Signature-scoring methods developed for bulk samples are not adequate for cancer single-cell RNA sequencing data

Nighat Noureen[1,2], **Zhenqing Ye**[1,2], **Yidong Chen**[1,2], **Xiaojing Wang**[1,2], **Siyuan Zheng**[1,2]*

[1]Greehey Children's Cancer Research Institute, UT Health San Antonio, San Antonio, United States; [2]Department of Population Health Sciences, UT Health San Antonio, San Antonio, United States

Abstract Quantifying the activity of gene expression signatures is common in analyses of single-cell RNA sequencing data. Methods originally developed for bulk samples are often used for this purpose without accounting for contextual differences between bulk and single-cell data. More broadly, few attempts have been made to benchmark these methods. Here, we benchmark five such methods, including single sample gene set enrichment analysis (ssGSEA), Gene Set Variation Analysis (GSVA), AUCell, Single Cell Signature Explorer (SCSE), and a new method we developed, Jointly Assessing Signature Mean and Inferring Enrichment (JASMINE). Using cancer as an example, we show cancer cells consistently express more genes than normal cells. This imbalance leads to bias in performance by bulk-sample-based ssGSEA in gold standard tests and down sampling experiments. In contrast, single-cell-based methods are less susceptible. Our results suggest caution should be exercised when using bulk-sample-based methods in single-cell data analyses, and cellular contexts should be taken into consideration when designing benchmarking strategies.

***For correspondence:** zhengs3@uthscsa.edu

Competing interest: The authors declare that no competing interests exist.

Funding: See page 9

Preprinted: 01 July 2021
Received: 06 July 2021
Accepted: 25 February 2022
Published: 25 February 2022

Reviewing Editor: C Daniela Robles-Espinoza, International Laboratory for Human Genome Research, Mexico

Editor's evaluation

In this revised manuscript, Noureen and collaborators benchmark five methods that are used in single-cell RNA sequencing data analyses, including single sample gene set enrichment analysis (ssGSEA), Gene Set Variation Analysis (GSVA), AUCell, Single Cell Signature Explorer (SCSE), and a newly developed method, Jointly Assessing Signature Mean and Inferring Enrichment (JASMINE). The authors test these in distinct cancer datasets and conclude that caution should be exercised when using bulk sample-based methods in single-cell data analyses, and cellular contexts should be taken into consideration. As a rapidly developing field in need of method benchmarking, we believe that this paper will be of great interest to the bioinformatics and single-cell data analysis communities.

Introduction

Single-cell RNA sequencing (scRNA-seq) is a powerful technology to study the ecosystems of normal and disease tissues. Gene expression signatures can be used to interrogate single cells for cell identities and other cellular properties (***Noureen et al., 2021a***) using signature-scoring methods. A key consideration for these methods is how to account for the variability of expression within the signature, a characteristic often associated with the overall expression variability of the interrogated sample. Despite widespread use and their benchmarking on certain applications (***Holland et al., 2020***; ***Zhang***

et al., 2020), limitations of these methods, including some initially developed for bulk samples, are incompletely understood in scRNA-seq analysis.

Compared with bulk-sample RNAseq, scRNA-seq data have high dropout rates (***Hicks et al., 2018***). Dropout rates are the opposite of gene counts, that is, the number of genes detected in a cell, a feature that is associated with cell differentiation status (***Gulati et al., 2020***). Variability in gene counts may cause imbalanced representation of non-expressed genes in gene signatures, resulting in scoring bias. Thus, we reason that in contexts where cells differ in differentiation status, failing to account for this variability may lead to misleading signature scores.

