ldots, p.$$

The sizes of synthetic metacells, denoted by $m_1, \ldots, m_M$, are randomly generated integers from the range 20 to 80 with the restriction that their mean equals to $\gamma^* = 50$. We then independently sampled synthetic single-cell feature counts from the fitted measurement model:

$$y_{i'j} \sim \text{Poisson}\left(\cdot; \omega_{i'j}\tilde{\mu}_{kj}\tilde{\theta}_{kj}\right), \quad i' \in \text{mc}_k,$$

where $\omega_{i'j}$ is the amplification factor for synthetic cell i' within the kth synthetic metacell, allowing for varying cell library sizes within the metacell, and mc_k denotes the index set of single cells assigned to the kth metacell ($|\text{mc}_k| = m_k$). To ensure that the cell library sizes in the semi-synthetic dataset closely

resemble those in the reference dataset, we independently sampled $\omega_{i'j}$ from the empirical distribution $\{\hat{\mu}_{ij}/\tilde{\mu}_{kj} : \ x_i = \tilde{x}_k; \ i = 1, \ldots, n\}$ for the kth metacell, where $\hat{\mu}_{ij}$ is the parameter estimate for the ith real cell, of the same type as the metacell, in the reference dataset, and $\tilde{\mu}_{kj}$ is the parameter of the kth metacell.

As a result of the above steps, the generated semi-synthetic dataset highly mimics the reference dataset (Supplementary Fig 1b–c).

4.5 Execution time of mcRigor

We report the execution time of mcRigor on an 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—the computation cost remains manageable and can be significantly reduced by employing parallel computing across different γ values. Additionally, the cost can be further lowered 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 (Bilous et al., 2024).

Table 1: The execution time of mcRigor (in minutes) for detecting dubious metacells and optimizing metacell method-hyperparameter configuration. Column names: "original": the time for implementing the original metacell methods; "mcRigor": the time for applying mcRigor to the generated metacell partition(s). "—" indicates the method is not applicable to the data (MetaCell and MetaCell2 are inapplicable to the scATAC-seq modality in "scMultiome" dataset; MetaCell2 is inapplicable to the "scRNA-seq + bulk ESC" dataset due to the small number of cells).

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	—	—	—	—	3.705	1.355	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

4.6 Implementation of metacell partitioning methods

In our analysis, we implemented four different metacell partitioning methods—SEACells (Persad et al., 2023), MetaCell (Baran et al., 2019), MetaCell2 (Ben-Kiki et al., 2022), and SuperCell (Bilous et al., 2022)—all of which mcRigor is compatible with. The SEACells method was executed using the Python package `SEACells` (version 0.3.3), following the detailed vignette available at `https://github.com/dpeerlab/SEACells/blob/main/notebooks/SEACell_computation.ipynb`. MetaCell was implemented with the R package `metacell` (version 0.3.7) following the vignette at `https://tanaylab.github.io/metacell/articles/a-basic_pbmc8k.html`, and MetaCell2 was implemented using the Python package `metacells` (version 0.9.5). The SuperCell method was carried out with the R package `SuperCell` (version 1.0).

Note that originally, only the SEACells method could be applied to identify metacell partitions for the scATAC-seq modality. To adapt the SuperCell method for scATAC-seq, we modified it by replacing its original normalization and dimensionality reduction steps with term frequency-inverse document frequency (TF-IDF) and singular value decomposition (SVD), respectively. However, since these two steps are built-in within the MetaCell and MetaCell2 methods, we were not able to modify them to accommodate scATAC-seq. Consequently, we used only SEACells and SuperCell for the scMultiome (RNA+ATAC) data (Table 1).

4.7 Projection of metacells onto the single-cell space for visualization

We visualized single cells with the identified metacells overlaid (Fig 1c, Supplementary Fig 4, Supplementary Fig 5a, Supplementary Fig 6-9b). Specifically, single cells were first plotted based on their two-dimensional embeddings (UMAP or t-SNE) using the `geom_scatter()` function from the R package `ggplot2` (v3.6.6). Each metacell was then positioned at the centroid of the single cells it encompasses in the two-dimensional embedding space and colored according to the predominant cell type within it. The dot size for each metacell reflects its actual metacell size, which is the number of single cells it includes.

