NAR Genomics and Bioinformatics, 2024, **6**, lqae124
https://doi.org/10.1093/nargab/lqae124
Advance access publication date: 24 September 2024
Computational Methods for Transcriptome Analysis

OXFORD

Comparative analysis of single-cell pathway scoring methods and a novel approach

Ruoqiao H. Wang [1] **and Juilee Thakar** [1,2,*]

[1]Department of Biomedical Genetics, University of Rochester, 601 Elmwood Ave, NY 14642, USA
[2]Department of Microbiology and Immunology, University of Rochester, 601 Elmwood Ave, NY 14642, USA
[*]To whom correspondence should be addressed. Email: juilee_thakar@urmc.rochester.edu

Abstract

Single-cell gene set analysis (scGSA) provides a useful approach for quantifying molecular functions and pathways in high-throughput transcriptomic data, facilitating the biological interpretation of complex human datasets. However, various factors such as gene set size, quality of the gene sets and the dropouts impact the performance of scGSA. To address these limitations, we present a single-cell Pathway Score (scPS) method to measure gene set activity at single-cell resolution. Furthermore, we benchmark our method with six other methods: AUCell, AddModuleScore, JASMINE, UCell, SCSE and ssGSEA. The comparison across all the methods using two different simulation approaches highlights the effect of cell count, gene set size, noise, condition-specific genes and zero imputation on their performance. The results of our study indicate that the scPS is comparable with other single-cell scoring methods and detects fewer false positives. Importantly, this work reveals critical variables in the scGSA.

Introduction

The increasing availability of single-cell RNA sequencing (scRNA-seq) data has necessitated the development of single-cell gene set analysis (scGSA) methods for biological interpretation. Although gene set analysis (GSA) approaches initially designed for bulk RNA-seq data are often used for scRNA-seq data (1–4), there is a significant shift towards methodologies specifically tailored to measure the gene set activities at the single-cell level (5–9). scGSA methods are rigorously assessed to address the unique challenges of scRNA-seq data, such as high noise levels, high dropout rate and large numbers of cells. The quantification at single-cell resolution with given gene sets can (i) improve post-hoc analysis by further clustering the cells based on their functionality, (ii) improve interpretation across multiple omics assays and (iii) lead to the discovery of novel pathways by minimizing the effect of averaging in heterogeneous data.

The commonly used scGSA methods are based on ranking or aggregating gene expression profiles. Although these methods offer valuable insights, they have several limitations, including sensitivity to the size of gene sets and the sparsity of the data, affecting the robustness and reproducibility of the results. The size of the gene sets varies depending on their function and source. In particular, the gene sets are obtained from multiple sources, such as databases, high-throughput data from the public domain and curation from published manuscripts, driving variation in gene set size. Moreover, most genes have low abundance at the single-cell level, and the ranking and aggregation do not always reveal true differences. Hence, in this study, we present a single-cell Pathway Score (scPS) method that improves our previously published single-person scoring method (10) to measure the activity of gene sets at the single-cell level. scPS uses principal component scores (PCs) weighted by their variance measured by principal component analysis (PCA) and the average gene set expression. PCs are the eigenvectors of the covariance matrix of the genes in the gene set and, during computation of the PCs, genes in the gene set underlying the variation at single-cell resolution obtain high weights. Hence, scPS can improve biological relevance by prioritizing the genes in the gene set using PCA (11,12).

The prevalence of excess zero or near-zero counts in the scRNA-seq data can obscure true biological signals and impact the gene set analysis. When applying PCA, cells with high proportions of zero counts may dominate certain principal components, distorting the representation of variation. Near-zero counts can introduce noise and inflate the importance of genes expressed at a low level, further impacting the interpretation of scGSA. To address this issue, various imputation methods have been developed, with scImpute being one of the prominent algorithms for zero imputation (13). scImpute effectively reduces dropout rates in scRNA-seq data by employing a sophisticated model that combines gamma and normal distributions. This model distinguishes between true zero expressions and technical dropouts, thereby providing a more accurate representation of actual gene expression levels. Hence, we also assess the performance of scGSA methods on scRNA-seq data with zero-preserving imputation.

Here, we revisit six commonly used scGSA methods: ssGSEA, UCell, AUCell, JASMINE, AddModuleScore and SCSE, and compare them with our novel approach, scPS. Briefly, we describe each method below. ssGSEA, AUCell,

Received: February 15, 2024. **Revised:** May 22, 2024. **Editorial Decision:** August 28, 2024. **Accepted:** September 3, 2024

UCell and JASMINE are ranking-based methods, while AddModuleScore and SCSE are count-based methods. Single-sample GSEA (ssGSEA) was initially developed for bulk RNA-seq data and calculated by the Kolmogorov–Smirnov-like random walk statistic (14). UCell calculates the Mann–Whitney U statistic (15). AUCell calculates the area under the curve (AUC) score among all ranked genes (16). Jointly Assessing Signature Mean and Inferring Enrichment (JASMINE) calculates the enrichment of the signature using the mean rank of the expressed genes, the proportion of expressed to non-expressed genes and the proportion of expressed genes in the gene set (17). AddModuleScore is a function in Seurat package and calculates scores for individual cells by aggregating the expression levels of each gene set, which are then subtracted by the aggregated expression of control feature sets randomly selected as background (18). The Single-Cell Signature Explorer (SCSE) is calculated as a sum of gene expression within gene sets divided by the sum of gene expression for each cell (19).

To compare different methods, we have used two simulation strategies to assess their specificity and sensitivity to differentiate cells with distinct gene set activity and detect true differential gene sets across biological conditions. This is the first time a detailed comparison of these seven methods has been performed. In particular, we assess the effect of the size of the cell counts, the composition of the gene sets, the condition-specific gene expression and zero imputation.

Materials and methods

To measure the accuracy of scGSA methods, we generated simulated data using two strategies where ground truth is known: (i) Splatter simulated data (SSD) were generated using the Splatter package (1.20.0) (20) based on the estimated parameters derived from the raw counts; and (ii) real-world simulated data (RWSD) were generated using log-transformed publicly available data (see below).

Data simulation and signal assignment

The 10X sequencing data from GSE164381 were loaded into R (version 4.2.1) (21) and processed with Seurat (version 4.2.0) (22–24) to generate simulated datasets. The data were normalized, clustered and annotated as described in the original paper (25). All non-detected genes were removed from the data, and an equal number of cells were randomly grouped into control and treatment groups. Sample sizes of 20, 50, 200 and 500 cells in each group were evaluated. Four scenarios of varying signal-to-non-signal ratios (SNRs) were developed. To model the differential expression, a signal was given to the chosen genes while other genes were the same between groups. The signal was defined as a 20% increase in gene expression in the treatment group compared with the control group. In scenario 1 (Figure 1A), the signal was given to at least 500 densely expressed genes, where the densely expressed genes were defined as the genes expressed in at least 90% of cells in SSD and 75% in RWSD. This criterion was applied independently for 10 replicates in both simulated datasets. In scenario 2 (Figure 1B), the signal was given to the randomly chosen 550 genes. For these two scenarios, the remaining genes were considered non-signal genes. To evaluate the zero imputation effect, we utilized scImpute (version 0.0.9) (13) with the recommended dropout threshold of 0.5 to mitigate the dropout rates in RWSD for scenarios 1 and 2. The additional two scenarios were designed to investigate the differences in the number of genes expressed across groups, which are frequently observed in cancer data. We introduced 250 (∼2%) more genes into the treatment group shown in scenario 3 and scenario 4 (Figure 1C, D). The expression of these added genes in the control group was set to 0, and no signal was assigned in scenarios 3 and 4. These four scenarios simulate the characteristics of scRNA-seq data, allowing us to evaluate scGSA performance.