Cancer cells exhibit stem-cell like properties ***Malta et al., 2018***; thus, cancer may represent such a context. To test this, we benchmark five signature-scoring methods in cancer single-cell datasets. The five methods included bulk-sample-based methods single sample gene set enrichment analysis (ssGSEA) and Gene Set Variation Analysis (GSVA) (implemented in the R GSVA package ***Hänzelmann et al., 2013***) and three single-cell-based methods, AUCell (***Aibar et al., 2017***), Single-Cell Signature Explorer (SCSE) (***Pont et al., 2019***), and Jointly Assessing Signature Mean and Inferring Enrichment (JASMINE), a new method we developed. We show that cancer cells consistently express more genes than normal cells, and this imbalance significantly biases results from ssGSEA and GSVA but largely spares single-cell-based methods. Our results caution against the use of bulk-sample-based methods in scRNA-seq analyses.

Results

Cancer cells demonstrate higher gene counts than normal cells

Several recent studies showed that gene counts of single cells are associated with cell differentiation status and cell identity (***Gulati et al., 2020***; ***Qiu, 2020***). Specifically, cells higher in the differentiation hierarchy express more genes. To examine if cancer and normal cells demonstrate similar imbalances, we collected 10 scRNA-seq datasets across different cancer types, technological platforms and sequencing coverage (***Supplementary file 1***). We also obtained cell type annotations from the original studies. We found that the average number of detected genes was significantly higher in tumor cells than in normal cells across all datasets (***Figure 1A***, p < 2.2e-16). This imbalance persisted when we separated normal cells into different cell populations (***Figure 1—figure supplement 1***). However, cancer-associated cells, including cancer-associated fibroblasts (CAF), tumor-associated macrophages (TAM), and tumor-related endothelial cells (TEC) had higher gene counts than other normal cell types, sometimes even comparable to malignant cells (***Figure 1—figure supplement 1***).

Bias of ssGSEA signature scores in cancer datasets

The reliability of signature-scoring methods is measured by their ability to identify the quantitative differences between two groups of samples at the gene-signature level. Here, we compare signature scores between tumor and normal cells. One issue is that scores from these methods differ in range and variance, thus making direct comparison challenging. To overcome this challenge, we used Cohen's d (***Cohen, 1988***), a measurement that normalizes mean differences with standard deviation, thus generating unitless contrast that can be compared across methods. For simplicity, we defined Cohen's d greater than one as upregulated, and less than –1 as down regulated. These criteria were used throughout this work.

We assembled 7503 gene sets from MSigDB that each has at least 20 genes (The term 'gene set' is used interchangeably with 'gene signature'; Methods). On average, single-cell methods reported similar proportions of up- and down-regulated gene sets (9% vs 12% of all input gene sets) (***Figure 1B***). In contrast, ssGSEA and GSVA reported 25 and 23% of the input gene sets as upregulated but only 6 and 5% as down regulated. In six out of the 10 datasets, ssGSEA and GSVA estimated more than twice as many upregulated gene sets than down gene sets. This pattern remained when we divided normal cells into different populations (***Figure 1—figure supplement 2***). Notably, in cell types with high gene counts such as TEC, TAM, and CAF, we did not observe significantly more upregulated gene sets by ssGSEA and GSVA (***Figure 1—figure supplement 2E,F***). These results collectively suggest ssGSEA and GSVA are sensitive to variability of gene counts common to scRNA-seq data.

We also found that Cohen's d from ssGSEA and GSVA showed consistent positive correlation with gene set sizes (***Figure 1C***). Single-cell methods did not demonstrate this pattern.

