Single-cell mean rank gene set scoring method for between-dataset comparison of scRNA-seq data

Dazhou Li[1], Guohao Meng[2*], Jieyu Wu[1], Guihui Tong[1]

1 Department of Pathology, First Affiliated Hospital of Guangzhou Medical University, Guangzhou, 510150, China

2 Department of Pathophysiology, Key Laboratory of Cell Differentiation and Apoptosis of Chinese Ministry of Education, Shanghai Jiaotong University School of Medicine, Shanghai, 200025, China

* Correspondence author

Email: mgh0525@outlook.com

Abstract

The surge in single-cell RNA sequencing (scRNA-seq) data offers a unique chance for researchers to understand functional changes in biological processes and diseases through gene set scoring across diverse datasets. Despite this, current methods for comparing scRNA-seq data at the signaling pathway level across datasets remain untested. To bridge this gap, we introduce the single-cell mean rank gene set scoring (scMRGSS) method, which assesses gene set activity between different scRNA-seq datasets. Leveraging gene expression ranks within each dataset, scMRGSS calculates mean rank scores for gene sets, enabling the comparison of their relative enrichment or depletion across datasets. Demonstrating its efficacy through simulated and real datasets, scMRGSS proves to be a simple yet informative tool for comparing gene set activity between cell types across diverse datasets. Its robustness against sequencing depth and dropout rate variations underscores its value for integrative scRNA-seq data analysis. Applying the method, we uncover that abnormal activity in oxidative phosphorylation and NF-κB signaling pathways in glioblastoma cancer cells may not solely stem from neurodevelopmental programs. Notably, the highest activity of these

pathways is observed in the mesenchymal cancer cell type, emphasizing the need to target specific cell types in glioblastoma drug development.

Keywords: single-cell analysis, RNA sequencing, gene set analysis, glioblastoma, NF-κB pathway

Introduction

Single-cell RNA sequencing (scRNA-seq) has revolutionized various fields of biological research by enabling the comprehensive profiling of gene expression at the individual cell level. Over the past decade, a substantial amount of scRNA-seq data has been accumulated through collaborative efforts in studying diverse biological systems and diseases(1–5). While scRNA-seq data offers several advantages in terms of scale and resolution, it presents unique challenges for data analysis and interpretation due to its noisy and zero-inflated nature(6). These challenges arise from inherent biological phenomena and the limited capture efficiency of the technology and are therefore difficult to overcome(7). As a result, specialized analytical methods are necessary to effectively process scRNA-seq data and extract meaningful biological insights.

A particularly noteworthy approach is the aggregation of gene expression profiles into biologically functional representations, commonly referred to as gene set analysis. This method aims to better comprehend the biological relevance of scRNA-seq data. A plethora of tools have been developed to estimate the activity of *a priori* gene sets or pathways based on scRNA-seq data, including single-cell signature explorer (SCSE)(8), AUCell(9), single-cell gene set enrichment analysis (scGSEA)(7), variance-adjusted Mahalanobis (VAM)(10), Pagoda2(11), and Vision(12). These tools enable researchers to assess the enrichment or depletion of specific gene sets in different cell populations, thereby providing insight into their functional differences. Considering the rapid increase in scRNA-seq data, it has become appealing to compare gene set activity underlying essential biological processes across various scRNA-seq datasets to identify commonalities and differences in cellular states. Regrettably, existing single-cell gene set scoring methods have yet to address this issue.

In this work, we present the single-cell mean rank gene set scoring (scMRGSS), a straightforward method for comparing gene set activity in single-cell RNA sequencing (scRNA-seq) data across different datasets. The scMRGSS method calculates a rank-based score for each gene set in a single cell mainly by averaging the ranks of genes belonging to that gene set. This simple approach facilitates easy interpretation of results, and by utilizing rank-based scores instead of raw or transformed expression values, it accounts for the noisy nature of scRNA-seq data. Our study demonstrates that scMRGSS is a reliable and robust method for comparing gene set activity between scRNA-seq datasets, as evidenced by simulations and real scRNA-seq datasets of cell lines and peripheral blood. Additionally, we applied the method to cancer biology and found that the up-regulation of oxidative phosphorylation and NF-κB signaling pathways in glioblastoma did not sorely reflect its resemblance to normal neurodevelopmental lineages, and there was notable heterogeneity across glioblastoma subtypes.

Methods

Single-cell mean rank gene set scoring

The scMRGSS method, which is adopted from the algorithm developed by Noureen et al.(13), calculates the normalized mean gene rank for each gene set among expressed genes in a cell, resulting in a score that ranges from 0 to 1. This method is robust to commonly used scRNA-seq normalization methods, such as counts per million (CPM), due to its reliance on gene ranks. In this study, scMRGSS was used to compare the scores between two datasets from different datasets in order to determine if there is a significant difference in the activity of a gene set between the datasets.

The method begins with basic data filtering, in which each dataset is subject to filtering of cells and genes in order to reduce technical and biological noise. The intersection of genes between the two datasets is selected as the input for the method. Gene sets are represented as a list of HUGO gene symbols, and those with less than the threshold proportion of expressed genes in a dataset (typically

0.6 in this study) are ignored for further procedure. Only expressed genes (with a count or expression value greater than 0) are considered for the calculation in order to minimize the influence of dropouts across cells in the same dataset. The genes are then ranked in each cell, and the rank-based score for each gene set is computed using the following formula:

$$score^{(s,c)} = \frac{\sum_{i=1}^{m} rank_i^{(s,c)}}{m \cdot n}$$

where m is the number of expressed genes of gene set s in cell c, and n is the total number of expressed genes in this cell. The metric is essentially the effective mean rank normalized by the number of expressed genes, and the normalization procedure allows for more comparable scores across cells.

Simulation of single-cell RNA-seq datasets and gene sets

The zero-inflated negative binomial (ZINB) model was employed to generate simulated single-cell RNA sequencing (scRNA-seq) datasets comprised of 4000 cells divided into four groups, each with varying library sizes or dropout rates. The VGAM R package was used to execute the simulation processes(14). Each group of cells expressed background genes and group-specific genes from eight gene sets, the sizes of which ranged from 50 to 120, with increments of 10 for both simulation scenarios. All genes not specific to the group were considered as background genes. The datasets, which corresponded to individual parameterisations, thus consisted of 2720 genes in total. The negative binomial (NB) distribution can be viewed as a gamma-Poisson mixture where the lambda parameter of Poisson distribution is distributed as a gamma distribution parameterised by shape and scale. For the library size simulation, the shape (size) parameter of the rzinegbin function was adjusted from 4 to 8 in increments of 1, while the munb parameter equalled product of shape and scale, where the scale parameter remained fixed at 3. The pstr0 parameter, which represents the probability of structural zero in the ZINB distribution, was set at 0.5. In the scenario of the dropout simulation, the shape and scale parameters were maintained at fixed values of 4 and 3, respectively.

Meanwhile, the pstr0 parameter varied from 0.2 to 0.8 in increments of 0.15. Similar to the distribution for group-specific genes in the dataset, background gene counts were also simulated using the ZINB distribution, of which the shape parameter was the same, the scale parameter was scaled by 0.5, and the pstr0 parameter was increased by 0.3 and bounded above by 0.9.