Gene set simulation

The four scenarios described above allow us to evaluate variations in the number of genes responding to the experimental condition and the number of cells with that response. In addition, gene sets with different variability characteristics were simulated. In scenario 1 and scenario 2, we introduced noise in gene sets using the non-signal genes (see the previous section for details) to evaluate the scGSA performance when a certain proportion of genes that are not true functionally related genes in the set were present. The gene set sizes ranging from 10 to 500 were simulated by randomly choosing the signal and non-signal genes. For example, to simulate 20% noise, 20% of the genes in a gene set are drawn from non-signal genes. The following noise levels were modeled: 0, 20, 50, 80 and 100%. For each noise and gene set size combination, we generated 100 gene sets. In total, 5000 gene sets were generated per data. In scenario 3 and scenario 4, we created gene sets with first-row label, second-row label and third-row label genes (condition-specific genes) as illustrated in Figure 1C and D. In the scenario 3 condition-specific genes shown in the third-row label were densely expressed unlike in the scenario 4. Approximately 2% of the genes in these gene sets were condition-specific genes. Note that in scenarios 3 and 4, none of the genes received signals.

scGSA score calculation and comparison

Single-cell Pathway Score (scPS)

scPS is a modification of our previously published single-person transcription factor score (spTFscore) (10). To compute the scPS score, PCA was applied to the gene expression matrix of the gene set. The score was then determined as follows:

$$scPS = \mu \times \sqrt{\sum_{i=1}^{m}(s_i - s_{min}) \times v_i} \quad (1)$$

where μ is the mean gene expression of the gene set, s_i is the unweighted principal component score (PC), s_{min} is the minimum of the s_i among all the cells, v_i is the percentage of variance explained by the PC and m is the number of PCs at which 50% cumulative variance is explained.

scGSA methods for the comparative analysis

The comparative analysis was performed using seven scGSA methods, namely scPS, AddModuleScore (Seurat version 4.2.0), AUCell (version 1.18.1), JASMINE, SCSE, UCell (version 2.0.1) and ssGSEA (GSVA version 1.44.5). All methods (except scPS and AddModuleScore) were calculated using the irGSEA package (version 2.1.5). Then the Wilcoxon test was applied to the scGSA scores to compute the gene sets with differential expression, and the P-values were adjusted for the false discovery rate (FDR). A gene set with an adjusted P-value < 0.05 was considered statistically significant between the two experimental conditions. The recovery rate was calculated by

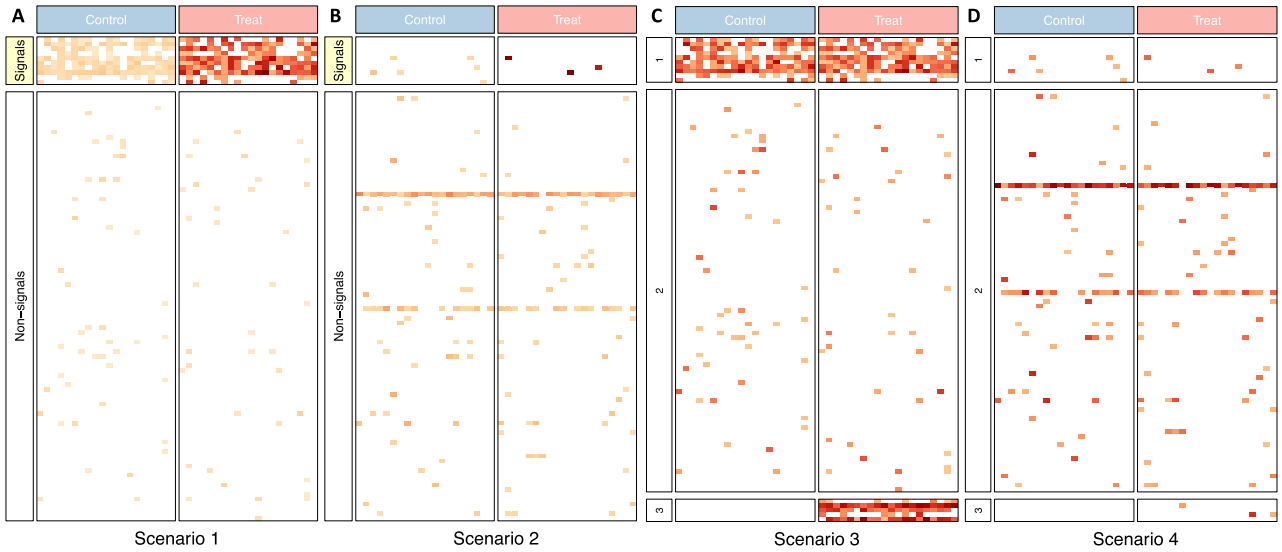

Figure 1. Data simulation schematics in the control (left) and treatment (right) groups. (**A**) Scenario 1, where the signal was assigned to the densely expressed genes. (**B**) Scenario 2, where the signal was assigned to the randomly chosen genes. The expression levels of these signal genes (yellow, top-row label) showed a 20% increase in the treatment group compared with the control group, while non-signal genes (white, bottom-row label) show no difference between the two conditions. (**C**) Scenario 3, the third-row label genes were densely expressed in the treatment group. (**D**) Scenario 4, the third-row label genes were randomly chosen in the treatment group. In C and D, the first-row label and second-row label genes showed no difference between the two conditions, while the third-row label genes were expressed only in the treatment group. Ten replicates for each scenario were generated.

dividing the number of identified gene sets by the total number of simulated gene sets.

Implementation of scPS in real data

The published peripheral blood mononuclear cell (PBMC) data from GSE198339 were used for the scPS implementation. The data were normalized, clustered and annotated as described in the original publication (26). The updated 2023 KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway from KEGGREST (version 1.42) and cell type-specific gene sets of B cells, T cells, natural killer (NK) cells and monocytes from Harmonizome (version 3.0) (27) were used.

Results

Effect of the cell count on scGSA performance

The cell count per group in the clustered scRNA-seq data could affect scGSA performance (28–30). To investigate the effect of cell count per group, we measured the recovery rate of the gene sets consisting of 100 signal genes, all of which received signals in the two scenarios described above (see the Materials and methods for more details). With an increase in cell count, the recovery rates of all except ssGSEA increased, whereas AUCell and JASMINE were minimally affected (Figure 2). In scenario 1, when densely expressed genes obtained the signal, scPS, AUCell, JASMINE and SCSE could identify at least 90% of gene sets for populations with $\geq$ 50 cells. In scenario 2, when the signal was dispersed among randomly chosen genes, scPS and SCSE were more affected by lower cell counts. However, the overall trend of the others remained consistent with scenario 1 (Figure 2A, B). Capturing meaningful variance for the PCA to calculate the scPS score was challenging due to the sparsity in the gene set expression matrix in scenario 2 (Supplementary Figure S1). However, the recovery rate of scPS exceeded 90% for clusters including $>$ 200 cells. Like scPS, the performance of AddModuleScore, SCSE and UCell improved as the number of cells increased in both scenarios, but UCell had a high standard deviation, attributable to the relatively high variability in the long tail of bottom-ranking genes. The results in RWSD were better than those in SSD, which had a higher signal-to-non-signal ratios (SNR) and relatively lower sparsity (Supplementary Figure S1E, F). All methods except ssGSEA generally showed improved performance with larger cell counts. This improvement was observed because larger cell counts enhanced the robustness of scGSA by boosting statistical power.