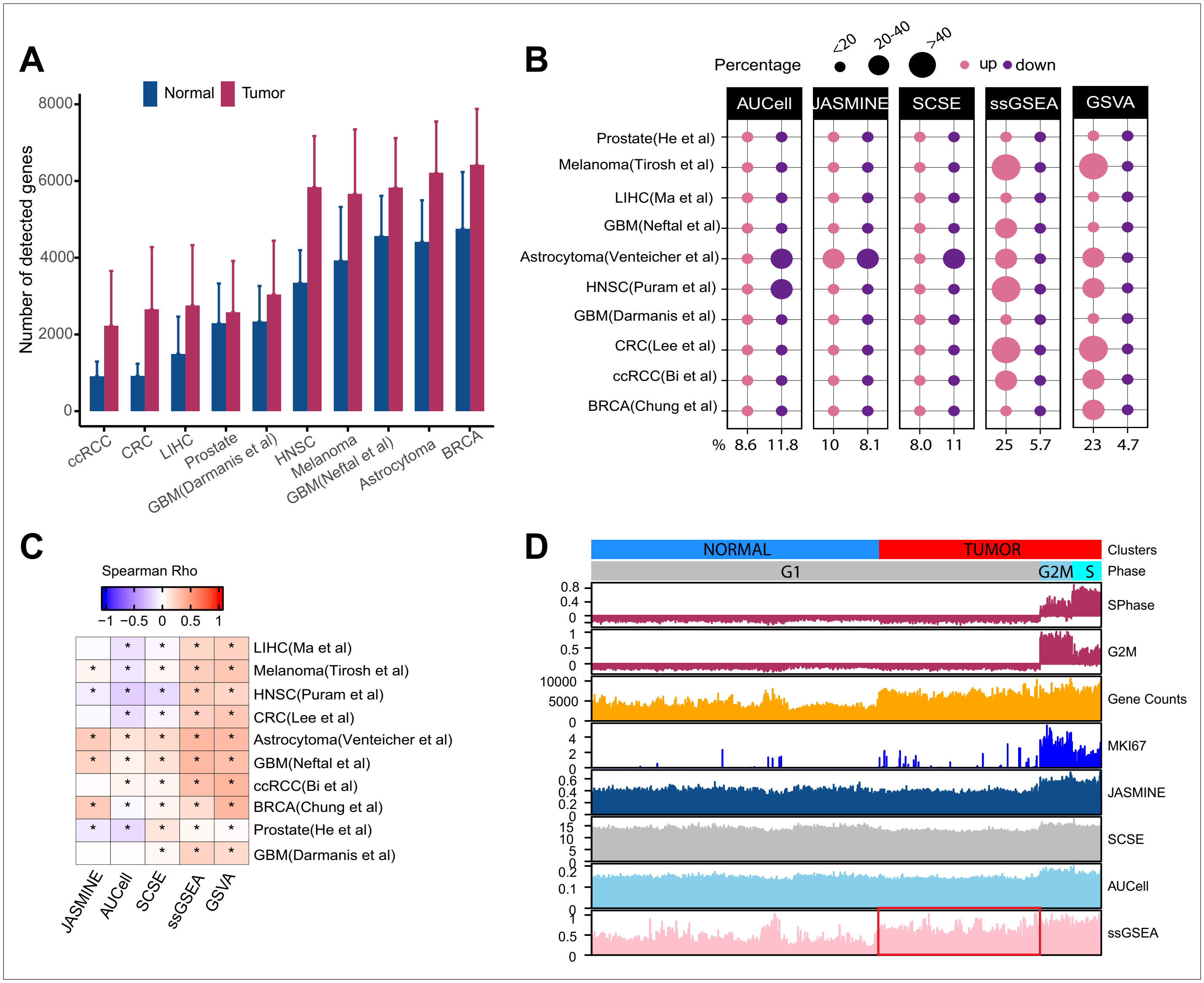

Figure 1. Gene count imbalances affect signature scoring. (**A**) The number of detected genes in tumor and normal cell populations in 10 single cell cancer RNAseq datasets. The height of each bar represents average, and whiskers represent standard deviation. In all cases, the difference is statistically significant (student t test, p < 2.2e-16). (**B**) Percentage of up and down regulated gene signatures in cancer cells relative to normal cells based on Cohen's d. Dot size corresponds to the percentage of all signatures tested (*n* = 7503). (**C**) Spearman correlation coefficients of Cohen's d with signature sizes across the datasets and methods. Asterisk (*) in each cell indicates p-value < 0.01. Color of the heatmap represents correlation coefficient. (**D**) Scores of a cell cycle gene set (GO:0007049) calculated using four methods along with MKI67 expression, gene counts, and cell cycle phases predicted by Seurat in Tumor and normal cell populations of HNSC dataset (GSE103322). The red box highlights non-cycling tumor cells that exhibit higher scores than non-cycling normal cells.