Performance evaluation

Simulation of scRNA-seq datasets was performed to evaluate the capacity of the proposed method in producing distinct gene set scores for different biological groups. The Splatter R package's splatSimulateGroups function(15) was utilised to generate datasets with two biological groups. The simulated dataset encompassed 4000 cells and 5000 genes, with approximately 20% of the genes designated as differentially expressed genes (DEGs). The DEGs exhibited a 50% probability of being down-regulated. Subsequent to dataset generation, random sampling procedures were employed to construct five distinct collections, each comprising 80 gene sets. These gene sets varied in size, ranging from 50 to 145 genes, with incremental steps of 5 genes. Importantly, the collections were distinguished by varying percentage of DEGs, spanning from 20% to 100%. For each gene set size, four gene sets were systematically generated using a consistent procedure for both biological groups. Specifically, within each group and gene set size, a group-specific gene set and a parallel group-unspecific gene set were simulated. The group-specific gene set was composed of group-specific genes sampled from the group-specific gene pool, along with background genes sampled from the group-unspecific gene pool. In contrast, the group-unspecific gene set solely comprised group-unspecific genes.

To quantitatively assess the performance, a diagnostic metric denoted as *d* was introduced, defined by the following equation:

$$d = \frac{((-\log 10(adj.p)/2) + max\{fc, 1/fc\})}{2}$$

where *adj.p* is the adjusted p-value, and *fc* is the fold change (see below). Subsequently, this metric *d* was employed to construct receiver operating characteristic (ROC) curve and precision-recall

(PR) curve for each DEG percentage using the EvalMetrics Julia package (https://github.com/VaclavMacha/EvalMetrics.jl). The objective was to meticulously evaluate the efficacy of scMRGSS in accurately classifying differences in gene set scores as either true or false under varying noise conditions, offering insights into the nuanced performance dynamics.

Public single-cell RNA sequencing datasets and gene sets

We utilised nine public scRNA-seq datasets in this study. These datasets were obtained from a variety of sources, including previous research studies and public databases. Four separate datasets consisting of Jurkat cells, 293T cells, a 50/50 mixture of Jurkat cells and 293T cells, a 99/1 mixture of Jurkat cells and 293T cells were obtained from a previous study(16). PBMC 10k 3p and PBMC 10k 5p datasets were downloaded from the dataset portal of 10x Genomics. Another human peripheral blood dataset was obtained from the Tabular Sapiens project(1). The glioblastoma and associated neural development datasets were obtained from previous study(17). Finally, another human fetal brain development dataset(18) was downloaded from the GEO database with accession number GSE162170. All dataset sources are listed in the Data Availability section. Furthermore, we obtained the Biocarta gene set collection, which consists of 292 gene sets, and the oxidative phosphorylation gene set of KEGG (v2023.1) from the MSigDB database(19).

Pre-processing of single-cell RNA sequencing datasets

The raw counts of scRNA-seq datasets were utilized as input for the method. For the human-derived datasets, we applied filters to exclude genes with numbers of expressed cells less than 2, and cells with numbers of expressed genes less than 2000. The threshold for filtering cells was determined based on the distribution of the number of genes for the simulated datasets. These procedures were pre-implemented in the scMRGSS Julia package and easily executable. We utilized pre-defined cell type labels when possible, and manually curated the datasets otherwise.

For the PBMC 10k 3p and BMC 10k 5p datasets, we employed the Seurat R package to cluster cells separately(20). We first removed low-quality cells based on the number of expressed genes and mitochondrial gene expression percentage, resulting in 8,827 and 9,894 cells in the two datasets, respectively. The sctransform method was applied to normalize the data(21), followed by linear dimensional reduction and cell clustering using the Louvain algorithm(22). The identified cell clusters were then embedded in non-linear low-dimensional space using the uniform manifold approximation and projection (UMAP) algorithm to better display local relationships. Next, we conducted cluster biomarker identification through gene differential expression analysis and compared the results with reference cell types to curate the cell clusters. We employed the SingleR package for this comparison(23). Cell labels were consistent among three reference sources for most identified cell clusters in the PBMC 10k 3p dataset (Supplementary Fig. S2A-C) and the PBMC 10k 5p dataset (Supplementary Fig. S2E-G), and cell labels were assigned in both datasets accordingly (Supplementary Fig. S2D,H).

The glioblastoma dataset contained curated cell labels, although they were incomplete. Only cells derived from the whole tumor samples were included in the study. The cells without labels and duplicates were excluded. In total, 18,430 cells were selected. The fetal brain cells at 24 post-conceptional weeks (pcw) were excluded from the brain development dataset (GSE162170)(18) to match the donor age in the fetal brain dataset associated with the glioblastoma study(17).

Comparison of gene set scores between two groups

To compare gene set scores between two groups, whether from the same dataset or not, we leveraged the results of hypothesis testing and the fold change. By default, the two-tailed Mann-Whitney U test implemented by the HypothesisTests Julia package (https://github.com/JuliaStats/HypothesisTests.jl) was applied to determine if the observed gene set score difference was statistically significant. The Bonferroni multiple hypothesis correction implemented by the MultipleTesting Julia package

(https://github.com/juliangehring/MultipleTesting.jl) was used to adjust p-values pertaining to the tested gene sets. An adjusted p-value of less than 0.01 was considered indicative of statistical significance. The gene set score difference was deemed biologically meaningful if the adjusted p-value was less than 0.01 and the fold change or its reciprocal was greater than 1.2 in the context of simulation study and applications in Jurkat and 293T cells as well as human peripheral blood cells. For the application in glioblastoma, the Bonferroni procedure was utilised to correct p-values for multiple group pair-wise comparisons regarding the selected gene set and the fold change threshold was set at 1.1. We also calculated the proportion of meaningfully differential gene sets group pair-wise and displayed the results on a heatmap using the Makie Julia package(24) for comparisons related to the cell lines and peripheral blood cells. Volcano plots were created to identify differential gene sets between the two groups using the EnhancedVolcano R package (https://bioconductor.org/packages/EnhancedVolcano). Venn diagrams were created to illustrate the overlapping differential gene sets using the ggVennDiagram R package(25).

Results

scMRGSS accurately identifies cell group-specific gene sets in simulation study

Different scRNA-seq preparation kits and sequencing platforms often produce data with different library sizes and dropouts levels. To assess the capacity of the method to detect difference in gene set activity between different conditions, we sought to perform simulation studies on the effects of library size and dropout rates on the performance of scMRGSS. Firstly, five scRNA-seq count datasets with various library sizes (or total count numbers) were simulated based on the ZINB model. Each dataset was composed of 2720 genes and 4000 cells, which were divided into four groups. Eight gene sets with increasing number of genes were simulated at the same time exclusively for one group. Density plots verified that the simulated datasets varied in terms of library size, while numbers of genes were comparable (Supp. Fig. S1A,B). As expected, cells within the same group showed higher similarity in the low-dimensional embedding and were clustered

together (Fig. 1A). Library size had a minor effect on cell clustering compared to the group condition (Fig. 1B). Each group of cells specifically expressed genes from the corresponding simulated gene sets (Fig. 1C). We combined the adjusted p-value and fold change to determine whether the detected differential gene sets were meaningful in the sense of biological context (see Methods). Proportions of biologically meaningful differential gene sets were larger in the same-group pairs than in the different-group pairs independent of library size, indicating that scMRGSS accurately identifies cell group-specific gene sets despite the challenge of shrinking library size (Fig. 1D).

Additionally, we conducted the simulation study on the impact of dropout rates using the ZINB model likewise. As dropout rates increased in the simulated datasets, both the total number of counts and the number of genes decreased (Supp. Fig. S1C,D). Group condition dominated the cell clustering compared to the dropout rate (Fig. 1E,F). Each cell group showed unique gene set activity pattern, with group exclusive gene sets only expressed in their respective groups (Fig. 1G). Proportions of differential gene sets were close to zero between the same groups from two datasets, and approximated 50% between different groups, in line with the theoretical values (Fig. 1H). This pattern was consistent regardless of the dropout levels, highlighting the robustness of scMRGSS in accurately identifying cell group-specific gene sets even in the presence of varying dropout rates. Taken together, these results indicated that scMRGSS is a robust method for detecting difference in gene set activity in scRNA-seq data in the simulation settings.