4.8 Data analysis

4.8.1 Cell line data for cell cycle analysis

We considered five scRNA-seq datasets, each from one cell line. Three of these datasets, from cell lines A549, H2228, and HCC827, were generated using scRNA-seq with the 10× Chromium protocol (Tian et al., 2019). The other two datasets, from cell lines HEK293T and Jurkat, were sequenced using the Illumina HiSeq2500 Rapid Run V2 platform (Zheng et al., 2017). Raw sequencing data for A549, H2228, and HCC827 are available at the NCBI Gene Expression Omnibus (GEO) with accession code GSE150728, while the raw data for HEK293T and Jurkat can be accessed at `http://support.10xgenomics.com/single-cell/datasets` under the titles "293T Cells" and "Jurkat Cells," respectively. These five datasets from cell lines A549, H2228, HCC827, HEK293T, and Jurkat contain 1,237, 744, 567, 2,885, and 3,258 single cells, respectively.

Cell-cycle phase annotation. We assigned each single cell to a specific cell-cycle phase using canonical gene markers provided by Tirosh et al. (2016). This cell-cycle phase annotation was performed with the `CellCycleScoring` function from the `Seurat` R package (version 5.1.0). The corresponding R code is: `CellCycleScoring(Seurat.obj, s.features = s.genes, g2m.features = g2m.genes)`, where `s.genes` and `g2m.genes` represent the canonical gene markers for the S and G2M phases, respectively. Each metacell was annotated with the cell-cycle phase that contained the highest fraction of single cells within this metacell.

4.8.2 COVID-19 PBMC data for gene co-expression analysis

The data was collected by applying scRNA-seq through the Seq-Well platform (Gierahn et al., 2017) to profile peripheral blood mononuclear cells (PBMCs) from 7 patients hospitalized for COVID-19 and 6 healthy controls (Wilk et al., 2020). The processed count data and corresponding metadata can be accessed from the COVID-19 Cell Atlas (`https://www.covid19cellatlas.org/index.patient.html`) under the label *Peripheral Blood Mononuclear Cells (PBMCs)*. Raw sequencing data are available at NCBI GEO with accession code GSE150728. The processed data contains 44,721 single cells, among which 28,094 are from COVID-19 patients and 16,627 are from healthy controls. We used a subset this process data, all the cells annotated as B cells, for metacell partitioning and gene co-expression analysis. Note that both metacell partitioning and gene co-expression calculation were performed on the COVID-19 group (3,028 cells) and healthy group (1,994 cells) separately.

Gene module co-expression enrichment testing. For each gene module with g genes, we viewd the $g \times g$ gene correlation values as observations. Between the two conditions (COVID-19 vs. healthy), we applied the Wilcoxon rank-sum test to determine if the average correlation in the COVID-19 group was significantly higher than in the healthy group. A gene module was considered enriched in co-expression for COVID-19 if the Wilcoxon rank-sum test yielded a small one-sided p-value. We note that while we used the p-value to indicate enrichment (with smaller p-values representing stronger enrichment), we acknowledge that the use of the Wilcoxon rank-sum test in this context is ad hoc, as the correlation values are not independent nor identically distributed.

4.8.3 Single-cell multiome data for gene regulatory inference

This single-cell multiome (scMultiome) dataset, generated by Persad et al. (2023), is archived at NCBI GEO with accession code GSE200046. Data were collected by performing single-cell multiome ATAC + gene expression on cryopreserved bone marrow stem/progenitor CD34+ cells from a healthy donor, utilizing the 10X Genomics Chromium system and the Chromium Next GEM Single Cell Multiome Reagent Kit. The preprocessed dataset consists of 6,881 hematopoietic stem and progenitor cells (HSPCs) from healthy bone marrow, sorted for the pan-HSPC marker CD34.

Peak-gene associations. Gene-peak pairs were identified using the `LinkPeaks` function from the `Signac` R package (version 1.13.0). For each gene, this function computes a correlation score and p-value between gene expression and the accessibility of each peak within a specified distance (set at 100,000 base pairs) from the gene's transcription start site (TSS).