The online version of this article includes the following source data and figure supplement(s) for figure 1:

Source data 1. Source data for ***Figure 1***.

Figure supplement 1. Bias in gene counts.

Figure supplement 2. Patterns of up and down regulated signatures.

Figure supplement 3. GSVA and ssGSEA comparison.

We observed that GSVA was slower than other methods. For instance, running the head and neck dataset with 200 threads on a 2.5 GHz Xeon processor took more than 3 days to complete. Because GSVA outputs were highly consistent with those of ssGSEA (***Figure 1—figure supplement 3***), and its running speed would not likely scale up to single cell datasets, we dropped GSVA from subsequent analyses.

We next use an example to illustrate how gene counts may affect signature scores and data interpretation. We scored the cell cycle gene signature from Gene Ontology (GO:0007049) in the head and neck cancer (HNSC) dataset (***Puram et al., 2017***). For comparison, we used Seurat (***Butler et al., 2018***) to identify cycling cells at G2M and S phases and also included the expression of KI67, a proliferation marker expressed by cycling cells. We observed all four methods reported a higher score in cycling cells (***Figure 1D***). However, ssGSEA scores were significantly higher in non-cycling cancer cells than in normal cells although neither are cycling (Student's *t* test, p < 2.2e-16). In contrast, the single-cell-based methods did not show this difference.

Sensitivity and specificity

We next generated gold standard up- and down-regulated gene sets at various sizes to benchmark detection sensitivity. To simulate real-world scenario, we added noises to each gene set so that higher noise levels attenuate more of their signals (Method).

We found gene set size had negligible impacts on detection rates (***Figure 2—figure supplement 1***). Noise levels, on the other hand, had a much bigger impact. For up gene sets, all methods were able to detect 50% of the gene sets on average even at the 80% noise level (***Figure 2A***). However, for down gene sets, ssGSEA performed worse than other methods (***Figure 2B***). Without noise, it only detected about 85% of the gene sets. At 80% noise level, it detected around 30% of the gene sets, compared to 70–80% by single-cell methods.

To benchmark detection specificity, we generated null datasets by randomly choosing 100 cancer cells from each dataset. We down sampled each cell to 50% coverage, resulting in lower gene counts. These cells were then normalized to the same total coverage. Because the down-sampled cells were the same cells with the original, no up or down gene sets were expected between the two groups.

We found AUCell and JASMINE outperformed SCSE and ssGSEA in specificity (***Figure 2C***). On average, AUCell detected 7.5 and 1.2% of the input gene sets as up and down; JASMINE detected 4 and 1%. In four datasets, SCSE overestimated down regulated gene sets (average 25.4%). ssGSEA grossly overestimated both up- and down regulated gene sets in all datasets (average 11 and 46%, respectively).

Next, we estimated how changes in gene count affected score stability. For each gene signature, we calculated coefficient of variance (CV) (***Figure 2—figure supplement 2***). We observed that single-cell methods were robust to our down sampling, with little changes in CV. However, in three datasets, ssGSEA had a higher CV in down-sampled cells than in original cells.

Comparing with consensus and computational efficiency

We next compared the three single-cell-based methods against consensus—gene sets that were identified as up- or down regulated by at least two methods. JASMINE aligned better with the consensus in most datasets (***Figure 2D*** and ***Figure 2—figure supplement 3***). AUCell overall had lower accuracies against the consensus, particularly for down regulated gene sets, likely because it is designed to score marker signatures (***Figure 2D***). When breaking down to sensitivity and specificity, AUCell showed higher false positive rates (***Figure 2—figure supplement 4A***). The three tools showed comparable sensitivity. Across the 10 datasets, SCSE and JASMINE showed better correlation (***Figure 2—figure supplement 4B,C***), thus explaining their relatively better alignment with the consensus.