Effect of gene set size on scGSA performance

We investigate influence of gene set sizes, which vary based on their sources, on scGSA performance (28,31,32). Gene sets curated from published omics studies or automated text-mining of biomedical literature range from a dozen to thousands. To exclude the effect of cell count, we assessed the recovery rate with a fixed 200 cells per group when all methods consistently performed well (Figure 2). Furthermore, the gene sets with a noise level of 0% were used here, indicating that all genes in the gene set received the signal. No gene set should be recognized if no signal was assigned to the signal genes, i.e. no difference in gene expression was found in the treatment and control groups (Supplementary Figure S2). When the signal was assigned a 20% increase in the treatment group, we found that almost all differentially expressed gene sets between sizes of 200 and 500 were identified by all methods except ssGSEA. However, in the zero-inflated data, the high proportion of zeros within smaller gene sets can affect the results of these methods. In scenario 1, scPS, JASMINE, AUCell and SCSE performed well with a smaller gene set and had an almost 100% recovery rate in all cases except 82.7 $\pm$ 11.3% for AUCell and 64.9 $\pm$ 8.2% for SCSE with the gene set size of 10 in SSD (Figure 3A), and even better in RWSD with a higher

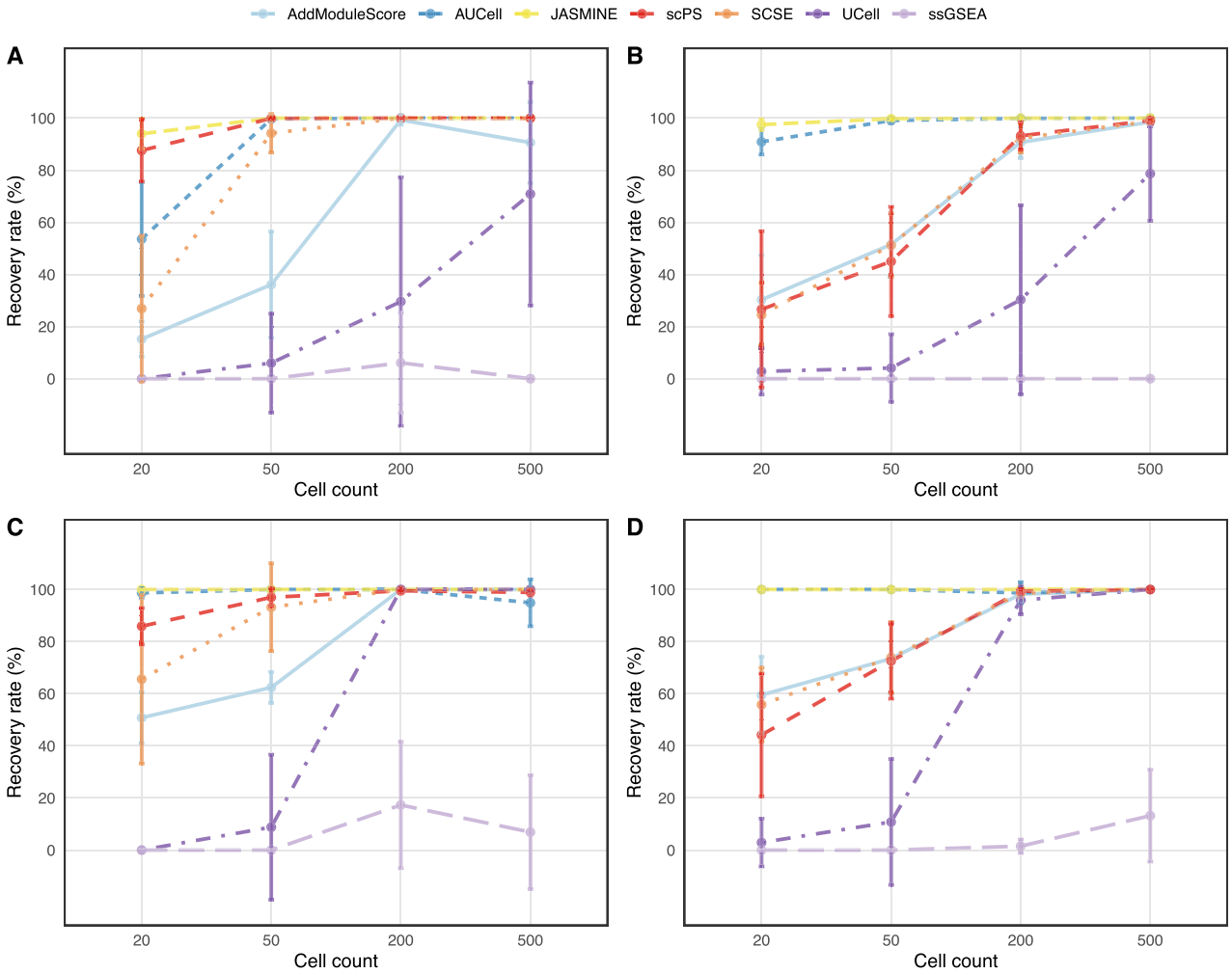

Figure 2. Effect of the cell count on scGSA performance. (**A**) Recovery rate of scGSA performance on SSD in scenario 1. (**B**) SSD in scenario 2. (**C**) RWSD in scenario 1. (**D**) RWSD in scenario 2. Recovery rate (*y*-axis) for varying cell count per group (*x*-axis) across the seven methods, with a gene set size of 100 and a noise level of 0%. The colors represent seven scGSA methods: AddModuleScore (light blue), AUCell (blue), JASMINE (yellow), scPS (red), SCSE (orange), UCell (purple) and ssGSEA (light purple).

SNR (Figure 3C; Supplementary Figure S1E). Furthermore, all methods except ssGSEA performed better in scenario 1 than in scenario 2 across all gene set sizes. This improved performance in scenario 1 can be attributed to the higher median number of cells expressing signal genes, which was 41.25% (32% for SSD and 44.5% for RWSD, Supplementary Figure S1G), compared with a lower median number of cells expressing signal genes—1.75% (Supplementary Figure S1H). Count-based methods such as AddModuleScore performed well with a gene set size >50 while UCell and ssGSEA performed poorly across all gene set sizes in SSD. However, the performance of UCell improved in RWSD (Figure 3C, D). Moreover, AUCell performed slightly better with RWSD in scenario 2, and better with zero-imputed data in scenarios 1 and 2 (Supplementary Figure S3A, B). AUCell performed better when the data had a higher SNR and lower dropout rate upon zero imputation because AUCell randomly assigned ranks to genes with the same expression, leading to skewed results. This randomization considered technical limitations in detecting genes with consistently low expression. Unlike AUCell, which relies solely on the ranking of genes within a set, JASMINE considers the mean rank of the gene set and the ratio of expressed and unexpressed genes within or outside the gene set. This dual consideration of both gene set rank and the comprehensive gene expression profile led to a higher recovery rate when the data had a higher dropout rate, like SSD. All methods had improved performance when zero imputation was applied, especially UCell and ssGSEA (Supplementary Figure S3A). Thus, gene set size can be an essential variable in the scGSA, and the results should be carefully interpreted.

Effect of the noise level on scGSA performance

Although gene sets are annotated molecular and cellular functions, they were frequently curated based on specific conditions (33,34), resulting in a likelihood that certain proportions of the genes are not always associated with the annotated biological function. Moreover, due to current technical limitations, some genes are not detected at single-cell resolution (35). In other words, not every gene within a given set accurately represented the underlying biological processes in heterogeneous scRNA-seq data (36). Therefore, we measured