4.8.4 Paired scRNA-seq and smRNA FISH cell line data for distinguishing between biological and non-biological zeros

Data were generated by Torre et al. (2018) using scRNA-seq via the Drop-seq platform on 8,498 cells from a melanoma cell line. Paired smRNA FISH data was obtained from the same cell line, measuring 26 drug resistance markers and housekeeping genes across 7,000 to 88,000 cells. These paired datasets are available at NCBI GEO with accession code GSE99330. 16 genes overlap between the two datasets, including seven drug-resistance markers (*C1S*, *FGFR1*, *FOSL1*, *JUN*, *RUNX2*, *TXNRD1*, *VCL*) and nine housekeeping genes (*BABAM1*, *GAPDH*, *LMNA*, *CCNA2*, *KDM5A*, *KDM5B*, *MITF*, *SOX10*, *VGF*). Since our focus is on the number of zero counts, no normalization was performed to account for the technical differences between the platforms.

Zero proportion calculation. For the 16 overlapping genes as a whole, the zero proportion was calculated as the proportion of zeros within the observation-by-gene matrix (spatial spot-by-gene matrix for the smRNA FISH data or metacell-by-gene matrix for the metacell data derived from scRNA-seq data). For each individual gene, the zero proportion was calculated as the proportion of zeros within the vector of that gene's counts across all observations (spatial spots for the smRNA FISH data or metacells for the metacell data derived from scRNA-seq data).

4.8.5 Paired scRNA-seq and bulk data of human progenitor cells for differential gene expression analysis

This dataset was obtained by performing scRNA-seq on snapshots of lineage-specific progenitor cells differentiated from H1 human embryonic stem cells (H1 hESC) (Chu et al., 2016). The original dataset, archived at NCBI GEO with accession code GSE75748, contains 1,018 single cells and 19 bulk RNA-seq samples from snapshot progenitors. We used a subset of this dataset comprising 350 scRNA-seq cells—212 from H1 hESC and 138 from definitive endoderm cells (DEC)—and six corresponding bulk RNA-seq samples (four from H1 hESC and two from DEC). Due to the small total cell number, we considered smaller granularity levels ($\gamma = 2, \ldots, 50$) for metacell partitioning optimization.

Identification of differentially expressed genes. The `FindMarkers` function from the `Seurat` R package (version 5.1.0) was used to identify differentially expressed (DE) genes between the two cell types—H1 hESC and DEC. For fair comparison, we employed the DESeq2 method to identify DE genes in both the bulk and scRNA-seq data. The R code used is: `FindMarkers(Seurat.obj, ident.1 = 'H1 hESC', ident.2 = 'DEC', test.use = 'DESeq2', min.cells.group = 2)`.

4.8.6 Zman-seq data for temporal trajectory analysis

This data, generated by Kirschenbaum et al. (2024), profiles immune cells in mouse glioblastoma using the Zman-seq technology, which tracks transcriptomic dynamics over time by introducing fluorescent time stamps into immune cells. Processed data and metadata, including time stamps, are available at NCBI GEO with accession code GSE232040. We used a subset of the processed data consisting of 2,431 intratumoral T and natural killer (NK) cells.

Metacell annotation. We annotated each metacell by first examining the highest fraction of single cells within this metacell that originated from the same cell type. If this fraction is above 80%, the metacell is annotated with the corresponding major cell type. Otherwise, manual annotation was performed using marker genes suggested by Kirschenbaum et al. (2024), including *S1pr5*+ for NK Chemotactic; *Prf1*+, *Gzma*+, and *Gzmb*+ for NK Cytotoxic; *Itga1*+, *Xcl1*+, *Eomes*-, and *Pmepa1*+ for NK Dysfunctional.

Identification of time-dependent genes. Following the pipeline from Kirschenbaum et al. (2024), we identified time-dependent genes and generated heatmaps for smoothed gene expression over time. The R functions for this analysis, along with step-by-step guidance, are available at `https://github.com/kenxie7/ZmanR`. Specifically, after the cTET value is obtained for each metacall, we identified top genes significantly correlated with cTET using Spearman's rank correlation. Over the obtained p-values (from the Spearman's rank correlation test), we then applied Benjamini-Hochberg (BH) correction to adjust for multiple testing and identify significant time-dependent genes (FDR = 0.05). Gene expression was smoothed across cTET using loess fitting and normalized to range from 0 to 1 across all metacells. With these smoothed expression, the `Heatmap` function from the R package `ComplexHeatmap` (version 2.18.0) was used to visualize and cluster genes with similar temporal patterns.

5 Data availability

The datasets used in this paper are all publically available. The COVID-19 PBMC dataset, scMulitome dataset, Dropseq + smRNA FISH dataset, scRNA-seq + bulk ESC dataset, and Zman-seq dataset are archived at the NCBI Gene Expression Omnibus (GEO) with accession code GSE150728, GSE200046, GSE99330, GSE75748, and GSE232040, respectively. For the cell line datasets, raw data for A549, H2228, and HCC827 are available at the NCBI GEO with accession code GSE150728, and the raw data for HEK293T and Jurkat can be accessed at `http://support.10xgenomics.com/single-cell/datasets` under the titles "293T Cells" and "Jurkat Cells," respectively.

6 Code availability

The R package `mcRigor` is available at `https://github.com/JSB-UCLA/mcRigor`. Implementations of SEACells, MetaCell, MetaCells, and Supercell, as well as the generated semi-synthetic dataset are also included in the GitHub repository.

References

10x Genomics (2019, July). 5k PBMCs from a healthy donor, Single Cell ATAC dataset analyzed using Cell Ranger ATAC v1.1.0.

Andrews, T. S. and M. Hemberg (2018). False signals induced by single-cell imputation. *F1000Research 7*, 1740.

Ashuach, T., D. A. Reidenbach, A. Gayoso, and N. Yosef (2022). Peakvi: A deep generative model for single-cell chromatin accessibility analysis. *Cell reports methods 2*(3), 100182.

Baran, Y., A. Bercovich, A. Sebe-Pedros, Y. Lubling, A. Giladi, E. Chomsky, Z. Meir, M. Hoichman, A. Lifshitz, and A. Tanay (2019). Metacell: analysis of single-cell rna-seq data using k-nn graph partitions. *Genome Biology 20*, 1–19.

Ben-Kiki, O., A. Bercovich, A. Lifshitz, and A. Tanay (2022). Metacell-2: a divide-and-conquer metacell algorithm for scalable scrna-seq analysis. *Genome Biology 23*(1), 100.

Bilous, M., L. Hérault, A. A. Gabriel, M. Teleman, and D. Gfeller (2024). Building and analyzing metacells in single-cell genomics data. *Molecular Systems Biology 20*, 744–766.

Bilous, M., L. Tran, C. Cianciaruso, A. Gabriel, H. Michel, S. J. Carmona, M. J. Pittet, and D. Gfeller (2022). Metacells untangle large and complex single-cell transcriptome networks. *BMC Bioinformatics 23*(1), 336.

Bornstein, C., S. Nevo, A. Giladi, N. Kadouri, M. Pouzolles, F. Gerbe, E. David, A. Machado, A. Chuprin, B. Tóth, et al. (2018). Single-cell mapping of the thymic stroma identifies il-25-producing tuft epithelial cells. *Nature 559*(7715), 622–626.

Buenrostro, J. D., B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature 523*(7561), 486–490.

Cai, T. T. (2017). Global testing and large-scale multiple testing for high-dimensional covariance structures. *Annual Review of Statistics and Its Application 4*(1), 423–446.

Cai, T. T. and Z. Ma (2013). Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli 19*(5B), 2359–2388.

Cao, J., D. A. Cusanovich, V. Ramani, D. Aghamirzaie, H. A. Pliner, A. J. Hill, R. M. Daza, J. L. McFaline-Figueroa, J. S. Packer, L. Christiansen, et al. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science 361*(6409), 1380–1385.

Chen, S., B. B. Lake, and K. Zhang (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology 37*(12), 1452–1457.

Chen, S. X., L.-X. Zhang, and P.-S. Zhong (2010). Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association 105*(490), 810–819.

Chen, Y., H. Wen, Y. Li, Y. Han, J. Tan, C. Guo, C. Cai, P. Liu, Y. Peng, Y. Liu, et al. (2023). A multi-omics analysis reveals clspn is associated with prognosis, immune microenvironment and drug resistance in cancers. *Biological Procedures Online 25*(1), 16.

Chu, C. P., A. A. Morin, and P. Pavlidis (2024). Limits to the inference of gene regulation from bulk tissue expression data. *bioRxiv*, 1–16.

Chu, L.-F., N. Leng, J. Zhang, Z. Hou, D. Mamott, D. T. Vereide, J. Choi, C. Kendziorski, R. Stewart, and J. A. Thomson (2016). Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biology 17*, 1–20.

Cuomo, A. S., G. Alvari, C. B. Azodi, S.-C. eQTLGen Consortium, D. J. McCarthy, and M. J. Bonder (2021). Optimizing expression quantitative trait locus mapping workflows for single-cell studies. *Genome Biology 22*(1), 188.

Cusanovich, D. A., R. Daza, A. Adey, H. A. Pliner, L. Christiansen, K. L. Gunderson, F. J. Steemers, C. Trapnell, and J. Shendure (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science 348*(6237), 910–914.

Dean, I., C. Y. Lee, Z. K. Tuong, Z. Li, C. A. Tibbitt, C. Willis, F. a. Gaspal, B. C. Kennedy, V. Matei-Rascu, R. Fiancette, et al. (2024). Rapid functional impairment of natural killer cells following tumor entry limits anti-tumor immunity. *Nature Communications 15*(1), 683.

Eraslan, G., L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis (2019). Single-cell rna-seq denoising using a deep count autoencoder. *Nature communications 10*(1), 390.

Fishilevich, S., R. Nudel, N. Rappaport, R. Hadar, I. Plaschkes, T. Iny Stein, N. Rosen, A. Kohn, M. Twik, M. Safran, et al. (2017). Genehancer: genome-wide integration of enhancers and target genes in genecards. *Database 2017*, bax028.

Gandhi, M. K., E. Lambley, J. Duraiswamy, U. Dua, C. Smith, S. Elliott, D. Gill, P. Marlton, J. Seymour, and R. Khanna (2006). 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.

Gierahn, T. M., M. H. Wadsworth, T. K. Hughes, B. D. Bryson, A. Butler, R. Satija, S. Fortune, J. C. Love, and A. K. Shalek (2017). Seq-well: portable, low-cost rna sequencing of single cells at high throughput. *Nature Methods 14*(4), 395–398.

Giladi, A., F. Paul, Y. Herzog, Y. Lubling, A. Weiner, I. Yofe, D. Jaitin, N. Cabezas-Wallscheid, R. Dress, F. Ginhoux, et al. (2018). Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis. *Nature Cell Biology 20*(7), 836–846.

Granja, J. M., M. R. Corces, S. E. Pierce, S. T. Bagdatli, H. Choudhry, H. Y. Chang, and W. J. Greenleaf (2021). Archr is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature genetics 53*(3), 403–411.

Grün, D., L. Kester, and A. Van Oudenaarden (2014). Validation of noise models for single-cell transcriptomics. *Nature Methods 11*(6), 637–640.

Hao, Y., T. Stuart, M. H. Kowalski, S. Choudhary, P. Hoffman, A. Hartman, A. Srivastava, G. Molla, S. Madad, C. Fernandez-Granda, et al. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology 42*(2), 293–304.

Hartl, C. L., G. Ramaswami, W. G. Pembroke, S. Muller, G. Pintacuda, A. Saha, P. Parsana, A. Battle, K. Lage, and D. H. Geschwind (2021). Coexpression network architecture reveals the brain-wide and multiregional basis of disease susceptibility. *Nature Neuroscience 24*(9), 1313–1323.

Huang, M., J. Wang, E. Torre, H. Dueck, S. Shaffer, R. Bonasio, J. I. Murray, A. Raj, M. Li, and N. R. Zhang (2018). Saver: gene expression recovery for single-cell rna sequencing. *Nature methods 15*(7), 539–542.

Jia, Y., X. Cheng, W. Liang, S. Lin, P. Li, Z. Yan, M. Zhang, W. Ma, C. Hu, B. Wang, et al. (2022). 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.

Jiang, R., T. Sun, D. Song, and J. J. Li (2022). Statistics or biology: the zero-inflation controversy about scrna-seq data. *Genome Biology 23*(1), 31.

Kirschenbaum, D., K. Xie, F. Ingelfinger, Y. Katzenelenbogen, K. Abadie, T. Look, F. Sheban, T. San Phan, B. Li, P. Zwicky, et al. (2024). Time-resolved single-cell transcriptomics defines immune trajectories in glioblastoma. *Cell 187*(1), 149–165.

Lähnemann, D., J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz, et al. (2020). Eleven grand challenges in single-cell data science. *Genome biology 21*, 1–35.

Lareau, C. A., F. M. Duarte, J. G. Chew, V. K. Kartha, Z. D. Burkett, A. S. Kohlway, D. Pokholok, M. J. Aryee, F. J. Steemers, R. Lebofsky, et al. (2019). Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nature Biotechnology 37*(8), 916–924.

Lareau, C. A., S. Ma, F. M. Duarte, and J. D. Buenrostro (2020). Inference and effects of barcode multiplets in droplet-based single-cell assays. *Nature Communications 11*(1), 866.

Li, H., A. M. van der Leun, I. Yofe, Y. Lubling, D. Gelbard-Solodkin, A. C. van Akkooi, M. van den Braber, E. A. Rozeman, J. B. Haanen, C. U. Blank, et al. (2019). Dysfunctional cd8 t cells form a proliferative, dynamically regulated compartment within human melanoma. *Cell 176*(4), 775–789.

Li, W. V. and J. J. Li (2018). An accurate and robust imputation method scimpute for single-cell rna-seq data. *Nature communications 9*(1), 997.

Liu, B., X. Hu, K. Feng, R. Gao, Z. Xue, S. Zhang, Y. Zhang, E. Corse, Y. Hu, W. Han, et al. (2022). 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.

LOC117038772 GeneCards (2016). `https://www.genecards.org/cgi-bin/carddisp.pl?gene=LOC117038772`.

Lopez, R., J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef (2018). Deep generative modeling for single-cell transcriptomics. *Nature methods 15*(12), 1053–1058.

Love, M. I., W. Huber, and S. Anders (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology 15*, 1–21.

Lu, S. and S. Keleş (2023). Debiased personalized gene coexpression networks for population-scale scrna-seq data. *Genome Research 33*(6), 932–947.

Ma, S., B. Zhang, L. M. LaFave, A. S. Earl, Z. Chiang, Y. Hu, J. Ding, A. Brack, V. K. Kartha, T. Tay, et al. (2020). Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell 183*(4), 1103–1116.

Macosko, E. Z., A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell 161*(5), 1202–1214.

Mao, H., M. Jia, J. X. Dou, H. Zhang, and P. V. Benos (2022). Coem: cross-modal embedding for metacell identification. *arXiv preprint arXiv:2207.07734*.

Matsuzaki, J., S. Gnjatic, P. Mhawech-Fauceglia, A. Beck, A. Miller, T. Tsuji, C. Eppolito, F. Qian, S. Lele, P. Shrikant, et al. (2010). 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.

Morabito, S., F. Reese, N. Rahimzadeh, E. Miyoshi, and V. Swarup (2023). hdwgcna identifies co-expression networks in high-dimensional transcriptomics data. *Cell Reports Methods 3*(6), 100498.

Pearl, J. R., C. Colantuoni, D. E. Bergey, C. C. Funk, P. Shannon, B. Basu, A. M. Casella, R. T. Oshone, L. Hood, N. D. Price, et al. (2019). Genome-scale transcriptional regulatory network models of psychiatric and neurodegenerative disorders. *Cell Systems 8*(2), 122–135.

Persad, S., Z.-N. Choo, C. Dien, N. Sohail, I. Masilionis, R. Chaligné, T. Nawy, C. C. Brown, R. Sharma, I. Pe'er, et al. (2023). Seacells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nature Biotechnology 41*(12), 1746–1757.

Pfister, D., N. G. Núñez, R. Pinyol, O. Govaere, M. Pinter, M. Szydlowska, R. Gupta, M. Qiu, A. Deczkowska, A. Weiner, et al. (2021). Nash limits anti-tumour surveillance in immunotherapy-treated hcc. *Nature 592*(7854), 450–456.

Picelli, S., O. R. Faridani, Å. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg (2014). Full-length rna-seq from single cells using smart-seq2. *Nature protocols 9*(1), 171–181.

Sarkar, A. and M. Stephens (2021). Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nature Genetics 53*(6), 770–777.

Sebé-Pedrós, A., B. Saudemont, E. Chomsky, F. Plessier, M.-P. Mailhé, J. Renno, Y. Loe-Mie, A. Lifshitz, Z. Mukamel, S. Schmutz, et al. (2018). Cnidarian cell type diversity and regulation revealed by whole-organism single-cell rna-seq. *Cell 173*(6), 1520–1534.

Song, D., Q. Wang, G. Yan, T. Liu, T. Sun, and J. J. Li (2024). scdesign3 generates realistic in silico data for multimodal single-cell and spatial omics. *Nature Biotechnology 42*(2), 247–252.

Stuart, T., A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija (2019). Comprehensive integration of single-cell data. *Cell 177*(7), 1888–1902.

Su, C., Z. Xu, X. Shan, B. Cai, H. Zhao, and J. Zhang (2023). Cell-type-specific co-expression inference from single cell rna-sequencing data. *Nature Communications 14*(1), 4846.

Tian, L., X. Dong, S. Freytag, K.-A. Lê Cao, S. Su, A. JalalAbadi, D. Amann-Zalcenstein, T. S. Weber, A. Seidi, J. S. Jabbari, et al. (2019). Benchmarking single cell rna-sequencing analysis pipelines using mixture control experiments. *Nature Methods 16*(6), 479–487.

Tirosh, I., A. S. Venteicher, C. Hebert, L. E. Escalante, A. P. Patel, K. Yizhak, J. M. Fisher, C. Rodman, C. Mount, M. G. Filbin, et al. (2016). Single-cell rna-seq supports a developmental hierarchy in human oligodendroglioma. *Nature 539*(7628), 309–313.

Torre, E., H. Dueck, S. Shaffer, J. Gospocic, R. Gupte, R. Bonasio, J. Kim, J. Murray, and A. Raj (2018). Rare cell detection by single-cell rna sequencing as guided by single-molecule rna fish. *Cell Systems 6*(2), 171–179.

Trevino, A. E., F. Müller, J. Andersen, L. Sundaram, A. Kathiria, A. Shcherbina, K. Farh, H. Y. Chang, A. M. Pașca, A. Kundaje, et al. (2021). Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell 184*(19), 5053–5069.

Van Dijk, D., R. Sharma, J. Nainys, K. Yim, P. Kathail, A. J. Carr, C. Burdziak, K. R. Moon, C. L. Chaffer, D. Pattabiraman, et al. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell 174*(3), 716–729.

Wilk, A. J., A. Rustagi, N. Q. Zhao, J. Roque, G. J. Martínez-Colón, J. L. McKechnie, G. T. Ivison, T. Ranganath, R. Vergara, T. Hollis, et al. (2020). A single-cell atlas of the peripheral immune response in patients with severe covid-19. *Nature Medicine 26*(7), 1070–1076.

Xi, N. M. and J. J. Li (2021a). Benchmarking computational doublet-detection methods for single-cell rna sequencing data. *Cell Systems 12*(2), 176–194.

Xi, N. M. and J. J. Li (2021b). Protocol for executing and benchmarking eight computational doublet-detection methods in single-cell rna sequencing data analysis. *STAR Protocols 2*(3), 100699.

Xia, L., C. Lee, and J. J. Li (2024). Statistical method scdeed for detecting dubious 2d single-cell embeddings and optimizing t-sne and umap hyperparameters. *Nature Communications 15*(1), 1753.

Xu, F., J. Liu, D. Liu, B. Liu, M. Wang, Z. Hu, X. Du, L. Tang, and F. He (2014). Lsectin expressed on melanoma cells promotes tumor progression by inhibiting antitumor t-cell responses. *Cancer Research 74*(13), 3418–3428.

Xue, A., S. Yazar, D. Neavin, and J. E. Powell (2023). Pitfalls and opportunities for applying latent variables in single-cell eqtl analyses. *Genome Biology 24*(1), 33.

Zheng, G. X., J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, et al. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications 8*(1), 14049.