Finally, we tested computing efficiency of the four methods. We randomly generated gene sets of various sizes and analyzed each gene set in a dataset consisting of 2000 cells. ***Figure 2—figure supplement 5*** shows the time and memory use averaged over 50 iterations for one gene set under the same hardware configuration (2.20 GHz CPU, 32 GB memory). Overall, SCSE was the most efficient computationally. JASMINE was the second fastest algorithm but required a memory similar to that of ssGSEA. AUCell was slightly faster than ssGSEA but required the most memory. Their performances were robust across signature sizes.

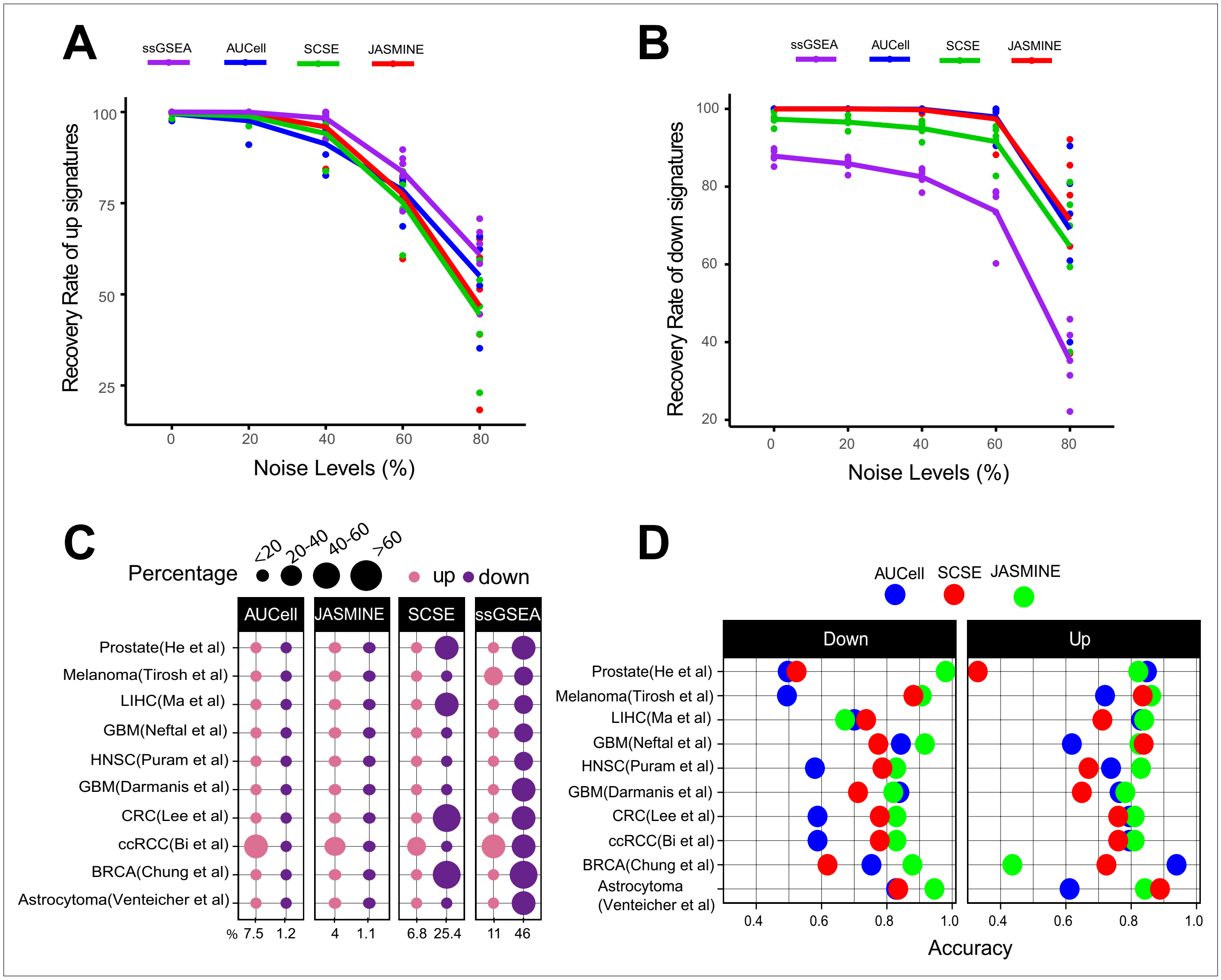

Figure 2. Sensitivity, specificity and accuracy. (**A**) Recovery rate for up gene signatures across five noise levels by the four methods. Each dot represents one dataset. At each noise level, average of all datasets is used to represent the performance of each method. (**B**) Similarly, for down signatures. (**C**) Percentages of false up and down signatures. The size of the dots corresponds to the percentages of all the signatures tested. Because the contrasting groups are generated by down sampling, no signatures are expected to be identified. The numbers below the heatmap are the average percentage. (**D**) Accuracy of the three methods, separated into up and down signatures. Accuracy is calculated as the agreement with consensus calls by at least two methods.

The online version of this article includes the following source data and figure supplement(s) for figure 2:

Source data 1. Source data for ***Figure 2***.

Figure supplement 1. Benchmarking sensitivity using simulated gene signatures.

Figure supplement 2. Coefficient of Variance.

Figure supplement 3. Comparison of calling results from the four methods across the seven datasets.

Figure supplement 4. Consistency with consensus and pairwise comparison.

Figure supplement 5. Evaluation of computing cost.

Dropouts affect ssGSEA scores

To further show the impact of dropouts on ssGSEA scores, we re-ran the down sampling experiment by setting down sampling rates to 20, 40, 60, and 80%. A down sampling rate of 80% indicates that the total sequencing depth of the down sampled cell is at 80% of the original cell. A lower down sampling rate creates more dropouts. We observed that as the down sampling rate shifted from low to high,