Simulation study demonstrates favourable performance of scMRGSS in yielding distinct scores for distinct conditions

We endeavoured to evaluate the method's capability to call true differential gene sets between two conditions using simulated datasets, taking into account the trade-off between sensitivity and specificity, as well as the trade-off between precision and recall. To assess the robustness of the method against noise, we incorporated varying proportions (0, 0.2, 0.4, 0.6, 0.8) of random

background genes into the group-specific gene sets. A diagnostic metric, integrating the adjusted p-value and fold change (see Methods), was developed to predict differential gene sets between the two conditions. We found that the 4000 simulated cells from the two groups displayed considerable overlap in a low-dimensional linear space, indicating a significant similarity between the two conditions (Fig. 2A). The scMRGSS demonstrated satisfactory sensitivity while maintaining high specificity even in the presence of 80% noise in the simulated gene sets (Fig. 2B). The PR curve is more informative than the ROC curve in imbalanced scenarios(26). The area under the PR curve surpassed 0.9 at all levels of noise (Fig. 2C), indicating that scMRGSS consistently exhibited high precision and recall in calling differential gene sets between the two conditions. Collectively, these findings demonstrate the robustness and favourable performance of scMRGSS in generating different gene set scores for two groups on simulated datasets.

scMRGSS differentiates biological conditions in human cell line and peripheral blood datasets

In order to evaluate the utility of scMRGSS for comparing gene set activity between two datasets in real-world situations, we applied the method to both human cell line and peripheral blood datasets. We used scMRGSS to calculate Biocarta gene set scores for Jurkat cells and 293T cells from four separate datasets and then determined the proportions of meaningful differential gene sets between each group pair. Our results revealed that the proportions of differential gene sets were substantially lower between same-cell groups compared to different-cell groups, indicating that scMRGSS effectively distinguishes between biological conditions (Fig. 3A). Furthermore, we examined the overlap of detected differential gene sets and found that only two differential gene sets overlapped between Jurkat cells from the Jurkat datasets and Jurkat cells from the 50/50 mixture dataset out of all dataset pairs of Jurkat cells (Fig. 3B). On the other hand, nine differential gene sets were shared among the four Jurkat-293T dataset pairs, and the majority of differential gene sets for a dataset pair overlapped with at least one dataset pair, suggesting consistent and reproducible differentiation between cell lines using scMRGSS (Fig. 3C).

We utilized a comparable analytical approach to assess the efficacy of scMRGSS in human peripheral blood datasets. This analysis encompassed CD14+ monocytes, CD16+ monocytes, naive CD4+ T cells, and B cells from three separate scRNA-seq sources generated by 10x Genomics microfluidic droplet and Smart-seq2 techniques(1). Similarly, we calculated the proportions of differential Biocarta gene sets between pairwise groups from the same or different datasets and observed similar patterns. In general, the proportions of differential gene sets were lower between same-cell groups compared to different-cell groups. Additionally, CD14+ monocytes and CD16+ monocytes shared a higher degree of similarity regarding this proportion compared to other cell group pairs (Fig. 3D), consistent with their intrinsic biological similarity and cell clustering results (Supplementary Fig. S2). The largest proportion of differential gene sets was observed between CD14+ monocytes and naive CD4+ T cells (Fig. 3D). We further investigated the detected differential gene sets in greater depth and found that the majority of differential gene sets were unique to each dataset pair in the cases of comparisons involving naive CD4+ T cells (Fig. 3E-H). However, when comparing CD16+ monocytes and naive CD4+ T cells, there was a higher overlap of differential gene sets among the four dataset combinations (Fig. 3I-M), indicating that the identified gene set score difference was consistent across different sources of scRNA-seq data. Given the above, our results illustrate that scMRGSS is effective in identifying differential gene sets in various cell types between scRNA-seq datasets.

scMRGSS reveals cancer cell heterogeneity of oxidative phosphorylation and NF-κB pathway activity in glioblastoma

Identifying differential gene sets and revealing potential heterogeneity in cellular activity across different cell types and sources can be of great utility to researchers. This is particularly true when comparing gene set activity in cells from various sources. One such application is the analysis of scRNA-seq datasets from glioblastoma samples. A previous study revealed that glioblastoma cancer cells resembled the human fetal brain lineages, including truncated radial glia (tRG), glial

progenitor cells (GPCs), oligo-lineage cells (OPCs), and interneurons, and hence the cancer cells were classified into mesenchymal, glial-progenitor, oligo-lineage, and neuronal subtypes(17). Another study uncovered four cell states and subtypes in glioblastoma along neurodevelopmental and metabolic axes, of which the mitochondrial cell state depends on the oxidative phosphorylation pathway for energy production(27).

NF-κB pathway is active in glioblastoma and plays an essential role in tumour progression and treatment resistance(28,29). We were interested in investigating if the up-regulated activity of oxidative phosphorylation and NF-κB pathways in glioblastoma could be explained by its resemblance to neurodevelopmental lineages and further if heterogeneity across cancer cell types exists regarding the activity of these pathways. In addition to the glioblastoma and neural development datasets included in the study of Couturier et al.(17), we incorporated another human brain development dataset(18) in the comparisons of gene set scores to demonstrate the consistency of our findings. The difference in gene set scores was viewed as significant in the biological sense if the related adjusted p-value was below 0.01 and the fold change exceeded 1.1. Our analysis revealed that gene set scores of the oxidative phosphorylation pathway were comparable between tRG from two neural development datasets (Fig. 4A). Notably, the gene set score of the mesenchymal cell type was higher when compared to tRG, whereas the scores of the other cancer cell types exhibited rising trends. When the scores of various cancer cell types were compared to those of GPCs and OPCs, respectively, this pattern remained (Supplementary Fig. S3A,B). This suggests that the up-regulated activity of the oxidative phosphorylation pathway in glioblastoma could not be merely attributed to its resemblance to neurodevelopmental lineages. Furthermore, the analysis revealed significant cellular heterogeneity within glioblastoma tumours with regard to the activity of the oxidative phosphorylation pathway.

The NF-κB pathway exhibited different patterns of activity across the various cancer cell types of glioblastoma. While tRG from two sources showed comparable scores of NF-κB pathway activity, the scores were significantly higher in the mesenchymal subtype compared to tRG (Fig. 4B).

However, the activity of the NF-κB pathway was lower in the neuronal subtype than in tRG, while the oligo-lineage and glial progenitor subtypes remained comparable to tRG. All other subtypes showed lower activity of the NF-κB pathway compared to the mesenchymal subtype, indicative of notable heterogeneity within glioblastoma with regard to the activity of the NF-κB pathway (Fig. 4B). Using GPCs and OPCs as the reference groups, respectively, we found similar patterns (Supplementary Fig. S3C,D). Collectively, these findings suggest that the up-regulated activity of the oxidative phosphorylation and NF-κB pathways in glioblastoma are not solely due to their resemblance to neurodevelopmental lineages. Moreover, the results also demonstrate the consistency of the method in comparing gene set scores between datasets.

Discussion

In recent years, the volume of scRNA-seq data has increased significantly across various research domains. This surge in data has made it feasible to compare the activity of gene sets of interest between different cell types across datasets. However, existing methods have not yet been tested to address this issue. In this study, we introduced scMRGSS, a simple yet effective approach for performing between-dataset comparisons of gene set activity using scRNA-seq data. The reliance of scMRGSS on the gene rank rather than the expression value confers it with greater robustness against commonly adopted normalisation and transformation methods, such as CPM, log transformation and z-score normalisation. This simplifies the analysis pipeline and enhances the applicability of the method to diverse datasets. Additionally, the normalisation of mean rank facilitates more consistent comparisons of cells between datasets with different scales of gene expression. We demonstrated the potential of scMRGSS in cancer research by identifying differences in the activity of oxidative phosphorylation and NF-κB pathways between glioblastoma cancer cell types and tRG, as well as highlighting the cellular heterogeneity within glioblastoma concerning these pathway activity.