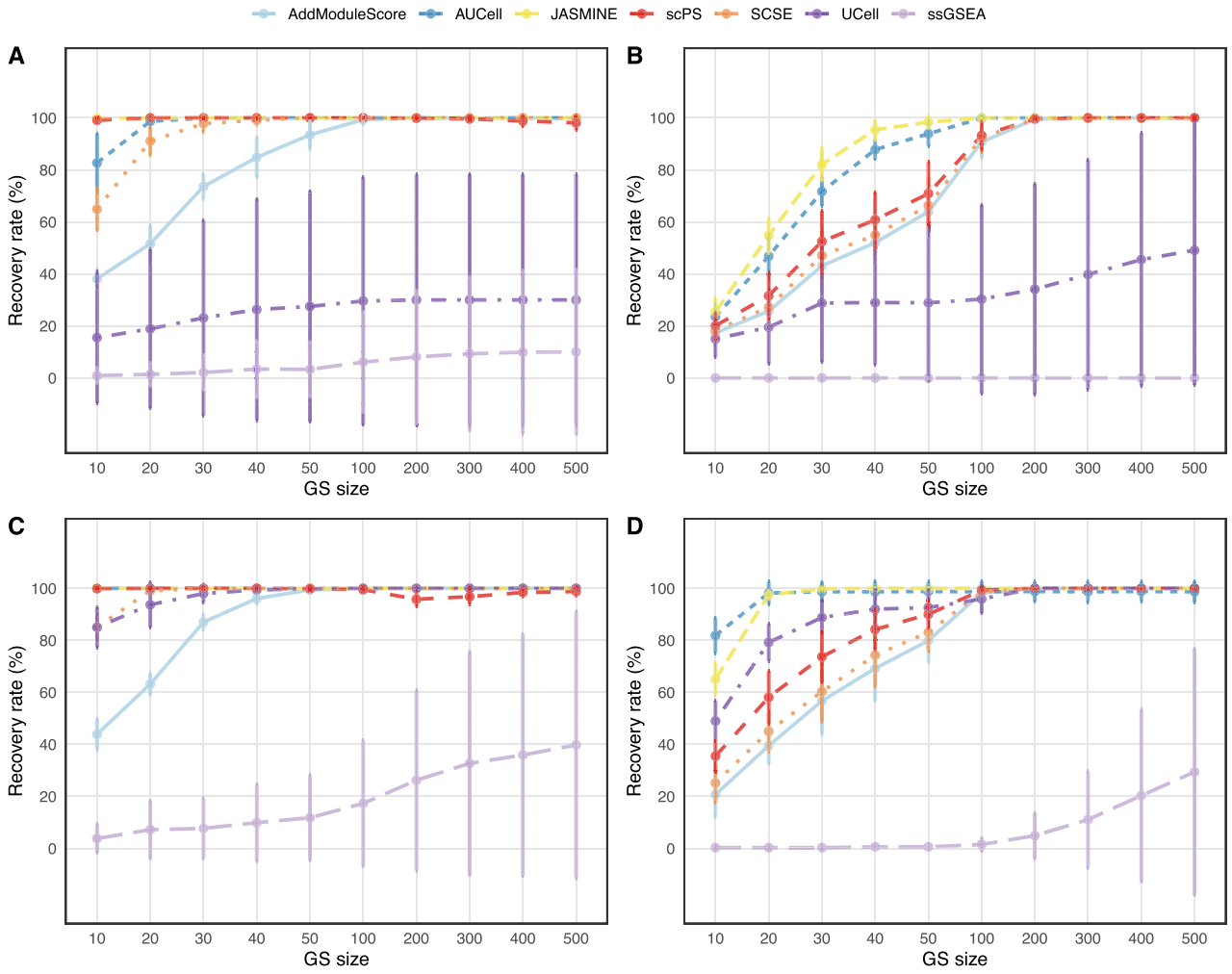

Figure 3. Effect of the gene set size on scGSA performance. (**A**) Recovery rate of scGSA performance on SSD in scenario 1. (**B**) SSD in scenario 2. (**C**) RWSD in scenario 1. (**D**) RWSD in scenario 2. Recovery rate (*y*-axis) for varying gene set size (*x*-axis) across the seven methods, with a cell count of 200 and a noise level of 0%.

the recovery rate of the gene set across different noise levels. This analysis utilized the simulated dataset with 200 cells per group and a gene set size of 100 in scenario 1 and scenario 2. scPS, JASMINE, AddModuleScore, AUCell and SCSE exhibited strong performance when applied to gene sets composed entirely of signal genes (noise level 0%) (see sensitivity in Table 1), and they consistently performed well in scenario 1 when the noise level ranges from 0 to 80% (Figure 4A, C). All methods performed better when the sparsity was reduced by applying zero imputation. scPS performed well at noise level 80% with and without zero-imputed data in scenario 1 compared with other methods, suggesting that it had an advantage in distinguishing subtle changes. Surprisingly, AddModuleScore performs worse at a noise level of 0% compared with 20% and 50% noise levels when using zero-imputed data, but this was not observed in data without zero imputation. Additionally, the performance of SCSE is significantly affected as the noise level increases in zero-imputed data, indicating that zero imputation methods can affect the data distribution and structure, potentially influencing how scGSA performs. The performance of UCell was poor in SSD but improved in RWSD and upon zero-imputed data, and ssGSEA barely detected any gene set in all cases (Supplementary Figure S3C).

In scenario 2, the recovery rate declined as the noise level increased across all methods. Nevertheless, scPS, JASMINE, AUCell, SCSE and AddModuleScore were able to identify >80% of gene sets with a noise level of 20%. However, AUCell, JASMINE, SCSE and UCell detected gene sets even if the genes in the gene set did not receive any signal (noise level 100%), especially in scenario 1 with and without zero-imputed data (see specificity in Table 1 and Supplementary Table S1). Despite a decrease in specificity in scenario 2, AUCell and JASMINE still detected random gene sets (Figure 4B, D). Therefore, JASMINE and AUCell had superior performance in identifying the differentially expressed gene set but detected more false-positive gene sets than other methods. This increased detection of false-positive gene sets by JASMINE and AUCell can be attributed to their reliance on gene expression rank without accounting for the magnitude of changes. Additionally, their sensitivity to minor variations or noise in the data further contributes to the false-positive detection. In contrast, scPS and AddModuleScore return a few false-positive gene sets at a 100% noise level in data with and without zero imputation (see the false-positive rate in Table 1 and Supplementary Table S1), which increases their accuracy and reliability.

Table 1. The sensitivity and false-positive rate of scGSA performance

	Sensitivity		False Positive Rate	
	SSD	RWSD	SSD	RWSD
Scenario 1				
JASMINE	100.0 ± 0.00	100.0 ± 0.00	74.4 ± 7.68	99.7 ± 0.48
AddModuleScore	99.3 ± 1.89	100.0 ± 0.00	3.5 ± 1.18	4.9 ± 1.66
AUCell	100.0 ± 0.00	100.0 ± 0.00	44.6 ± 13.66	100.0 ± 0.00
scPS	100.0 ± 0.00	99.5 ± 0.53	4.9 ± 2.47	4.1 ± 1.97
SCSE	100.0 ± 0.00	100.0 ± 0.00	25.6 ± 7.75	40.0 ± 6.51
UCell	29.6 ± 47.67	100.0 ± 0.00	0.8 ± 1.62	11.8 ± 5.73
ssGSEA	6.1 ± 19.29	17.3 ± 24.25	0.2 ± 0.63	0.7 ± 1.16
Scenario 2				
JASMINE	100.0 ± 0.00	100.0 ± 0.00	6.0 ± 2.45	25.0 ± 6.78
AddModuleScore	90.7 ± 6.04	98.2 ± 2.39	2.7 ± 1.70	2.9 ± 1.52
AUCell	99.9 ± 0.32	98.7 ± 4.11	2.5 ± 1.65	17.0 ± 3.89
scPS	93.2 ± 5.29	99.3 ± 0.82	2.5 ± 1.72	3.5 ± 2.32
SCSE	92.3 ± 5.54	98.6 ± 1.65	2.6 ± 1.26	2.7 ± 1.64
UCell	30.3 ± 36.25	95.9 ± 5.40	1.3 ± 2.26	1.8 ± 1.99
ssGSEA	0.0 ± 0.00	1.5 ± 2.59	0.0 ± 0.00	0.0 ± 0.00

Sensitivity is the recovery rate at a noise level of 0%, whereas false-positive rate is the recovery rate at a noise level of 100% with the gene set size of 100 and cell count of 200.