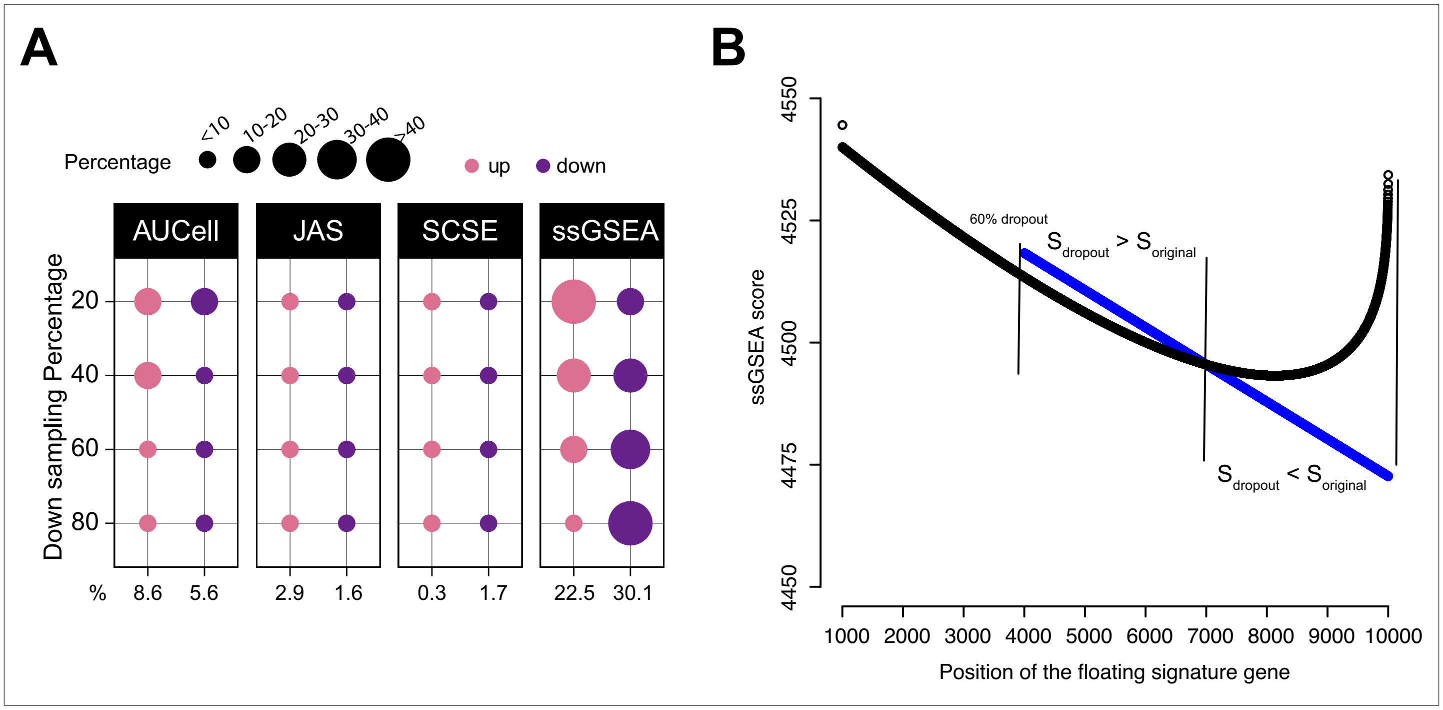

Figure 3. Impact of dropouts on ssGSEA signature scoring. (**A**) Percentages of up and down regulated gene signatures in original cells relative to down sampled cells for four levels of down sampling (20, 40, 60, and 80%) based on Cohen's d. Dot size corresponds to the percentage of all signatures tested (n = 7503) in Head and Neck (***Puram et al., 2017***). (**B**) Effect of dropouts on ssGSEA scoring using a dummy expression matrix. The black line denotes the cell without any dropouts, and the blue line denotes the same cell with a 60% dropout rate. Note that for the gene signature, the first 99 genes are fixed. The x axis reflects the position of the last signature gene. When the gene is at rank <4000. The two cells give identical scores. However, after entering dropout zone, the scores start to deviate.

The online version of this article includes the following source data and figure supplement(s) for figure 3:

Source data 1. Source data for ***Figure 3***.

Figure supplement 1. Down sampling levels affect signature scoring.

Figure supplement 2. An example showing ssGSEA score changes.

the pattern of deregulated gene sets by ssGSEA also shifted from more up gene sets towards more down gene sets (***Figure 3A*** and ***Figure 3—figure supplement 1***). This observation strongly associates dropouts with ssGSEA scores. In ***Figure 3—figure supplement 2***, we provide an example using real data to show how ssGSEA scores change when limited to cells with comparable gene counts.

To understand how dropouts may affect ssGSEA scoring, we simulated a dummy expression profile of a cell with 10,000 genes. We then set different dropout rates to the cell ranging from no dropout to 80% dropout. Genes were ranked from high to low expression to mimic ssGSEA calculation and the gene rankings were identical at each dropout rate. We then assembled a gene signature of 100 genes. The first 99 genes were drawn from the top 1000 genes, and the last gene was added to the signature from the 1001st gene onward to the bottom of the list. This way, the signature consists of 99 highly expressed genes and a floating gene. As the floating gene moves toward the dropout zone, we can observe how the signature score changes, especially around the dropout zone. The codes to reproduce these data can be found at our Github repository (Methods). ***Figure 3B*** shows the result at the 60% dropout rate. