GeneCover: A Combinatorial Approach for Label-free Marker Gene Selection

An Wang[1], Stephanie Hicks[3,4,5,6], Donald Geman[1,2], and Laurent Younes[1,2]

[1] Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore MD 21218, USA
[2] Center for Imaging Science, Johns Hopkins University, Baltimore MD 21218, USA
[3] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[4] Department of Biomedical Engineering, Johns Hopkins School of Medicine, Baltimore, MD, USA
[5] Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA
[6] Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD, USA

Abstract. The selection of marker gene panels is critical for capturing the cellular and spatial heterogeneity in the expanding atlases of single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics data. Most current approaches to marker gene selection operate in a label-based framework, which is inherently limited by its dependency on predefined cell type labels or clustering results. In contrast, existing label-free methods often struggle to identify genes that characterize rare cell types or subtle spatial patterns, and they frequently fail to scale efficiently with large datasets. Here, we introduce geneCover, a label-free combinatorial method that selects an optimal panel of minimally redundant marker genes based on gene-gene correlations. Our method demonstrates excellent scalability to large datasets and identifies marker gene panels that capture distinct correlation structures across the transcriptome. This allows geneCover to distinguish cell states in various tissues of living organisms effectively, including those associated with rare or otherwise difficult-to-identify cell types. We evaluate the performance of geneCover across various scRNA-seq and spatial transcriptomics datasets, comparing it to other unsupervised algorithms to highlight its utility and potential in diverse biological contexts.

Keywords: Marker Gene Selection · Single-Cell RNA-seq · Spatial Transcriptomics

1 Introduction

The identification of marker genes plays a critical role in advancing our understanding of cellular and spatial heterogeneity at the transcriptomic level. With the continuous expansion of scRNA-seq and spatial omics data, the ability to identify informative marker gene panels has become essential for characterizing distinct cell states and their spatial distribution within tissues. These insights are fundamental for unraveling complex biological processes and for constructing comprehensive cellular atlases across various tissues and organisms.

Current approaches to marker gene selection can be broadly categorized into three types: generative, label-based and label-free. Generative methods [1,2,3,4,5] build statistical models of gene expression in which cell types and marker genes enter as latent variables. They typically rely on heavy computation and are not easily scalable. Label-based methods rely on predefined cell type labels to identify marker genes that differentiate between these cell types. Notable examples of such label-based methods include Seurat [6] differentially expressed gene (DEG) analysis, scGenefit [7], RankCorr [8], CellCover [9] etc. While effective, these methods have inherent limitations due to their reliance on clustering-based cell type labeling or manual annotation. Clustering-based labeling typically focuses on identifying cell sub-populations that exhibit significant variability at the transcriptomic level, as principal component analysis (PCA) is often applied prior to graph-based clustering. This emphasis on high-variability features may obscure the detection of subtle or rare cell types, as the genes characterizing these populations often do not display dominant variability patterns. Moreover, manual annotation at single-cell resolution requires expert knowledge, making the process both time-consuming and resource-intensive.

In contrast, most existing label-free marker gene selection methods, which do not rely on predefined cell type labels, adopt an imputation-based objective. These methods aim to select gene panels that effectively recover the underlying structure of the entire transcriptome. For example, PERSIST [10] selects genes that are maximally predictive of the overall gene expression profile using a concrete autoencoder network. Similarly, SCMER [11] identifies an optimal gene set by preserving the graph structure defined by pairwise cell similarity scores. GeneBasis [12] employs a greedy algorithm to select a gene panel that maintains the distance between the data manifold of the full transcriptome and that formed by the selected genes. Additionally, DUBStepR [13] uses stepwise regression on the gene-gene correlation matrix to predict the correlation matrix of the full gene set from the selected genes, iteratively regressing out the gene that explains the largest amount of variance in the residuals from the previous step. While these imputation-based methods provide effective unsupervised solutions by selecting features that preserve the global structure of complex, high-dimensional omics data, the gene panels they produce often reflect a broad, global representation of the data, yet are less sensitive to small cell populations. This is because the data structure these methods seek to recover is predominantly influenced by genes with high variability, or, as in the case of DUBStepR, the selection process itself is driven by explained variance. Consequently, these imputation-based approaches may overlook genes that are crucial for identifying rare cell types or capturing fine spatial organization, limiting their ability to detect nuanced biological signals.

Given the preceding discussion, both label-based methods and imputation-based label-free approaches face limitations in capturing biological signatures from rare sources of variability. To facilitate the discovery of genes associated with all sources of transcriptomic variability, we introduce geneCover, a label-free correlation-based marker gene selection method designed for single-cell RNA-seq and spatial transcriptomics data. GeneCover is motivated by the observation that, within highly heterogeneous gene expression profiles, groups of genes that characterize specific cell states or spatial organizations exhibit similar expression patterns, forming unique correlation structures among the transcriptome. This observation, which has also been made in methods such as DUBStepR [13], was independently recognized in our work. It highlights the potential of correlation-based methods for capturing both local and global signals by identifying distinct correlation groups formed by genes associated with rare and major cell types separately.

To capture these diverse genome-wide correlation structures, geneCover employs a minimal set-covering approach applied to the pairwise gene correlation matrix. This combinatorial strategy provides a globally optimal solution for the identification of minimally redundant gene panels that represents each distinct correlation structure, effectively characterizing unique spatial and cellular expression patterns at both local and global scales. By focusing on gene-gene correlations, geneCover is capable of identifying markers with subtle variations in transcriptomic activity that imputation-based methods often overlook, thereby providing a comprehensive tool for studying complex biological systems.

We demonstrate that geneCover enhances the detection of transcriptionally distinct cell types and spatially organized cell populations. To evaluate the performance of our method, we conducted an extensive benchmark analysis across multiple scRNA-seq and spatial transcriptomics datasets, comparing geneCover with leading imputation-based label-free marker gene selection methods. GeneCover consistently outperforms or matches the best imputation-based methods in recovering cell identities and spatial organization. Notably, geneCover increases the resolution of rare cell types and subtle spatial structures compared to imputation-based approaches. As illustrations, we show that geneCover successfully distinguishes the hippocampal subfields in a mouse brain Visium HD dataset [16,17], a level of resolution that other imputation-based methods fail to achieve. In a second example, analyzing a breast cancer scRNA-seq dataset [18], we show that geneCover identifies a transcriptionally distinct immune cell subpopulation, potentially linked to dendritic cells, which is absent from the original cell type annotation.

In addition to its accuracy, geneCover demonstrates superior scalability compared to existing label-free approaches, allowing it to handle large and complex datasets with significantly reduced computational overhead compared to competing methods. This scalability makes geneCover well-suited for use with the latest whole transcriptome spatial transcriptomics technologies at single-cell resolution, such as Visium HD, enabling the exploration of cellular heterogeneity at an unprecedented level of detail. Our method thus represents a robust, scalable, and comprehensive tool for marker gene selection in cutting-edge transcriptomic studies.

2 Methods

Notation In the analysis of spatial transcriptomics data, each tissue section is decomposed into a series of discrete locations, referred to as "spots," at different cellular resolutions depending on the sequencing technology. For scRNA-seq data, each individual cell serves as the fundamental unit of analysis. To unify these concepts, we denote both spots and cells as basic units in the target dataset, indexed from 1 to N. The gene expression levels across these units are represented as $X_j = (x_j^{(1)}, \ldots, x_j^{(N)}) \in \mathbb{R}^N$, where $j \in [\![d]\!] = \{1, \ldots, d\}$ denotes the jth gene out of the total d genes, and $x_j^{(i)}$ is the expression of gene j in the ith unit. We can convert X_j to its rank representation and denote it as $R(X_j)$, where the ith element of $R(X_j)$ records the rank of $x_j^{(i)}$ in X_j.

Each gene j is also associated with a weight w_j, reflecting the cost of its inclusion in the marker panel. To account for the gene-gene correlations, we define $\rho(j, j')$ as a measure of the correlation between X_j and $X_{j'}$, where $j, j' \in [\![d]\!]$. By default, we will set $w_j = 1$ for $j \in [\![d]\!]$ and use Spearman's correlation

$$\rho(j, j') = \frac{\text{COV}(R(X_j), R(X_{j'}))}{\sigma(R(X_j))\sigma(R(X_{j'}))}$$

as the correlation measure. Given a subset of genes $G \subseteq [\![d]\!]$ [1] and a correlation threshold λ, we define the neighborhood of gene $j \in G$ as $M_{G,j}^{\lambda} = \{j' \in G : \rho(j, j') \geq \lambda\}$. This correlation structure is encoded in a binary adjacency matrix $A_G^{\lambda} \in \{0, 1\}^{|G| \times |G|}$ such that $A_G^{\lambda}(j, j') = 1$ if $j' \in M_{G,j}^{\lambda}$ and $A_G^{\lambda}(j, j') = 0$ otherwise. Similarly, we denote $\boldsymbol{\rho_G}$ and $\boldsymbol{w_G}$ as the correlation matrix and weight vector with genes in G.

Minimal Weight Set Covering The minimal-weight set covering problem aims to identify a subset of genes $J \subseteq G$ that covers the remaining transcriptome while minimizing the total weights of the selected genes. Let $M_{G,j}$ denote the set of neighboring genes for each $j \in G$. A minimal set cover is a subset $J \subset G$ of minimum cardinality such that $\bigcup_{j \in J} M_{G,j} = G$. In other words, the covering requires each gene $j' \in G$ to belong to at least one $M_{G,j}$ where $j \in J$. The weighted version of the problem minimizes $\sum_{j \in J} w_j$ over the possible covering set J.

Minimal weight set covering is a classical problem in combinatorial optimization [19] and can be formulated as an integer programming problem. Introduce the binary vector $\boldsymbol{u} \in \{0, 1\}^{|G|}$, where $u_j = 1$

[1] $[\![d]\!]$ represents the whole transcriptome, and G is the index set of remaining genes after processing.

indicates that gene $j \in G$ is selected for the marker panel. Given the thresholding parameter λ, the integer programming formulation is

$$\begin{aligned} &\min_{u} \boldsymbol{w_G} \cdot \boldsymbol{u} \\ &\text{s.t.} \sum_{k \in G} A_G^{\lambda}(k,l) u_k \geq 1, \quad \forall l \in G \end{aligned} \tag{1}$$

We solve this integer programming problem using the Gurobi optimizer [20]. The optimal covering set is $J_G(\lambda) = \{j \in G : u_j = 1\}$.

Refinement and Size Adjustment The optimal solution $J_G(\lambda)$ obtained from the integer programming formulation includes genes that only cover themselves, i.e., $j \in J_G(\lambda)$ if $M_{G,j}^{\lambda} = \{j\}$. However, genes that correlate only with themselves or with few other genes often exhibit noisy expression patterns and contribute limited biological insight. To increase the robustness of our marker panel, we exclude genes $j \in J_G(\lambda)$ that cover fewer than m genes where $m > 1$. The final marker gene panel is thus refined as

$$J_G^*(\lambda) = \{j \in J_G(\lambda) : |M_{G,j}^{\lambda}| \geq m\}.$$

To obtain a marker gene panel J_G^* of pre-defined size k, we perform a binary search on parameter λ until $|J_G^*(\lambda)| = k$.

Expansion One may also wish to expand the marker panel to capture a broader set of genes representing multiple genes from each correlated gene group. Since minimal set covering identifies a compact set of genes that characterize distinct correlated gene groups, we can expand the panel by iteratively selecting additional genes from these groups. To achieve this, at each iteration $t+1$, we remove the lastest optimal marker panel $J_{G_t}^*$ from the set of genes considered in the previous iteration G_t and then run the minimal set covering on the remaining genes $G_{t+1} = G_t \setminus J_{G_t}^*$ with marker panel refinement to identify a new panel $J_{G_{t+1}}^*$ of pre-selected size that captures additional genes from the remaining correlated gene groups. The expanded marker panel is obtained by repeating this process over multiple iterations until the desired panel size or coverage is achieved.

GeneCover Algorithm The geneCover algorithm takes as input the whole-transcriptome gene-gene correlation matrix $\boldsymbol{\rho} \in \mathbb{R}^{d \times d}$, a positive weight vector $\boldsymbol{w} \in \mathbb{R}^d$, a target subset of the transcriptome $G \subseteq [\![d]\!]$, the marker panel refinement parameter m and the pre-selected marker panel size k or a non-decreasing sequence of sizes $\{k_t\}_{1:T}$ for successive expansions. It is fully described in Algorithm 1.

3 Results

To systematically evaluate the performance of geneCover, we applied our method across a range of distinct biological systems captured by both scRNA-seq and spatial transcriptomics datasets from multiple protocols (See Supplementary Section 1 for details of dataset processing and marker panels generation from other label-free methods):

- The **CBMC** CITE-Seq dataset [21] is a multimodal single-cell analysis derived from cord blood mononuclear cells (CBMCs). It combines cell-surface protein expression with transcriptomic data, offering a rich source of immune cell population information in the cord blood system.
- The **scFFPE breast cancer** dataset [18] provides a detailed molecular characterization of breast cancer, identifying 15 distinct cell types to improve understanding of tumor progression and immune interactions.
- The **DLPFC** dataset [22] offers spatial mapping of gene expression across the six layers of the human dorsolateral prefrontal cortex using the 10x Genomics Visium platform. With manual histological layer annotation, this dataset is commonly used as a benchmark for evaluating spatial transcriptomics methods.
- The **mouse brain Visium HD** dataset [17] is derived from an FFPE brain tissue block of an eight-week-old male mouse, providing a high-resolution, whole-transcriptome spatial mapping of multiple brain regions.

Algorithm 1 GeneCover

procedure GeneCover($k, \boldsymbol{\rho}, \boldsymbol{w}, G, m$)
  $J^* = \emptyset$
  $\lambda_{\min} = 0$
  $\lambda_{\max} = 1$
  while $|J^*| \neq k$ **do**
    $\lambda = \frac{1}{2}(\lambda_{\min} + \lambda_{\max})$
    $A_G^\lambda = \mathbf{1}_{\boldsymbol{\rho_G} > \lambda}$ ▷ Binarize $\boldsymbol{\rho_G}$ based on threshold λ
    u = MinimalSetCover($\boldsymbol{w_G}, A_G^\lambda$) ▷ Solution to Formulation (1) with weights $\boldsymbol{w_G}$ and matrix A_G^λ
    $J^* = \{j \in G : u_j = 1, |M_{G,j}^\lambda| \geq m\}$
    if $|J^*| < k$ **then**
      $\lambda_{\max} = \lambda$
    else
      $\lambda_{\min} = \lambda$
    end if
  end while
  Return J^*
end procedure

procedure Iterative-GeneCover($\{k_t\}_{1:T}, \boldsymbol{\rho}, \boldsymbol{w}, G, m$)
  for t = 1: T **do**
    $J_G^* =$ GENECOVER($k_t, \boldsymbol{\rho}, \boldsymbol{w}, G, m$)
    $J^* = J^* \cup J_G^*$
    $G = G \setminus J_G^*$
  end for
  Return J^*
end procedure

The following sections present an experimental evaluation of geneCover. First, we perform a benchmark analysis comparing geneCover with other leading label-free marker gene selection methods on the DLPFC dataset, demonstrating its effectiveness in identifying spatially organized gene expression patterns. Next, we illustrate that geneCover enhances the resolution of single-cell and spatial transcriptomic discoveries, focusing on its ability to uncover nuanced cell types and spatial organizations in the CBMC, mouse brain Visium HD, and scFFPE breast cancer dataset. Finally, we discuss the superior scalability of geneCover, showcasing its ability to efficiently handle large datasets.

3.1 Benchmark

To benchmark the performance of geneCover against other label-free gene selection methods, we conducted an experiment using the DLPFC dataset. We compared geneCover with five other methods: geneBasis, PERSIST, SCMER, and DUBStepR, which are label-free imputation-based marker gene selection methods, as well as Highly Variable Genes (HVGs). For each method, we obtained marker gene panels from the gene expression profile of the DLPFC dataset and restricted the log-normalized count matrix to the selected genes. We then performed principal component analysis (PCA), retaining 50 principal components based on these gene panels. To assess the clustering performance, we applied the Leiden [23] algorithm in SCANPY [24] with default parameters to these principal components across 30 random seeds and computed the average normalized mutual information (NMI) between the resulting clusters and the manually annotated histological layers. This benchmarking procedure allowed us to quantify how well each method's marker gene panel could recover the morphological structure of the prefrontal cortex.

The results, shown in Figure 1A, indicate that geneCover consistently outperforms the other methods across different marker panel sizes, with geneBasis being the closest competitor. Specifically, geneCover maintains this performance advantage across different marker panel sizes, demonstrating its robustness and effectiveness in recovering spatially organized structures in the DLPFC.

Despite being a label-free method, geneCover is able to identify layer-enriched signals that closely align with the manually annotated layers in the DLPFC dataset. For example, in Figure 1D, the white matter

region is predominantly defined by the geneCover marker MOBP. KRT17 shows higher expression in Layer 6, while the elevated expression of NEFH marks the boundary between layer 3 and 4. Likewise, layer 3 is enriched for CARTPT, and KRT19 is highly expressed in layer 1. These findings highlight geneCover's ability to detect biologically meaningful signals even without pre-defined labels. Additionally, as label-free marker gene selection methods like geneCover identify genes from all sources of variability, some genes may capture biological processes that are not directly associated with anatomical structures. For instance, we observe that SCGB2A2 is the most differentially expressed gene in Cluster 3. Interestingly, its expression shows spatial variability, even though it does not correspond to any of the annotated layers.

In summary, the benchmark analysis validates the effectiveness of geneCover in recovering biologically meaningful structures in spatial transcriptomics data. Its performance in this benchmark underscores its potential for providing more accurate and biologically relevant insights into additional spatial transcriptomic and scRNA-seq data.

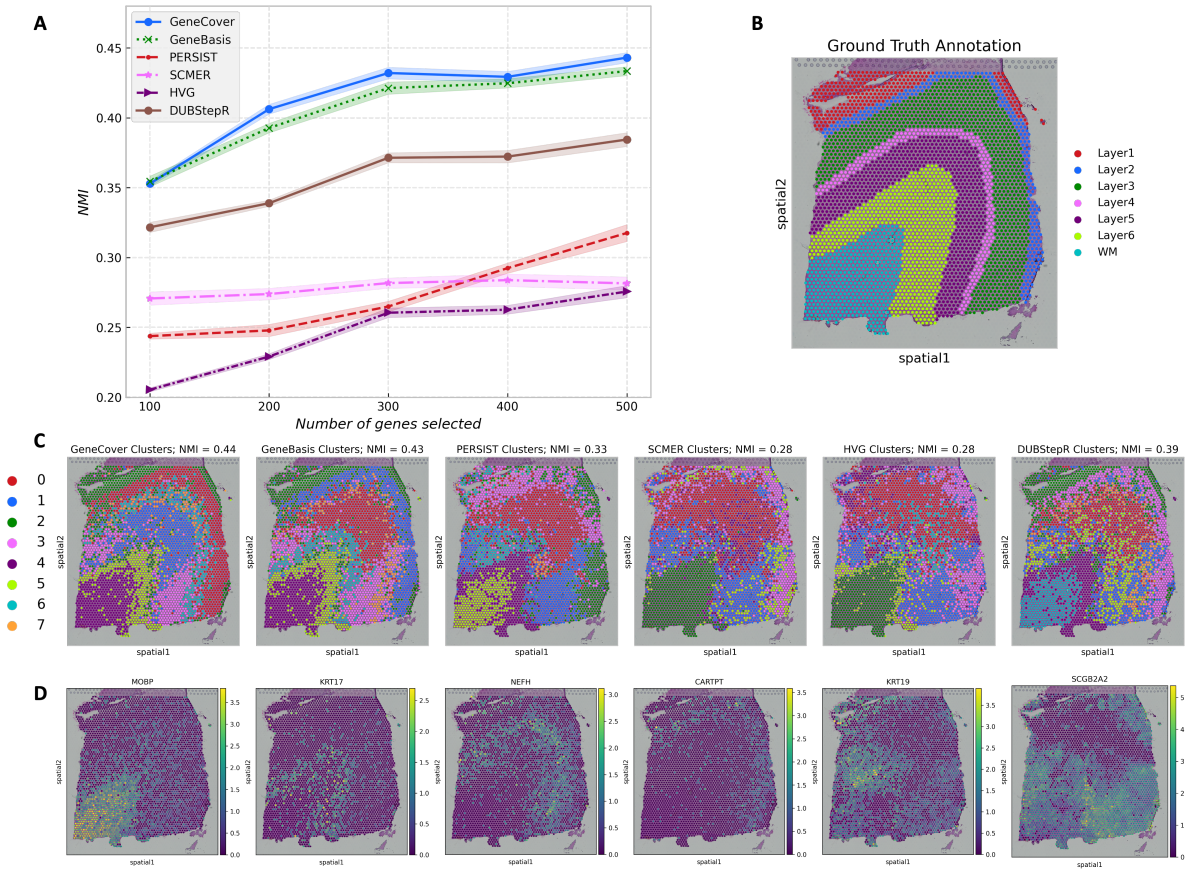

Fig. 1: **Benchmarking performance of geneCover and competing methods on the DLPFC dataset.** To obtain the GeneCover marker panel, we apply Iterative-GeneCover in Algorithm 1 with parameters: $\{k_t\}_{1:T} = \{100, 100, 100, 100, 100\}$, $m = 3$. (A) NMI versus the size of the marker gene panel for all label-free methods. (B) Manually annotated histological layers of the DLPFC. (C) Leiden clusters obtained using 500 genes selected by each method with the same random seed. (D) Expression of selected marker genes from geneCover.

3.2 GeneCover Improves Resolution in Single Cell and Spatial Transcriptomics Discovery

CBMC In this section, we highlight how geneCover discovers a minimally redundant set of marker genes that characterize the diverse cell types in the CBMC dataset. Using the same benchmarking procedure as in

the previous subsection, we find that geneCover consistently achieves the best performance, or aligns with the top-performing method, in recovering cell types once the marker panel size increases beyond 150 genes, as shown in Figure 2A.

When restricting the marker panel to 50 genes for each method, geneCover identifies a minimally redundant set of genes that effectively captures the diverse cellular architecture within the cord blood dataset. Figure 2C and 2E show the Spearman's correlation matrices for the first 50 genes identified by geneBasis and geneCover, respectively, with genes reordered using hierarchical clustering. Notably, geneCover identifies a more distinct set of correlated gene groups, as indicated by the clearer diagonal blocks in Figure 2C compared to Figure 2E. Moreover, the correlated gene groups identified by geneCover are visibly smaller, indicating its ability to reduce redundancy through the minimal set-covering approach. We also observe that geneCover selects significantly fewer redundant marker genes for certain cell populations. For example, while SCMER, PERSIST, DUBStepR (Supplementary Figure 1A, 1B, 1C), and geneBasis identify multiple highly correlated markers for the mouse cell population, geneCover selects only the MOUSE-ENO1 gene to represent this group. This highlights how geneCover efficiently explores the complex correlation structures within the omics data and selects a compact, non-redundant set of marker genes.

Importantly, Figures 2D and 2F illustrate that the markers identified by geneCover are more specific to individual cell types, whereas the markers from geneBasis are more broadly expressed across multiple cell types but less informative about the cell types. For instance, geneCover identifies CDKN1C as uniquely characterizing the CD16+ Mono cell type, while CLEC10A is exclusively expressed in DC cells. In contrast, the markers identified by geneBasis tend to favor widely expressed genes with broader variation patterns. Although this selection may improve clustering accuracy—since geneBasis achieves strong performance with 50 markers (as shown in Figure 2A)—it does not summarize a rich portfolio of cell-type-specific expression patterns, potentially limiting deeper biological insights. Due to the nature of the conventional clustering pipeline, particularly the PCA step, which emphasizes dominant variability in transcriptomic activity, the contribution of highly specific genes—typically exhibiting local variability patterns—to identifying transcriptionally distinct cell clusters is often diminished.

Notably, geneCover is also capable of distinguishing hierarchical gene expression patterns. For example, KIAA0101 (Figure 2D, fourth gene from right to left) is expressed in both the multiples and erythroid cell populations, while RHAG (Figure 2D, fifth gene from right to left) is uniquely expressed in erythroid cells. Even though KIAA0101 encompasses the variation pattern of RHAG, geneCover is still able to distinguish these two gene groups with overlapping yet distinct correlation structures.

Mouse Brain Visium HD We demonstrate that geneCover can effectively resolve hippocampal subfields in the mouse brain. We selected 200 marker genes using geneCover, geneBasis, and DUBStepR and applied the Leiden clustering algorithm on a cell-neighborhood graph, which was generated from the expression profile restricted to these marker genes (Figure 3A, 3B). Here, we avoid using principal components, which could downplay the contribution of marker genes that characterize spatial organization with very low abundance. As a comparison, we also applied the conventional clustering pipeline using 200 principal components from the whole transcriptome, followed by Leiden clustering (Figure 3C).

While all methods manage to identify the dentate gyrus in the mouse brain, geneCover uniquely divides the CA1-CA3 subiculum into two distinct regions (Figure 3A & Supplementary Figure 2A), while none of geneBasis, whole-transcriptome + PCA, and DUBStepR is capable of resolving this important hippocampal subregion, regardless of the random seeds used for clustering (Supplementary Figure 2B, 2C, 2D). This distinction is significant as the CA1-CA3 subiculum plays a crucial role in hippocampal function, contributing to memory formation and spatial navigation. Importantly, the two clusters identified by geneCover are transcriptionally distinct, as evidenced by the marker genes within the geneCover panel (Figure 3G). For example, Fibcd1 expression is uniquely localized to geneCover cluster 10 (Figure 3A, top right), which corresponds to the first division of the CA1-CA3 subiculum, while Chgb shows the highest expression in cluster 12 (Figure 3A, middle right), representing the second division of the CA1-CA3 subiculum.

To further quantify how well geneCover resolves these delicate spatial organizations, we compared its performance to geneBasis using the 15 graph-based clusters provided by 10X Genomics (Supplementary Figure 3) as a reference. We focused specifically on four reference clusters with the smallest cell abundances (clusters 12–15 in Figure 3E), each comprising less than 1% of the total cell population. We matched clusters from geneCover and geneBasis to the 10X Genomics clusters using the F1 score (see Supplementary Section

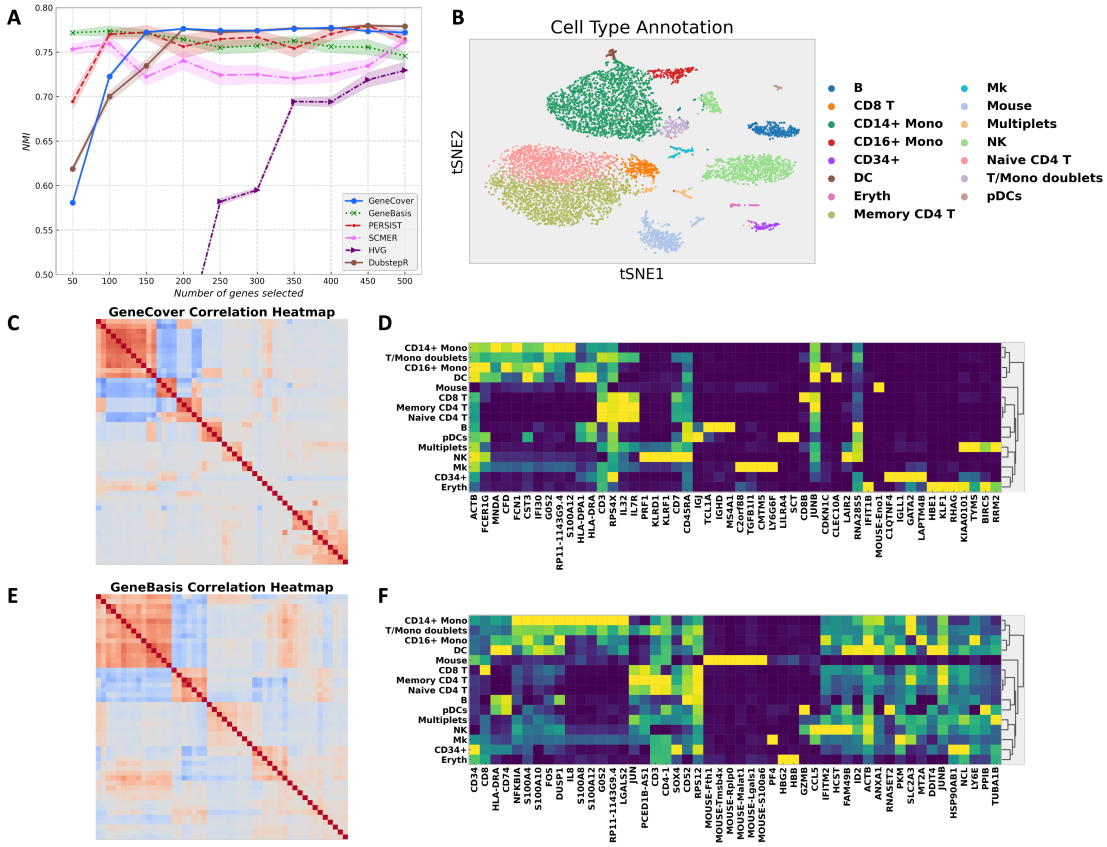

Fig. 2: **GeneCover identifies a minimally redundant set of marker genes characterizing diverse cell types in the CBMC dataset:** To obtain the GeneCover marker panel, we apply Iterative-GeneCover in Algorithm 1 with parameters: $\{k_t\}_{1:T} = \{50, 50, 50, 50, 50, 50, 50, 50, 50, 50\}$, $m = 3$. (A) NMI versus marker panel size for different label-free methods. (B) Cell type annotations for the CBMC dataset. (C) Spearman correlation heatmap of the 50 geneCover marker genes identified by geneCover, with gene reordered by hierarchical clustering. D) Expression matrix of geneCover markers, with same ordering as in (C), in cell types. Gene expression is standardized to $[0, 1]$ range. The color intensity represents the level of the normalized expression. (E) Same as (C) but for geneBasis markers. (F) Same as (D) but for geneBasis markers.

2 for details). According to Figure 3F, the clusters identified by geneCover consistently demonstrate better matching qualities with the four rarest reference 10X Genomics clusters based on F1 score (see Supplementary Figure 4 for matching quality comparisons on all 10X Genomics clusters). This result highlights geneCover's ability to enhance the resolution of spatial transcriptomics discovery, particularly in identifying highly refined spatial organizations, using a compact and minimally redundant set of genes.

scFFPE Breast Cancer GeneCover markers facilitate the identification of a transcriptionally distinct immune cell subpopulation that was absent from the original cell type annotations in the scFFPE breast cancer dataset. Specifically, within the originally labeled Macrophage 1 cell population (Figure 4A), geneCover uniquely identifies a subpopulation (Cluster 12 in Figure 4C) using 300 markers, a distinction that geneBasis struggles to achieve (Supplementary Figure 6). Moreover, we demonstrate that geneCover can reliably identify this potential immune subpopulation even with marker panels reduced to 100 or 200 genes. This performance remains robust across different random seeds used for clustering (Supplementary Figure 5).

Based on differential expression analysis of geneCover cluster 12, we hypothesize that this immune cell population may be related to dendritic cells, as geneCover identifies CD1C, CLEC10A, and FCER1A, which are all well-established markers of dendritic cells, as marker genes for this cluster (Figure 4E). CD1C is a

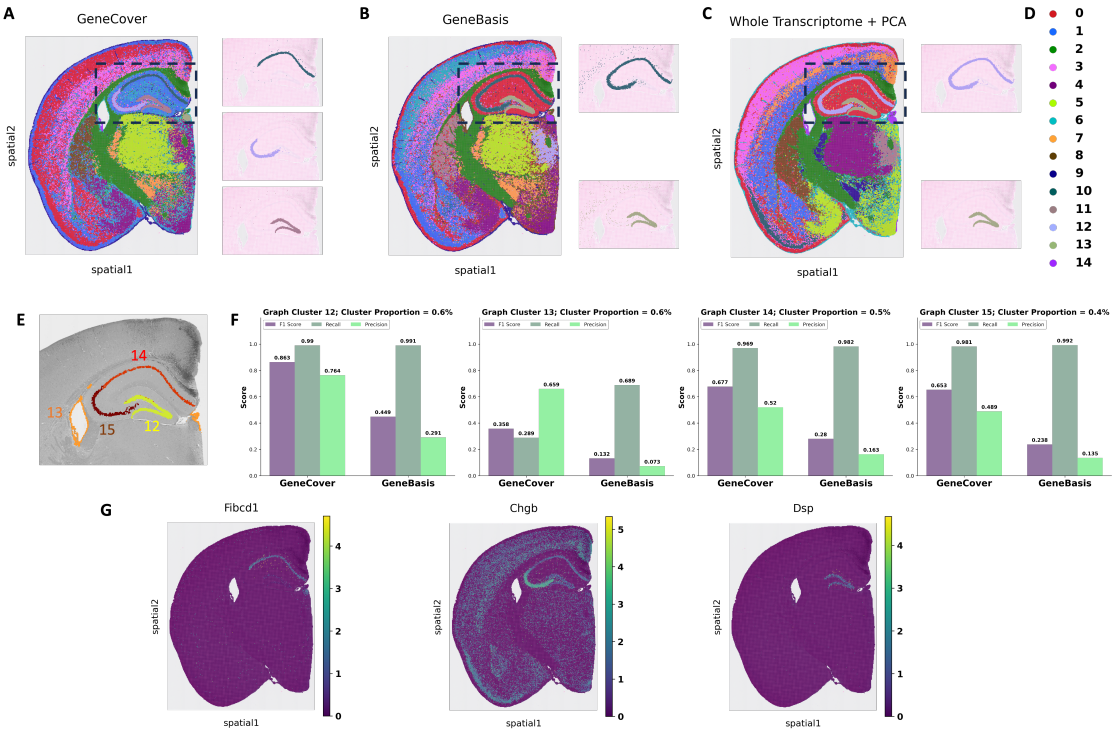

Fig. 3: **GeneCover resolves hippocampal subfields in the mouse brain:** To obtain the GeneCover marker panel, we apply Iterative-GeneCover in Algorithm 1 with parameters: $\{k_t\}_{1:T} = \{80, 60, 60\}$, $m = 3$. (A) Leiden clusters learned from 200 geneCover markers, with a highlight of clusters in the CA1-CA3 subiculum and dentate gyrus regions. (B) Same as Panel A but for 200 geneBasis markers. (C) Same as Panel A but for 200 principal components from the whole transcriptome. (D) Legend for clusters in Panel A, B, and C. (E) Four graph-based clusters provided by 10x Genomics with the lowest cell abundance. (F) Comparison of geneCover and geneBasis in resolving rare spatial organization in the mouse brain. (G) Differentially expressed geneCover markers for geneBasis clusters 10, 11, and 12.

marker of conventional dendritic cells type 2 (cDC2), which are crucial for presenting antigens and initiating immune responses. CLEC10A is specifically expressed on CD1C+ dendritic cells, enhancing their cytokine secretion in response to toll-like receptor stimulation, which contributes to their role in immune surveillance [25]. Notably, CLEC10A is also identified by geneCover in the CBMC dataset, where it is exclusively expressed in dendritic cells (Figure 2D, 16th gene from right to left). Lastly, FCER1A encodes the alpha chain of the high-affinity IgE receptor, which is expressed on dendritic cells and plays a critical role in mediating allergic responses by promoting antigen presentation and activation of immune cells in response to IgE-bound allergens [26]. Together, the identification of these marker genes suggests that the transcriptionally distinct immune cell subpopulation uncovered by geneCover may represent a previously unrecognized subset of dendritic cells within the tumor microenvironment.

The ability of geneCover to enhance the resolution of omics data analysis offers the potential for novel hypotheses regarding subtle cell types, shedding light on cell populations that may have been overlooked in previous studies.

3.3 Scalability

In this section, we demonstrate the scalability of geneCover when applied to large omics datasets. As shown in Table 1, geneCover significantly outperforms other label-free marker gene selection methods in terms of

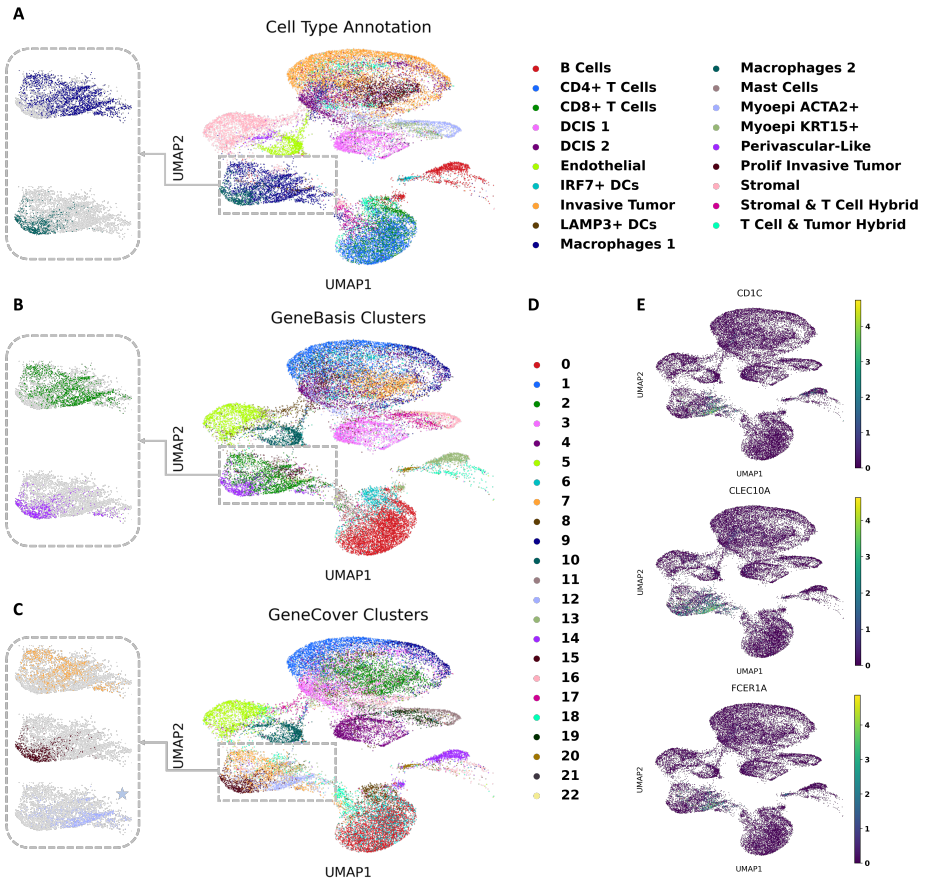

Fig. 4: **GeneCover Markers Facilitate the Identification of a Transcriptionally Distinct Immune Cell Subpopulation in Breast Cancer:** To obtain the GeneCover marker panel, we apply Iterative-GeneCover in Algorithm 1 with parameters: $\{k_t\}_{1:T} = \{100, 100, 100\}$, $m = 3$. (A) UMAP visualization of cell type annotations provided by the dataset, with a zoom-in on the Macrophage 1 and Macrophage 2 subpopulations. (B) Data-driven Leiden clusters learned from 300 geneBasis markers using the standard pipeline. (C) Same as Panel B, but for 300 geneCover markers. (D) Legend for the clusters in Panels B and C. (E) Differentially expressed genes for geneCover Cluster 12.

run time across the four datasets of consideration. Here, we include only label-free methods that allow the specification of marker panel size.

In particular, for the mouse brain Visium HD dataset, which contains approximately 100,000 bins, geneCover completes its task in just 134.7 seconds, making it approximately 500 times faster than the runner-up, SCMER, which requires 18.2 hours. Additionally, geneBasis takes over 93 hours to generate the marker panel, and PERSIST is unable to handle the dataset due to memory overflow, regardless of the various batch sizes tested. This stark contrast in performance highlights the limitations of imputation-based methods in efficiently processing large omics datasets.

The impressive scalability of geneCover can be attributed to its focus on gene-gene correlations. The most computationally intensive step is the calculation of the correlation matrix, which scales with the number of cells. However, this step can typically be executed efficiently using parallel computing. More importantly, because the input dimension for the minimal set covering problem is determined solely by the number of genes, the run time of the set covering algorithm remains invariant to the number of cells and depends only on the number of genes being considered. In contrast, the three imputation-based methods have time complexities that scale with both the number of cells and the number of genes. As a result, their run times increase considerably as the dataset size grows, making them significantly slower on large-scale datasets.

As advances in high-throughput whole-transcriptome spatial transcriptomics continue to push cellular resolution to new levels, the scalability of marker gene selection methods becomes increasingly critical. GeneCover is well-suited to adapt to these evolving data modalities, ensuring robust and efficient marker selection in the face of rapidly growing data complexity.

	DLPFC	**CBMC**	**scFFPE BC**	**Mouse Brain**
GeneCover	**22.1 s**	**18.3 s**	**35.4 s**	**134.7 s**
GeneBasis	515.4 s	2193.2 s	17.5 h	93.8 h
SCMER	241.6 s	1961.4 s	2.4 h	18.2 h
PERSIST	3.8 h	7.5 h	5.5 h	N/A

Table 1: **Empirical Run Time of label-free Marker Gene Selection Methods on Omics Datasets:** The time is measured in either seconds (s) or hours (h). For each method, we obtained 100 markers for the DLPFC and CBMC datasets, 300 markers for the scFFPE breast cancer dataset, and 200 markers for the mouse brain Visium HD dataset. GeneCover and geneBasis were run using the CPU, while SCMER and PERSIST were run on the GPU. N/A indicates memory overflow. All experiments were conducted on an Intel Core i9 13900K CPU or an NVIDIA GeForce RTX 4090 GPU.

4 Discussion

In this work, we propose geneCover as a novel label-free marker gene selection algorithm for single-cell and spatial transcriptomics. GeneCover employs a minimal weight set covering approach to identify a minimally redundant panel of marker genes that represent distinct correlation structures within the transcriptome. Our study demonstrates that geneCover provides a robust and scalable solution for label-free marker gene selection, outperforming existing label-free methods in both computational efficiency and accuracy in resolving glandular spatial organization or rare cell types in spatial transcriptomics and scRNA-seq data.

GeneCover excels at identifying a minimally redundant marker panel that captures various sources of transcriptional variability. More importantly, by focusing on exploring distinct gene-gene correlation groups through minimal set covering, geneCover makes the discovery of granular biological signals as feasible as identifying marker genes with large variations across cell types. This allows geneCover to significantly enhance the resolution of scRNA-seq and spatial transcriptomics discovery. Notably, geneCover markers enable the division of the highly refined CA1-CA3 subiculum in the mouse brain hippocampus into two distinct regions—a distinction that other existing strategies fail to achieve. In addition, using geneCover markers, we identify a transcriptionally distinct immune cell subpopulation characterized by dendritic cell markers, suggesting the potential to uncover previously unrecognized cell types or subpopulations. The identification of finely organized and transcriptionally distinct cell subpopulations is crucial for advancing our understanding of tissue heterogeneity, disease progression, and cellular dynamics. By enhancing resolution and enabling the detection of rare or spatially distinct cell types, geneCover offers an invaluable tool to explore complex biological systems at high resolution.

Beyond its improved resolution, geneCover demonstrates superior scalability, achieving significantly faster empirical run times compared to other label-free marker gene selection methods, particularly on large omics datasets. As cellular resolution continues to expand in whole-transcriptome spatial transcriptomics, the ability to process larger and more complex datasets efficiently will be critical. GeneCover offers a highly practical solution for modern high-throughput analyses and is well-positioned to adapt to these growing data modalities, paving the way for nuanced biological insights and deeper exploration of complex systems.

Code Availability The Python implementation of geneCover is publicly available on GitHub at: https://github.com/ANWANGJHU/GeneCover

Acknowledgments. The work is supported by NSF Award 2124230. We would like to express our gratitude to Jean Fan, Luigi Marchionni, and Caleb Hallinan for their valuable discussions on dataset selection and computational experiment design. We are also grateful to Hongyu Cheng for insightful discussions on integer programming.

References

1. Lopez, R., et al.: Deep generative modeling for single-cell transcriptomics. Nature methods **15**(12), 1053-1058 (2018)
2. Sun, Z., et al.: DIMM-SC: a Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. Bioinformatics **34**(1), 139-146 (2018)
3. duVerle, D. A., et al.: CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data. BMC bioinformatics **17**, 1-17 (2016)
4. Grønbech, C. H., et al.: scVAE: variational auto-encoders for single-cell gene expression data. Bioinformatics **36**(16), 4415-4422 (2020)
5. Armingol, E., et al.: The diversification of methods for studying cell–cell interactions and communication. Nature Reviews Genetics **25**(6), 381-400 (2024)
6. Hao, Y., et al.: Integrated analysis of multimodal single-cell data. Cell **184**(13), 3573-3587 (2021)
7. Dumitrascu, B., et al.: Optimal marker gene selection for cell type discrimination in single cell analyses. Nature communications **12**(1), 1186 (2021)
8. Vargo, A. H., et al.: A rank-based marker selection method for high throughput scRNA-seq data. BMC bioinformatics **21**, 1-51 (2020)
9. Ji, L., et al.: CellCover Captures Neural Stem Cell Progression in Mammalian Neocortical Development. bioRxiv (2024)
10. Covert, I., et al.: Predictive and robust gene selection for spatial transcriptomics. Nature Communications **14**(1), 2091 (2023)
11. Liang, S., et al.: Single-cell manifold-preserving feature selection for detecting rare cell populations. Nature computational science **1**(5), 374-384 (2021)
12. Missarova, A., et al.: geneBasis: an iterative approach for unsupervised selection of targeted gene panels from scRNA-seq. Genome biology **22**, 1-22 (2021)
13. Ranjan, B., et al.: DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data. Nature Communications **12**(1), 5849 (2021)
14. Chen, K. H., et al.: Spatially resolved, highly multiplexed RNA profiling in single cells. Science **348**(6233), aaa6090 (2015)
15. Gary, S.: Step away from stepwise. Journal of Big Data **5**(1), 1-12 (2018)
16. 10x Genomics: Visium HD Spatial Gene Expression (2024), https://www.10xgenomics.com/products/visium-hd-spatial-gene-expression
17. 10x Genomics: Visium HD Spatial Gene Expression Library, Mouse Brain FFPE (2024). https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-of-mouse-brain-he
18. Janesick, A., et al.: High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. Nature Communications **14**(1), 8353 (2023)
19. Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C.: Introduction to algorithms. MIT press (2022)
20. Gurobi Optimization, LLC: Gurobi Optimizer Reference Manual (2024), https://www.gurobi.com
21. Stoeckius, M., et al.: Simultaneous epitope and transcriptome measurement in single cells. Nature methods **14**(9), 865-868 (2017)
22. Maynard, K. R., et al.: Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nature neuroscience **24**(3), 425-436 (2021)
23. Traag, V. A., et al.: From Louvain to Leiden: guaranteeing well-connected communities. Scientific reports **9**(1), 1-12 (2019)
24. Wolf, F. A., et al.: SCANPY: large-scale single-cell gene expression data analysis. Genome biology **19**, 1-5 (2018)
25. Heger, L., et al.: CLEC10A is a specific marker for human CD1c+ dendritic cells and enhances their toll-like receptor 7/8-induced cytokine secretion. Frontiers in immunology **9**, 744 (2018)
26. Prussin, C., et al.: Omalizumab treatment downregulates dendritic cell FceRI expression. Journal of allergy and clinical immunology **112**(6), 1147-1154 (2003)