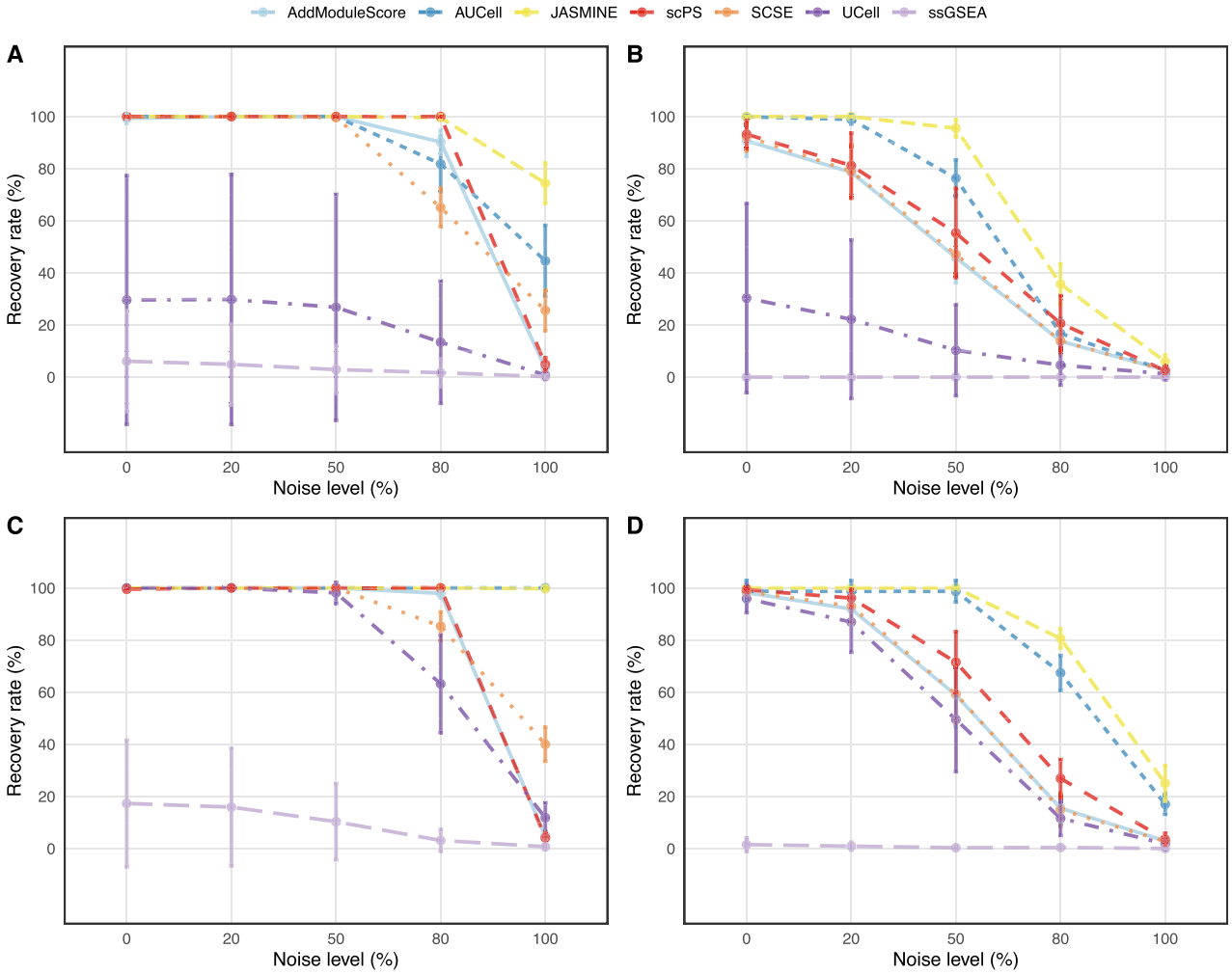

Figure 4. Effect of the noise level on scGSA performance. (**A**) Recovery rate of scGSA performance on SSD in scenario 1. (**B**) SSD in scenario 2. (**C**) RWSD in scenario 1. (**D**) RWSD in scenario 2. Recovery rate (*y*-axis) for varying noise level (*x*-axis) across the seven methods, with a cell count of 200 and a gene set size of 100.

Effect of the condition-specific genes on scGSA performance

To analyze the effect of the unequal number of detected genes across groups on scGSA performance, we developed scenarios 3 and 4. Briefly, no gene received any signal in these scenarios, indicating that they were not differentially expressed in any specific condition. Instead, gene sets were chosen based on their expression patterns; in scenario 3, genes were densely expressed across all the cells, and in scenario 4, genes were randomly chosen (see the Materials and methods for more details). This experiment aimed to test if condition-specific variation in expression drove the holistic enrichment results when gene set genes might not be differentially expressed. These scenarios represented observations across cancer cell types (17,37). It is important to note that none of the methods could effectively identify any gene set in the absence of a signal in scenarios 1 and 2 (Supplementary Figure S2). However, in scenario 3, all methods except scPS detected gene sets without condition-specific genes (Figure 5A, C). These methods, with the exception of scPS, always gave a positive result if there was an imbalance in the number of detected genes per group. Unlike scenario 3, the recovery rate decreased in scenario 4 (Figure 5B, D), although all methods except scPS could still detect a few random gene sets in RWSD.

We also analyzed how condition-specific genes influenced scGSA performance. When a higher proportion of genes exclusively expressed in one condition increased in the gene set, scPS demonstrated its capability to discern the underlying differences in scenario 3. The performance of ssGSEA improved when dealing with unequal numbers of genes between two groups due to its intrinsic ranking procedures. Nearly all gene sets without condition-specific genes were detectable by other methods across different gene set sizes (Supplementary Figure S3), even when there were no differences between genes in gene sets. In scenario 4, we might require a higher percentage of condition-specific genes to identify the differences successfully (2% condition-specific genes in the gene set). The rising quantity of condition-specific genes consistently improved the ability to identify the gene sets by scPS. In conclusion, scPS will correctly identify gene sets, such as the cell cycle set, which has condition-specific expression in cancer samples, whereas other methods will incorrectly identify gene sets that do not have condition-specific expression.

Analyses of peripheral blood mononuclear cells (PBMCs)

To test the performance of these methods on real scRNA-seq data, we used human PBMCs measuring impact of atherosclerosis (AS) in HIV-positive individuals (26). All methods were used to annotate cellular populations using cell type-specific gene sets and to detect functional differences in individuals with and without AS by KEGG pathways, respectively. The cell cluster annotations were compared with those from the original study (26). scPS successfully identified these cell clusters in a cell type-specific manner (Supplementary Figures S5 and S6). Notably, T-cell clusters demonstrated the highest scPS scores when assessed against the T-cell-specific gene set, effectively annotating the cell types (Figure 6A). Similarly, all other methods presented here also identified the cell types correctly (Supplementary Figures S7–S12).

The seven scGSA methods identified a few overlapping KEGG pathways to be significantly different between AS-positive and -negative groups in CD8+ T cells, and the ssGSEA identified fewer pathways compared with other methods (adjusted *P*-value threshold 0.05 shown in Supplementary Figure S13 and 0.01 presented in Figure 6B). ssGSEA exclusively identified 3 out of the 329 gene sets at an adjusted *P*-value < 0.05, one of which was the taste-transduction pathway, which we used as an example to investigate ssGSEA scores (Figure 6C). As a comparison, the hepatitis C pathway was also evaluated, which was only found by scPS and in the original paper (26). The ratio of non-zero genes among the data was low in both groups (Figure 6C, top). The size of the taste-transduction pathway is 86 genes, with only 15 genes expressed in the data, while 117 out of 158 genes from the hepatitis C pathway are expressed. Furthermore, the average gene expression of the hepatitis C pathway exhibited a normal distribution, with most of the cells expressing gene set genes at the > 0 level, whereas most of the taste-transduction pathway genes had zero expression following a negative binomial distribution (Figure 6C, bottom). Thus, ssGSEA identified small-size pathways with low expression. Moreover, the capacity of ssGSEA to accurately identify the significant pathway diminishes when subjected to a stringent threshold for the adjusted *P*-value. In contrast, both scPS and UCell exhibited greater capabilities in this regard (Figure 6B). Importantly, the MAPK signaling pathway and the JAK–STAT signaling pathway, both of which were identified by the original paper, were also identified using the scPS, but not by other methods. Furthermore, several interesting pathways were identified using the scPS that were not identified in the original analysis. These pathways include: the adrenergic signaling in cardiomyocytes pathway, the mTOR signaling pathway, the NF-κB signaling pathway and the chemokine signaling pathway. The identification of these pathways in the context of the CD8+ T-cell cluster from HIV-positive individuals with and without AS highlights the potential of our method to uncover significant biological processes that may have been overlooked in previous studies. For instance, the mTOR and NF-κB signaling pathways are crucial for T-cell function and activation, which are vital in both HIV progression and the development of atherosclerosis. The adrenergic signaling and chemokine signaling pathways are also known to play roles in cardiovascular and immune responses, respectively. This reveals that the scPS can provide deeper insights into the pathophysiology of HIV and atherosclerosis. Thus, scPS is one of the robust methods that has applicability to a wider range of datasets.

Discussion

Single-cell gene set analysis (scGSA) provides a powerful approach to unraveling the intricacies of cellular processes and heterogeneity from high-throughput datasets (16,38). By examining individual cells, scGSA enables the identification of gene sets or pathways that might be overlooked in pseudo-bulk analyses, particularly highlighting the enrichment or depletion of functions in rare cell types (39,40). Integration with other single-cell omics data enhances the comprehensiveness of analyses and, by focusing on gene sets, scGSA improves statistical power, especially in scenarios where individual genes may not exhibit significant changes but collectively contribute to specific biological functions (2,6,8). Furthermore, scGSA facilitates personalized medicine by uncovering molecular signatures associated with individual cells, aiding in tailored therapeutic interventions (41). Overall, scGSA serves as a crucial

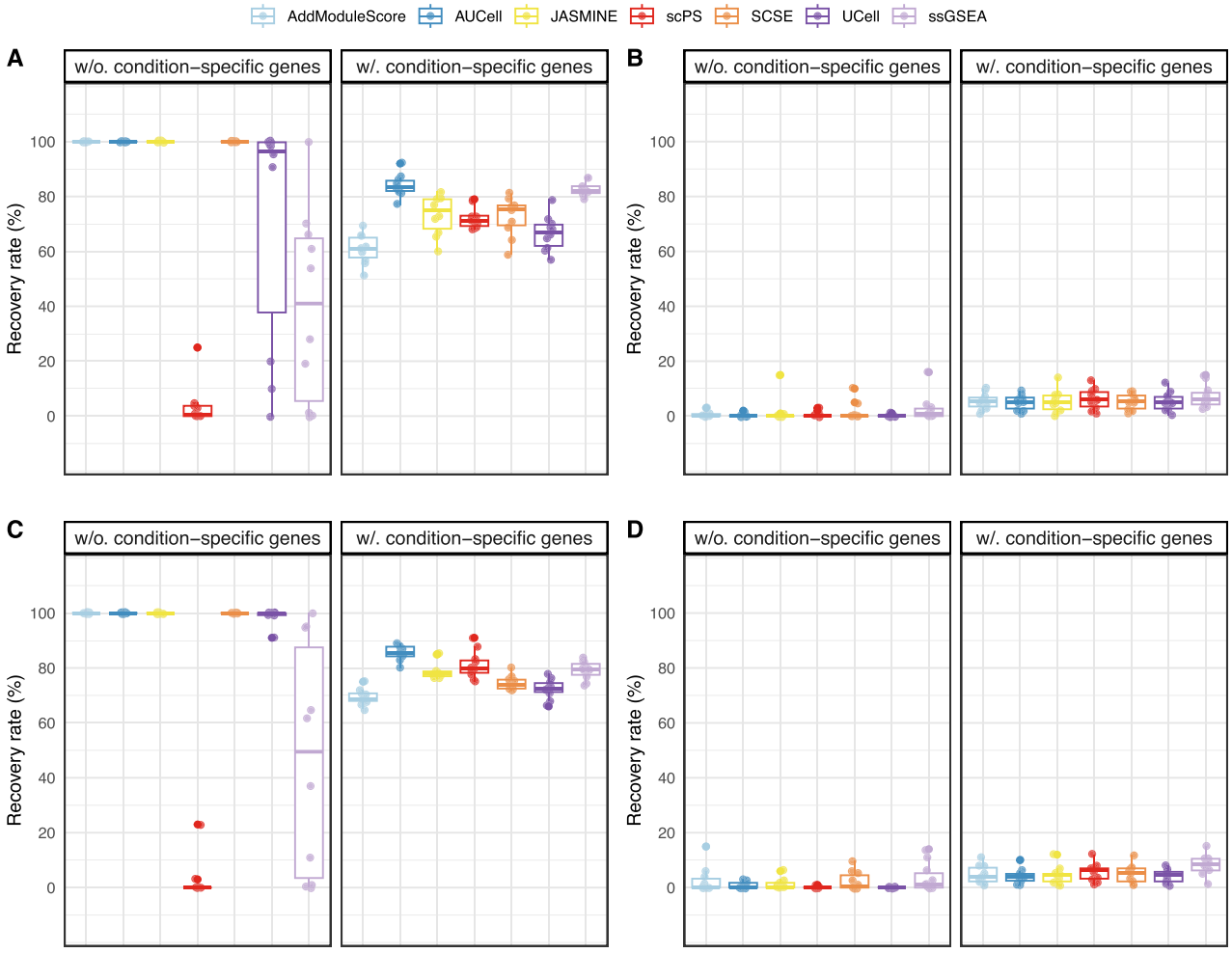

Figure 5. Effect of the condition-specific genes on scGSA performance. (**A**) Recovery rate of scGSA performance on SSD in scenario 3. (**B**) SSD in scenario 4. (**C**) RWSD in scenario 3. (**D**) RWSD in scenario 4. Recovery rate (*y*-axis) for the gene sets without the condition-specific genes (left) and with the condition-specific genes (right) across the seven methods, with a cell count of 200 and a gene set size of 100. For SSD and RWSD, no signal was assigned to the signal genes in scenario 3 (densely expressed genes) and scenario 4 (random genes).

tool for advancing our understanding of cellular biology and disease mechanisms at a finer resolution. Here, we propose a new method, scPS, and provide a detailed comparison of existing scGSA methods.

By capturing cellular heterogeneity and measuring transcripts in large numbers of cells, the scRNA sequencing technique alleviates concerns about statistical power. As it captures the diversity of cell types and states, even analyses with a relatively modest number of cells can yield meaningful insights into molecular functions defined by gene sets. Even when dealing with smaller sample sizes, the diversity captured at the single-cell level compensates for the potential loss in statistical power observed in bulk analyses. All methods, except ssGSEA, performed well with relatively smaller cell counts, and the recovery rates vary with the number of cells expressing the genes in the gene set. With sparse data below 50 cells, where PCA cannot provide accurate covariance estimates, scPS is not the best method, whereas AUCell and JASMINE may be better choices.

The large gene sets are beneficial for comprehensive inference of a biological function, especially for transcription factor gene sets with thousands of genes involved. However, these gene sets are generalized, and only a fraction of the genes within the gene set may be pertinent to the experimental conditions in reality. All methods except ssGSEA performed well when the gene sets consisted of genes densely expressed across the cells (Figure 3A, C; Supplementary Figure S4A). However, when genes in the gene set are sparsely expressed, scGSA performance improved with the gene set size (Figure 3B, D; Supplementary Figure S4B). All the methods except ssGSEA performed well when the gene set size was >100 in scenario 1 and scenario 2. We recommend applying AUCell and JASMINE to any gene set size less than 50 but greater than 10 for a sparse gene expression matrix of the gene set. Although scPS is not optimal for a sparse gene set matrix, in the size range from 10 to 50 zero imputation improves the recovery rates. Importantly, scPS has a low false-positive discovery rate (Figures 4 and 5; Supplementary Figure S4C, D).

Interestingly, in scenarios 3 and 4, large gene sets, even without the condition-specific expression, led to a higher false-positive discovery rate (Figure 5). If there is an unequal number of genes between two conditions, methods other than scPS have a high false-positive discovery rate. Notably, scPS weighs genes using their PCA loadings and thus achieves lower

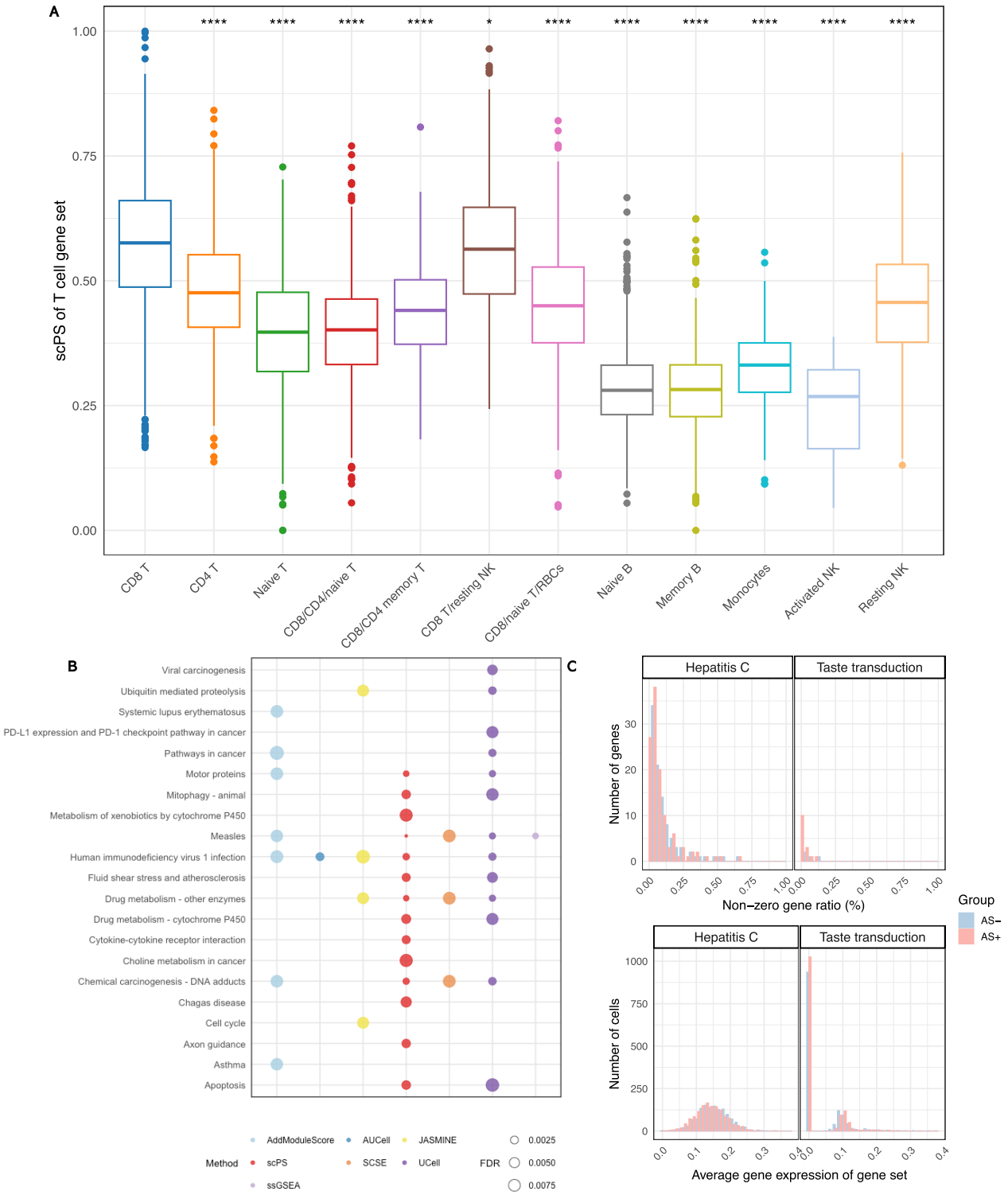

Figure 6. Application of scPS to PBMC data from HIV-positive people with and without atherosclerosis (AS). (**A**) scPS performance of a T-cell cluster-specific gene set across all cell subpopulations. Two-sample Wilcoxon signed-rank test, * indicates a *P*-value < 0.05, while **** signifies a *P*-value < 0.0001. (**B**) KEGG pathways (*y*-axis) dysregulated in CD8+ T cells from HIV-positive people with and without AS across the seven methods with an adjusted *P*-value of 0.01. (**C**) The distribution of the sparsity and average gene expression of the gene set in CD8+ T cells from HIV-positive people with (red) and without AS (blue). The top panel shows the number of genes (*y*-axis) across the gene set's non-zero gene ratio (*x*-axis). The bottom panel shows the number of cells (*y*-axis) across the average gene expression (log-normalized) of the gene set (*x*-axis).

false-positive discovery rates in all cases, even when genes were expressed in a condition-specific manner. The risk lies in the potential dilution of signal specificity, as several canonical pathways are generalized collections of genes associated with biological function without carefully considering the context of investigation (34). These gene sets, nevertheless, are crucial in investigating biological function in high-throughput datasets. Our results show that rank-based methods are susceptible to detecting false-positive gene sets because they rely on the relative ranking of genes rather than the magnitude of changes. Significant variability or noise in the data may lead to fluctuations in gene rankings, causing these methods to identify false-positive gene sets. As we discussed before in cases where the gene set size is large, random variations or noise may lead to the identification of false positives, especially if the ranking of individual genes within the gene set is not robust. scPS used the PCs and their variance to capture the gene set activity between and within diverse cell populations by accounting for the variations among cells within the gene set to estimate the biological function comprehensively. Furthermore, the incorporation of actual expression levels contributes to the biological relevance of the observed enrichments. scPS not only captures the nuances of cellular heterogeneity but also enhances the performance of scGSA by navigating the challenges posed by sparsity, ultimately providing a more accurate portrayal of biological processes at single-cell resolution. Additionally, with the highest specificity, scPS can identify the truly active and differently expressed gene sets, lowering the false-positive gene set detection rate. Most importantly, scPS utilizes Seurat data structures, streamlining implementation as Seurat is one of the most commonly used packages for scRNA-seq data analysis.

In conclusion, this manuscript provides a decisive comparison across scGSA methods with two different types of simulated data and four scenarios representing real biological data. In particular, four scenarios model varying levels of dropout events. ssGSEA, a method originally developed for bulk RNA-seq, did not perform well and has been reported in (17). In addition, we show that the application of imputation methods can improve the performance of all the methods. Thus, scPS performs best in controlling for false positives, and other methods specifically developed for single-cell data perform reasonably well.

Data availability

The data used by this study can be accessed in previously published studies at GSE164381 (25) and GSE198339 (26). All code and models are available at https://github.com/Thakar-Lab/scPS and https://doi.org/10.5281/zenodo.13620619.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

The authors thank the anonymous reviewers for their valuable suggestions.

Author contributions: RW and JT: conceptualization, data curation, formal analysis and writing—original draft, writing—review & editing. RW: implemented the method, code development and data analysis. JT: funding acquisition, project administration, resources and supervision

Funding

The National Institutes of Health [R01HL160229 to J.T.].

Conflict of interest statement

None declared.

References

1. Maleki,F., Ovens,K.L., Hogan,D.J., Rezaei,E., Rosenberg,A.M. and Kusalik,A.J. (2019) Method choice in gene set analysis has important consequences for analysis outcome. *J. Bioinform. Comput. Biol.*, **17**, 1940010.
2. Maleki,F., Ovens,K., Hogan,D.J. and Kusalik,A.J. (2020) Gene set analysis: challenges, opportunities, and future research. *Front. Genet.*, **11**, 654.
3. Das,S., McClain,C.J. and Rai,S.N. (2020) Fifteen years of gene set analysis for high-throughput genomic data: a review of statistical approaches and future challenges. *Entropy*, **22**, 427.
4. Geistlinger,L., Csaba,G., Santarelli,M., Ramos,M., Schiffer,L., Turaga,N., Law,C., Davis,S., Carey,V., Morgan,M., *et al.* (2021) Toward a gold standard for benchmarking gene set enrichment analysis. *Brief. Bioinform.*, **22**, 545–556.
5. Zhang,Y., Ma,Y., Huang,Y., Zhang,Y., Jiang,Q., Zhou,M. and Su,J. (2020) Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. *Comput. Struct. Biotechnol. J.*, **18**, 2953–2961.
6. Ma,Y., Sun,S., Shang,X., Keller,E.T., Chen,M. and Zhou,X. (2020) Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies. *Nat. Commun.*, **11**, 1585.
7. Lukassen,S., Ten,F.W., Adam,L., Eils,R. and Conrad,C. (2020) Gene set inference from single-cell sequencing data using a hybrid of matrix factorization and variational autoencoders. *Nat. Mach. Intell.*, **2**, 800–809.
8. Zhao,K. and Rhee,S.Y. (2023) Interpreting omics data with pathway enrichment analysis. *Trends Genet.*, **39**, 308–319.
9. Franchini,M., Pellecchia,S., Viscido,G. and Gambardella,G. (2023) Single-cell gene set enrichment analysis and transfer learning for functional annotation of scRNA-seq data. *NAR Genom. Bioinform.*, **5**, lqad024.
10. Cornwell,A., Palli,R., Singh,M.V., Benoodt,L., Tyrell,A., Abe,J.-I., Schifitto,G., Maggirwar,S.B. and Thakar,J. (2021) Molecular characterization of atherosclerosis in HIV positive persons. *Sci. Rep.*, **11**, 3232.
11. Yao,L., Jayasinghe,R.G., Lee,B.H., Bhasin,S.S., Pilcher,W., Doxie,D.B., Gonzalez-Kozlova,E., Dasari,S., Fiala,M.A., Pita-Juarez,Y., *et al.* . (2022) Comprehensive characterization of the multiple myeloma immune microenvironment using integrated scRNA-seq, CyTOF, and CITE-seq analysis. *Cancer Res. Commun.*, **2**, 1255–1265.
12. Moignard,V., Macaulay,I.C., Swiers,G., Buettner,F., Schütte,J., Calero-Nieto,F.J., Kinston,S., Joshi,A., Hannah,R., Theis,F.J., *et al.* (2013) Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nat. Cell Biol.*, **15**, 363–372.
13. Li,W.V. and Li,J.J. (2018) An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nat. Commun.*, **9**, 997.
14. Barbie,D.A., Tamayo,P., Boehm,J.S., Kim,S.Y., Moody,S.E., Dunn,I.F., Schinzel,A.C., Sandy,P., Meylan,E., Scholl,C., *et al.* (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, **462**, 108–112.
15. Andreatta,M. and Carmona,S.J. (2021) UCell: robust and scalable single-cell gene signature scoring. *Comput. Struct. Biotechnol. J.*, **19**, 3796–3798.
16. Aibar,S., González-Blas,C.B., Moerman,T., Huynh-Thu,V.A., Imrichova,H., Hulselmans,G., Rambow,F., Marine,J.-C., Geurts,P.,

Aerts,J., *et al.* (2017) SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*, **14**, 1083–1086.
17. Noureen,N., Ye,Z., Chen,Y., Wang,X. and Zheng,S. (2022) Signature-scoring methods developed for bulk samples are not adequate for cancer single-cell RNA sequencing data. *Elife*, **11**, e71994.
18. Tirosh,I., Izar,B., Prakadan,S.M., Wadsworth,M.H., Treacy,D., Trombetta,J.J., Rotem,A., Rodman,C., Lian,C., Murphy,G., *et al.* (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, **352**, 189–196.
19. Pont,F., Tosolini,M. and Fournié,J.J. (2019) Single-Cell Signature Explorer for comprehensive visualization of single cell signatures across scRNA-seq datasets. *Nucleic Acids Res.*, **47**, e133.
20. Zappia,L., Phipson,B. and Oshlack,A. (2017) Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.*, **18**, 174.
21. R Core Development Team (2010) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.
22. Hao,Y., Hao,S., Andersen-Nissen,E., Mauck,W.M., Zheng,S., Butler,A., Lee,M.J., Wilk,A.J., Darby,C., Zager,M., *et al.* (2021) Integrated analysis of multimodal single-cell data. *Cell*, **184**, 3573–3587.
23. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck,W.M., Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.
24. Butler,A., Hoffman,P., Smibert,P., Papalexi,E. and Satija,R. (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.*, **36**, 411–420.
25. Hoehn,K.B., Ramanathan,P., Unterman,A., Sumida,T.S., Asashima,H., Hafler,D.A., Kaminski,N., Dela Cruz,C.S., Sealfon,S.C., Bukreyev,A., *et al.* (2021) Cutting edge: distinct B cell repertoires characterize patients with mild and severe COVID-19. *J. Immunol.*, **206**, 2785–2790.
26. Palshikar,M.G., Palli,R., Tyrell,A., Maggirwar,S., Schifitto,G., Singh,M.V. and Thakar,J. (2022) Executable models of immune signaling pathways in HIV-associated atherosclerosis. *NPJ Syst. Biol. Appl.*, **8**, 35.
27. Rouillard,A.D., Gundersen,G.W., Fernandez,N.F., Wang,Z., Monteiro,C.D., McDermott,M.G. and Ma'ayan,A. (2016) The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, **2016**, baw100.
28. Liu,Q., Dinu,I., Adewale,A.J., Potter,J.D. and Yasui,Y. (2007) Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, **8**, 431.
29. Maciejewski,H. (2014) Gene set analysis methods: statistical models and methodological differences. *Brief. Bioinform.*, **15**, 504–518.
30. De Leeuw,C.A., Neale,B.M., Heskes,T. and Posthuma,D. (2016) The statistical properties of gene-set analysis. *Nat. Rev. Genet.*, **17**, 353–364.
31. Dinu,I., Potter,J.D., Mueller,T., Liu,Q., Adewale,A.J., Jhangri,G.S., Einecke,G., Famulski,K.S., Halloran,P. and Yasui,Y. (2009) Gene-set analysis and reduction. *Brief. Bioinform.*, **10**, 24–34.
32. Maleki,F., Ovens,K., McQuillan,I. and Kusalik,A.J. (2019) Size matters: how sample size affects the reproducibility and specificity of gene set analysis. *Hum. Genomics*, **13**, 42.
33. Lachmann,A., Rizzo,K.A., Bartal,A., Jeon,M., Clarke,D.J. and Ma'ayan,A. (2023) PrismEXP: gene annotation prediction from stratified gene–gene co-expression matrices. *PeerJ*, **11**, e14927.
34. Evangelista,J.E., Xie,Z., Marino,G.B., Nguyen,N., Clarke,D.J. and Ma'ayan,A. (2023) Enrichr-KG: bridging enrichment analysis across multiple libraries. *Nucleic Acids Res.*, **51**, W168–W179.
35. Qiu,P. (2020) Embracing the dropouts in single-cell RNA-seq analysis. *Nat. Commun.*, **11**, 1169.
36. Jiang,R., Sun,T., Song,D. and Li,J.J. (2022) Statistics or biology: the zero-inflation controversy about scRNA-seq data. *Genome Biol.*, **23**, 31.
37. Guo,W., Wang,D., Wang,S., Shan,Y., Liu,C. and Gu,J. (2021) scCancer: a package for automated processing of single-cell RNA-seq data in cancer. *Brief. Bioinform.*, **22**, bbaa127.
38. Osborn,R.M., Leach,J., Zanche,M., Ashton,J.M., Chu,C., Thakar,J., Dewhurst,S., Rosenberger,S., Pavelka,M., Pryhuber,G.S., *et al.* (2023) Preparation of noninfectious scRNAseq samples from SARS-CoV-2-infected epithelial cells. *PLoS One*, **18**, e0281898.
39. Ianevski,A., Giri,A.K. and Aittokallio,T. (2022) Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat. Commun.*, **13**, 1246.
40. Shi,X., Zhang,L., Li,Y., Xue,J., Liang,F., Ni,H.-W., Wang,X., Cai,Z., Shen,L.-H., Huang,T., *et al.* (2022) Integrative analysis of bulk and single-cell RNA sequencing data reveals cell types involved in heart failure. *Front. Bioeng. Biotechnol.*, **9**, 779225.
41. Chen,Z., Chen,H., Yu,L., Xin,H., Kong,J., Bai,Y., Zeng,W., Zhang,J., Wu,Q. and Fan,H. (2021) Bioinformatic identification of key pathways, hub genes, and microbiota for therapeutic intervention in *Helicobacter pylori* infection. *J. Cell. Physiol.*, **236**, 1158–1183.

Received: February 15, 2024. **Revised:** May 22, 2024. **Editorial Decision:** August 28, 2024. **Accepted:** September 3, 2024