The efficacy of the method in identifying disparate gene sets between biological conditions while concurrently regulating false positive rates was demonstrated through both simulated and genuine data analysis. The performance of scMRGSS is contingent upon the premise that the technical bias of mRNA capture and amplification between the two data sources remains consistent or varies only marginally. To elucidate, when the biological variations between the two datasets are equivalent, the discrepancy in gene rank on average is negligible. This assumption is generally applicable to data derived from the 10x Genomics and Smart-seq2 platforms, particularly when obtained from the same platform. Nonetheless, when employing the method to compare pathway activity of two cell types from two datasets generated by distinct scRNA-seq protocols or platforms, it would be prudent to include one shared cell type between the two datasets to assess the validity of the assumption. In such cases, single-cell dataset integration approaches, such as Scanorama(16) an scGen(30), may be required to account for the technical bias. In light of the challenge in establishing strict rules for determining thresholds in adjusted p-values and fold changes, the interpretation of disparities in gene set scores demands meticulous attention. For instance, assuming the difference in gene set scores remains constant, larger group sizes often yield smaller p-values. Given that gene set scores range between 0 and 1, the upper limit for the fold change must be constrained by a value depending upon the reference score. Accumulating datasets from analogous sources will contribute valuable insights towards identifying pattern of gene set activity difference. scMRGSS is designed for gene set analysis following the clustering and labelling of cells within datasets. Although it is a non-parametric and rank-invariant method, it is essential to note that basic gene and cell filtering is necessary to ensure the reliability and accuracy of the analysis results. This is because noise in the data could potentially impact the gene rank(31) and the mean and dispersion of gene set scores within a cell type when performing hypothesis testing. As a general guideline, it is advisable to set the filtering threshold based on the distribution of the data and *a priori* knowledge about the cells of interest. For datasets generated by common scRNA-seq platforms such as 10x Genomics and Smart-seq2, we typically recommend removing genes with the number of

cells below 2 and removing cells with the number of expressed genes below 2000. Additionally, the percentage of mitochondrial genes may be integrated into the filtering scheme to account for potential cell stress or contamination. Furthermore, it is recommended to include gene sets whose percentage of expressed genes in both datasets exceeds some threshold, such as 0.6 in this study, to avoid identifying differential gene sets of little biological value depending on only a small subset of genes. The exact percentage may depend on both the gene set and the datasets.

In this study we also sought to investigate the relationships of the activity of oxidative phosphorylation and NF-κB pathways between neurodevelopmental lineages and cancer cell types in glioblastoma. Previous research has shown that glioblastoma cells epitomise the normal neurodevelopmental process and resemble several lineages during the process, including tRG, GPCs, OPCs and interneurons(17). It is established that most glioblastoma cancer cells produce energy through oxidative phosphorylation as opposed to glycolysis(32). The key role of oxidative phosphorylation in glioblastoma is also emphasized by the potent inhibitory effect of the oxidative phosphorylation inhibitor on glioblastoma cancer cells(33). Our findings suggest that the metabolic feature of the glioblastoma messenchymal cell type is not associated with the neurodevelopmental programs, as we observed elevated oxidative phosphorylation activity in the messenchymal cell type compared to brain development lineages. We speculate this cancer cell type may overlap with a recently identified mitochondrial subtype that depends exclusively on oxidative phosphorylation for energy production(27). Though not as strong as in messenchymal cells, other glioblastoma cancer cell types also showed increasing trends in oxidative phosphorylation activity, indicating glioblastoma cellular heterogeneity in energy metabolism. Aberrant NF-κB activation in glioblastoma contributes to cancer cell proliferation, invasion, mesenchymal differentiation, and resistance to radiotherapy(34). We found that NF-κB activity were significantly elevated in the messenchymal cell type of glioblastoma relative to neurodevelopmental lineages, but not other subtypes. This suggests that the neurodevelopmental programs hijacked by glioblastoma cannot fully explain NF-κB pathway activation in all cancer cell types. The heterogeneity of glioblastoma

also highlights the need for a more precise approach to target the messenchymal cancer subtype when developing drugs aimed at the NF-κB pathway in glioblastoma treatment.

In brief, we have presented a simple and efficient method to compare gene set activity amongst cell types from various scRNA-seq datasets. By applying this approach to glioblastoma, it has been discovered that the disease's aberrant oxidative phosphorylation and NF-κB pathway activity is not exclusively caused by neurodevelopmental programs. This finding has implications for precision medicine strategies aimed at addressing particular cancer subtypes.

Availability of data

All scRNA-seq datasets used in this study are available for download. We used the following public datasets:

pbmc_10k_3p dataset from 10x Genomics: https://cf.10xgenomics.com/samples/cell-exp/4.0.0/Parent_NGSC3_DI_PBMC/Parent_NGSC3_DI_PBMC_filtered_feature_bc_matrix.h5;

pbmc_10k_5p dataset from 10x Genomics: https://cf.10xgenomics.com/samples/cell-vdj/5.0.0/sc5p_v2_hs_PBMC_10k/sc5p_v2_hs_PBMC_10k_filtered_feature_bc_matrix.h5;

Tabular Sapiens dataset from figshare: https://figshare.com/articles/dataset/Tabula_Sapiens_release_1_0/14267219;

Glioblastoma dataset: https://github.com/mbourgey/scRNA_GBM;

Human neural development dataset: https://github.com/mbourgey/scRNA_GBM;

Human cortex development dataset: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE162170.

All human gene sets in the study were downloaded from MSigDB (https://www.gsea-msigdb.org/gsea/msigdb).

Simulated datasets have been deposited on figshare with the following DOI https://doi.org/10.6084/m9.figshare.24886047.

Availability of code

scMRGSS is developed in Julia programming language v1.9.3 (https://julialang.org/) and available at https://github.com/giuseppedelnapalle/scmrgss, while the scripts to reproduce results of the manuscript can be accessed on zenodo at the DOI https://doi.org/10.5281/zenodo.10418687.

Funding

This work was carried out independent of funding.

Acknowledgements

We would like to thank Jiawei Li for helpful questions and comments about the manuscript.

Author contributions

G.M. conceived the problem and the algorithm. D.L. and G.M. developed the method and performed the computational experiments. D.L., G.M., J.W. and G.T. wrote, reviewed and revised the manuscript. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing in interests.

References

1. THE TABULA SAPIENS CONSORTIUM. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science. 2022 May 13;376(6594):eabl4896.

2. Domínguez Conde C, Xu C, Jarvis LB, Rainbow DB, Wells SB, Gomes T, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science. 2022 May 13;376(6594):eabl5197.

3. Eraslan G, Drokhlyansky E, Anand S, Fiskin E, Subramanian A, Slyper M, et al. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. Science. 2022 May 13;376(6594):eabl4290.

4. Yao Z, Liu H, Xie F, Fischer S, Adkins RS, Aldridge AI, et al. A transcriptomic and epigenomic cell atlas of the mouse primary motor cortex. Nature. 2021 Oct;598(7879):103–10.

5. Jardine L, Webb S, Goh I, Quiroga Londoño M, Reynolds G, Mather M, et al. Blood and immune development in human fetal bone marrow and Down syndrome. Nature. 2021 Oct;598(7880):327–31.

6. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert JP. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun. 2018 Jan 18;9(1):284.

7. Franchini M, Pellecchia S, Viscido G, Gambardella G. Single-cell gene set enrichment analysis and transfer learning for functional annotation of scRNA-seq data. NAR Genom Bioinform. 2023 Mar;5(1):lqad024.

8. Pont F, Tosolini M, Fournié JJ. Single-Cell Signature Explorer for comprehensive visualization of single cell signatures across scRNA-seq datasets. Nucleic Acids Res. 2019 Dec 2;47(21):e133.

9. Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nat Methods. 2017 Nov;14(11):1083–6.

10. Frost HR. Variance-adjusted Mahalanobis (VAM): a fast and accurate method for cell-specific gene set scoring. Nucleic Acids Res. 2020 Sep 18;48(16):e94.

11. Lake BB, Chen S, Sos BC, Fan J, Kaeser GE, Yung YC, et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nat Biotechnol. 2018 Jan;36(1):70–80.

12. DeTomaso D, Jones MG, Subramaniam M, Ashuach T, Ye CJ, Yosef N. Functional interpretation of single cell similarity maps. Nat Commun. 2019 Sep 26;10(1):4376.

13. Noureen N, Ye Z, Chen Y, Wang X, Zheng S. Signature-scoring methods developed for bulk samples are not adequate for cancer single-cell RNA sequencing data. Elife. 2022 Feb 25;11:e71994.

14. Yee TW. The VGAM Package for Categorical Data Analysis. Journal of Statistical Software. 2010 Jan 5;32:1–34.

15. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol. 2017 Sep 12;18(1):174.

16. Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat Biotechnol. 2019 Jun;37(6):685–91.

17. Couturier CP, Ayyadhury S, Le PU, Nadaf J, Monlong J, Riva G, et al. Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. Nat Commun. 2020 Jul 8;11(1):3406.

18. Trevino AE, Müller F, Andersen J, Sundaram L, Kathiria A, Shcherbina A, et al. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. Cell. 2021 Sep 16;184(19):5053-5069.e23.

19. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011 Jun 15;27(12):1739–40.

20. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021 Jun 24;184(13):3573-3587.e29.

21. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019 Dec 23;20(1):296.

22. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech. 2008 Oct;2008(10):P10008.

23. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019 Feb;20(2):163–72.

24. Danisch S, Krumbiegel J. Makie.jl: Flexible high-performance data visualization for Julia. Journal of Open Source Software. 2021 Sep 1;6(65):3349.

25. Gao CH, Yu G, Cai P. ggVennDiagram: An Intuitive, Easy-to-Use, and Highly Customizable R Package to Generate Venn Diagram. Front Genet. 2021;12:706907.

26. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.

27. Garofano L, Migliozzi S, Oh YT, D'Angelo F, Najac RD, Ko A, et al. Pathway-based classification of glioblastoma uncovers a mitochondrial subtype with therapeutic vulnerabilities. Nat Cancer. 2021 Feb;2(2):141–56.

28. Ji J, Ding K, Luo T, Zhang X, Chen A, Zhang D, et al. TRIM22 activates NF-κB signaling in glioblastoma by accelerating the degradation of IκBα. Cell Death Differ. 2021 Jan;28(1):367–81.

29. Xiang J, Alafate W, Wu W, Wang Y, Li X, Xie W, et al. NEK2 enhances malignancies of glioblastoma via NIK/NF-κB pathway. Cell Death Dis. 2022 Jan 14;13(1):58.

30. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. Nat Methods. 2019 Aug;16(8):715–21.

31. Foroutan M, Bhuva DD, Lyu R, Horan K, Cursons J, Davis MJ. Single sample scoring of molecular phenotypes. BMC Bioinformatics. 2018 Nov 6;19(1):404.

32. Bonnay F, Veloso A, Steinmann V, Köcher T, Abdusselamoglu MD, Bajaj S, et al. Oxidative Metabolism Drives Immortalization of Neural Stem Cells during Tumorigenesis. Cell. 2020 Sep 17;182(6):1490-1507.e19.

33. Shi Y, Lim SK, Liang Q, Iyer SV, Wang HY, Wang Z, et al. Gboxin is an oxidative phosphorylation inhibitor that targets glioblastoma. Nature. 2019 Mar;567(7748):341–6.

34. Soubannier V, Stifani S. NF-κB Signalling in Glioblastoma. Biomedicines. 2017 Jun 9;5(2):29.

Figures and Legends

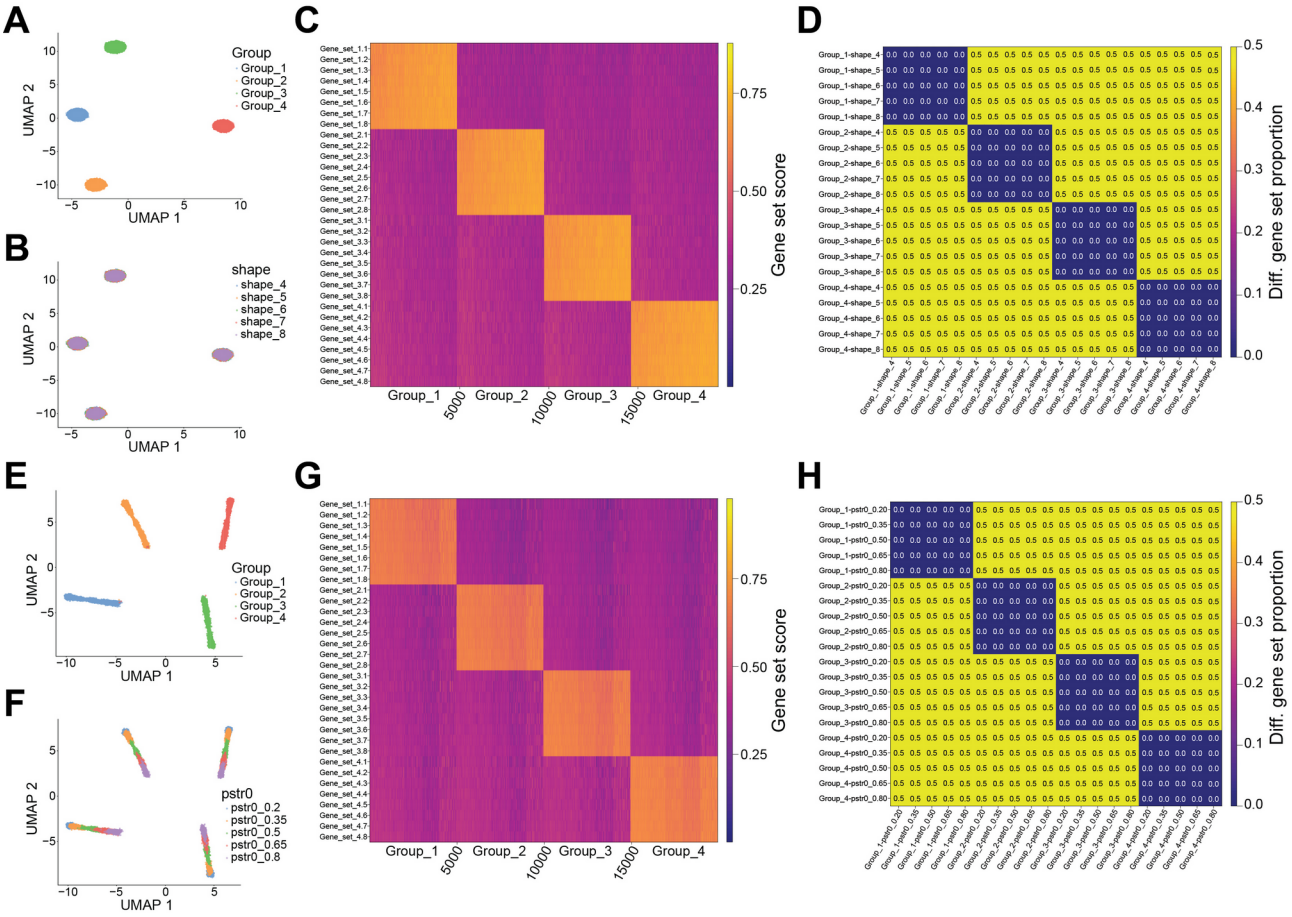
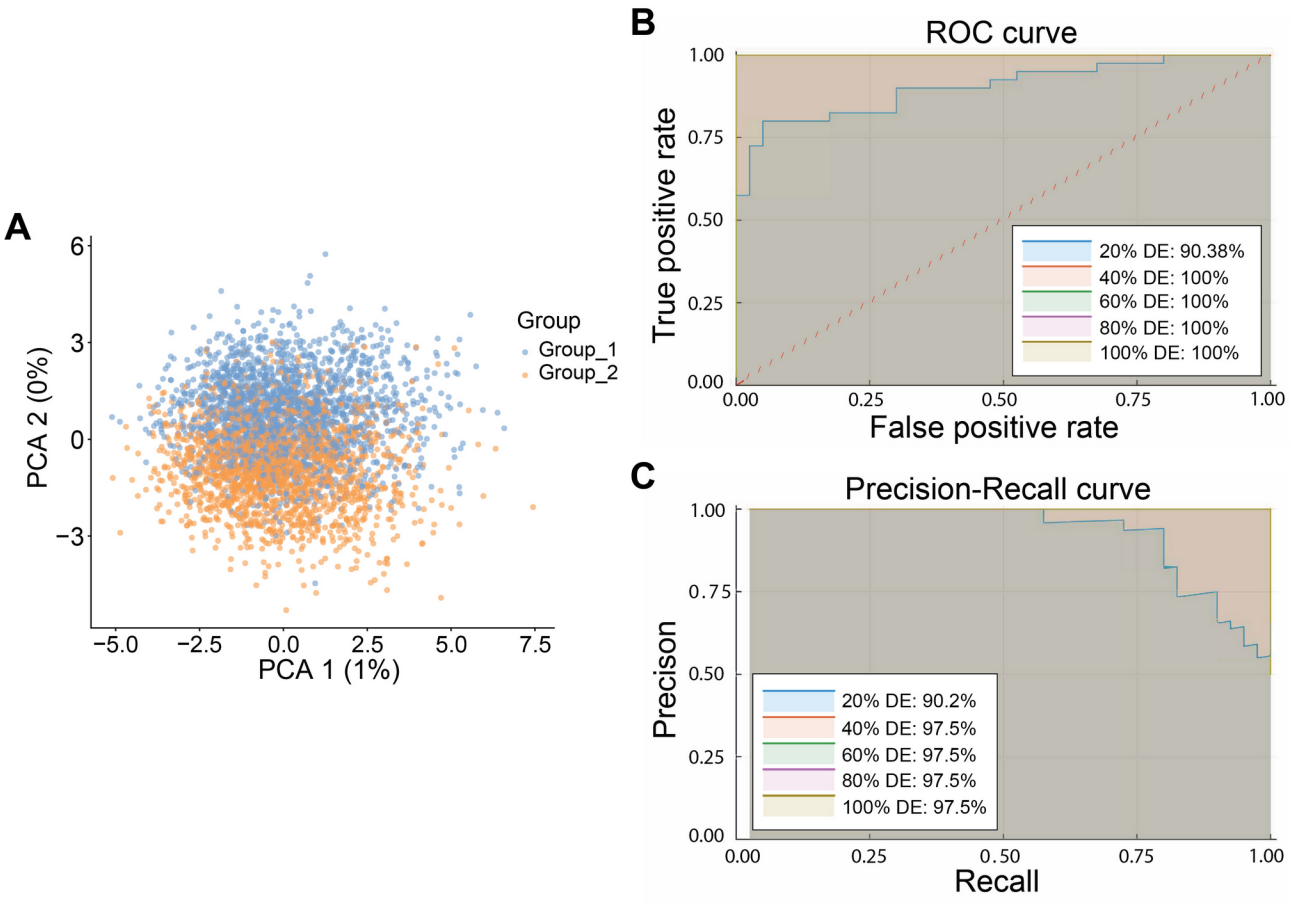
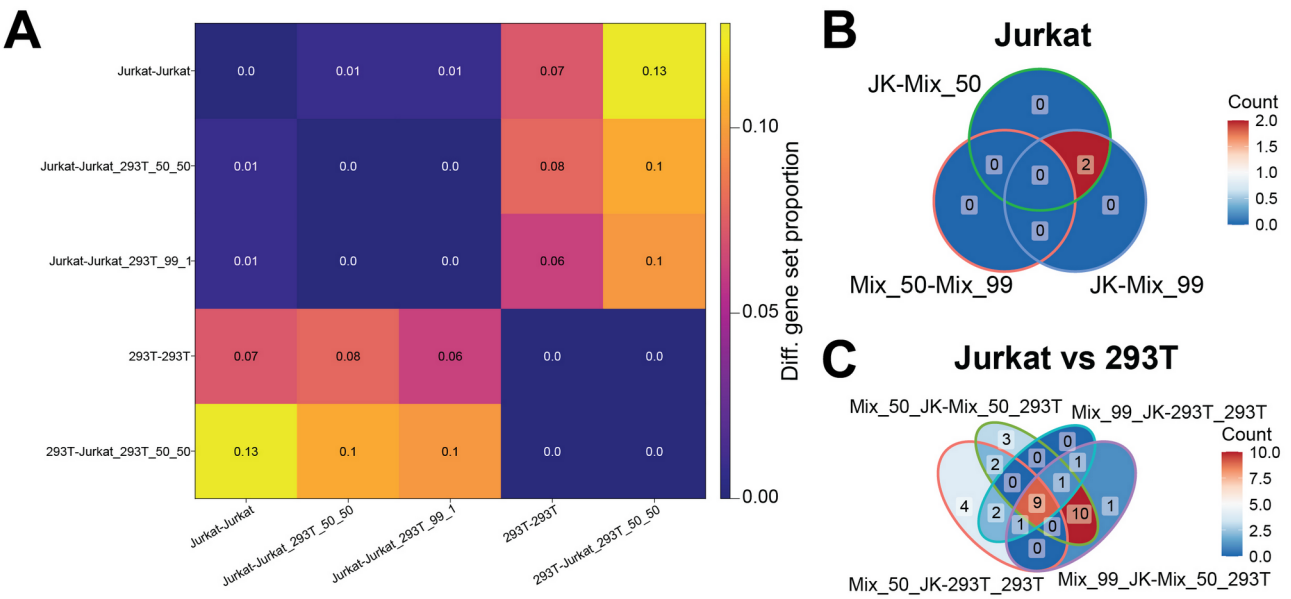


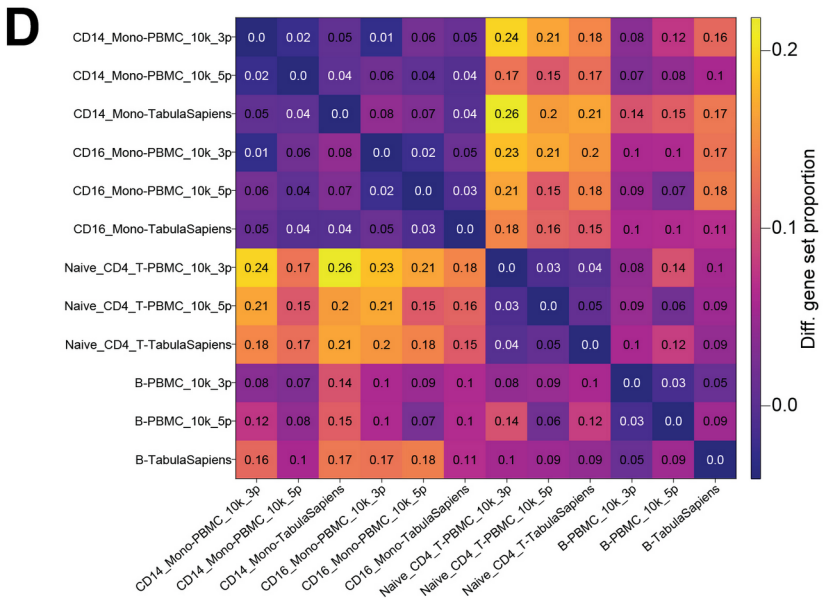
Figure 1. Simulation study of the single-cell mean rank gene set scoring (scMRGSS) method. (**A**,**B**) UMAP representation of 20,000 simulated cells categorised into four groups across five datasets with varying library sizes, labelled by group (**A**) and the shape (size) parameter (**B**) for the zero-inflated negative binomial (ZINB) model, respectively. Each shape parameter value corresponds to a distinct dataset. (**C**) Gene set activity of 32 simulated gene sets computed by scMRGSS across 20,000 cells from five distinct datasets as presented in (**A**,**B**). Rows represent gene sets while columns represent cells in the heatmap. The hierarchical clustering algorithm groups cells into four categories mirroring the simulated groups. (**D**) Heatmap illustrating the

proportions of biologically meaningful differential gene sets between groups from the same or distinct datasets as observed in (**A**,**B**). The difference in gene set activity is considered valid if the associated adjusted p-value is below 0.01 and the fold change or its reciprocal is larger than 1.2. (**E**,**F**) UMAP plots depicting 20,000 simulated cells categorised into four groups across five datasets with varying dropout rates, labelled by group (**E**) and the pstr0 parameter (**F**) for the zero-inflated negative binomial (ZINB) model, respectively. pstr0 controls the probability of structural zero in the ZINB model, and each pstr0 value corresponds to a different dataset. (**G**) Heatmap presenting the activity of 32 simulated gene sets computed by scMRGSS across 20,000 cells from five distinct datasets as shown in (**E**,**F**). Rows represent gene sets while columns represent cells. The hierarchical clustering algorithm categorises cells into four groups as expected. (**H**) Proportions of biologically meaningful differential gene sets between groups from same or distinct datasets as depicted in (**E**,**F**).



458 **Figure 2.** Performance evaluation of scMRGSS. (**A**) Principal component analysis (PCA) biplot of
459 4000 simulated cells from two groups. (**B**) Receiver operating characteristic (ROC) curve of five
460 classifiers associated with different differential expressed gene (DEG) percentage (ranging from
461 20% to 100%) of the group-specific gene sets. For each DEG percentage, 20 group 1-specific gene
462 sets, 20 group 2-specific gene sets, and 40 unspecific gene sets of varying sizes spanning from 50 to
463 145 genes were generated from the simulated single-cell RNA sequencing (scRNA-seq) dataset in
464 (**A**). A diagnostic metric *d* incorporating the adjusted p-value and the fold change was introduced to
465 formulate classifying models for each DEG percentage to predict the veracity of the observed
466 differences in gene set scores between the two groups. The ROC curve serves as a visual
467 representation of the performance of the classifying models. (**C**) Precision-recall (PR) curve of the
468 five classifiers as described in (**B**).
469


A
Jurkat-Jurkat
Jurkat-Jurkat_293T_50_50
Jurkat-Jurkat_293T_99_1
293T-293T
293T-Jurkat_293T_50_50
Diff. gene set proportion
B
Jurkat
JK-Mix_50
Mix_50-Mix_99
JK-Mix_99
Count
C
Jurkat vs 293T
Mix_50_JK-Mix_50_293T
Mix_99_JK-293T_293T
Mix_50_JK-293T_293T
Mix_99_JK-Mix_50_293T
Count


D
CD14_Mono-PBMC_10k_3p
CD14_Mono-PBMC_10k_5p
CD14_Mono-TabulaSapiens
CD16_Mono-PBMC_10k_3p
CD16_Mono-PBMC_10k_5p
CD16_Mono-TabulaSapiens
Naive_CD4_T-PBMC_10k_3p
Naive_CD4_T-PBMC_10k_5p
Naive_CD4_T-TabulaSapiens
B-PBMC_10k_3p
B-PBMC_10k_5p
B-TabulaSapiens
Diff. gene set proportion

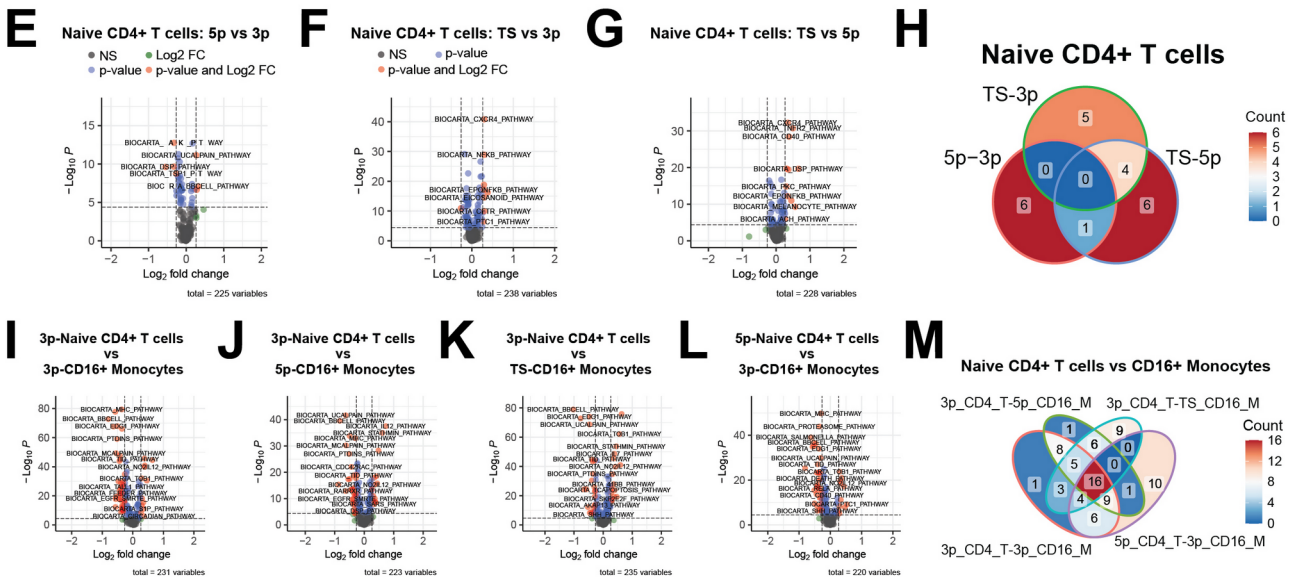

E
Naive CD4+ T cells: 5p vs 3p
NS
Log2 FC
p-value
p-value and Log2 FC
Log2 fold change
total = 225 variables
F
Naive CD4+ T cells: TS vs 3p
total = 238 variables
G
Naive CD4+ T cells: TS vs 5p
total = 228 variables
H
Naive CD4+ T cells
TS-3p
5p−3p
TS-5p
Count
I
3p-Naive CD4+ T cells vs 3p-CD16+ Monocytes
total = 231 variables
J
3p-Naive CD4+ T cells vs 5p-CD16+ Monocytes
total = 223 variables
K
3p-Naive CD4+ T cells vs TS-CD16+ Monocytes
total = 235 variables
L
5p-Naive CD4+ T cells vs 3p-CD16+ Monocytes
total = 220 variables
M
Naive CD4+ T cells vs CD16+ Monocytes
3p_CD4_T-5p_CD16_M
3p_CD4_T-TS_CD16_M
3p_CD4_T-3p_CD16_M
5p_CD4_T-3p_CD16_M
Count

Figure 3. scMRGSS distinguishes biological groups in real datasets. (**A**) Heatmap displaying the proportions of biologically meaningful differential Biocarta gene sets between cell lines from four datasets comprising Jurkat and/or 293T cells. scMRGSS was applied to a collection of four datasets consisting of one entirely of Jurkat cells (Jurkat), one entirely of 293T cells (293T), a 50/50 mixture of Jurkat and 293T cells (Jurkat_293T_50_50), and a 99/1 mixture of Jurkat and 293T cells (Jurkat_293T_99_1). Cell line names and Dataset name are split by a hyphen ("-") in each row or column. The difference in gene set scores is considered meaningful if the adjusted p-value is below 0.01 and the fold change or its reciprocal is larger than 1.2. (**B**) Venn diagram illustrating how differential gene sets between Jurkat cells from three different datasets as described in (**A**) overlap. (**C**) Overlap of differential gene sets between Jurkat and 293T cells from four different datasets. Mix_50: Jurkat_293T_50_50; Mix_99: Jurkat_293T_99_1; JK: Jurkat cells; 293T: 293T dataset or 293T cells. (**D**) Proportions of biologically meaningful differential Biocarta gene sets between human peripheral blood cells from the PBMC_10k_3p, PBMC_10k_5p and TabularSapienes datasets. The analysis takes into account four distinct cell groups: CD14+ monocytes, CD16+ monocytes, naive CD4+ T cells and B cells. An identical process was carried out as in (**A**). (**E**-**G**) Identification of differential gene sets of naive CD4+ T cells between PBMC_10k_5p and PBMC_10k_3p datasets (**E**), between TabularSapienes and PBMC_10k_3p datasets (**F**), and between TabularSapienes and PBMC_10k_5p datasets (**G**), respectively. (**H**) Overlap of differential gene sets between naive CD4+ T cells from three different datasets depicted in (**D**). (**I**-**L**) Identification of differential gene sets between naive CD4+ T cells and CD16+ monocytes from the three datasets. Pairwise comparisons of gene set activity were conducted between naive CD4+ T cells and CD16+ monocytes of the PBMC_10k_3p dataset (**I**), between naive CD4+ T cells of the PBMC_10k_3p dataset and CD16+ monocytes of the PBMC_10k_5p dataset (**J**), between naive CD4+ T cells of the PBMC_10k_3p dataset and CD16+ monocytes of the TabularSapienes dataset (**K**), between naive CD4+ T cells of the PBMC_10k_5p dataset and CD16+ monocytes of the PBMC_10k_3p dataset (**L**), respectively. (**M**) Overlap of differential gene sets between naive CD4+

T cells and CD16+ monocytes from the three datasets. 3p: PBMC_10k_3p; 5p: PBMC_10k_5p; TS: TabularSapienes; CD4_T: naive CD4+ T cells; CD16_M: CD16+ monocytes.

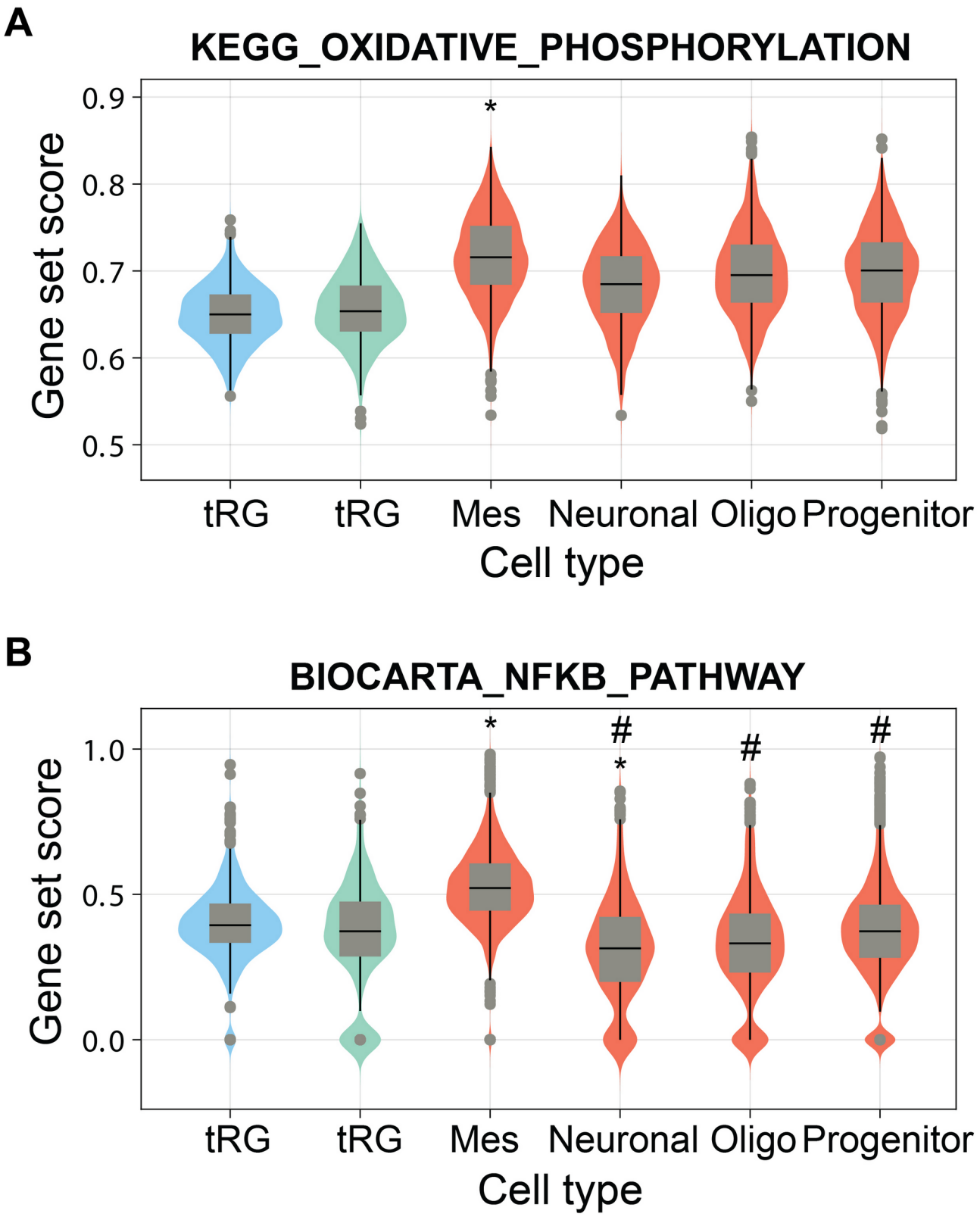


Figure 4. Application of scMRGSS in glioblastoma. (**A**-**B**) Gene set activity of the KEGG oxidative phosphorylation (**A**) and the Biocarta NF-κB (**B**) pathways estimated by scMRGSS in distinct cancer cell types of glioblastoma compared to truncated radial glia (tRG) of two datasets. tRG of the cortex development dataset (first column) and the messenchymal cancer cells are chosen as reference to calculate p-values. Bonferroni procedure was used to adjust the p-values. The difference in gene set scores is considered meaningful if the adjusted p-value is below 0.01 and the fold change or its reciprocal is larger than 1.1. * denotes the difference is meaningful when compared to tRG, whereas # denotes the difference is meaningful when compared to the messenchymal cancer cells.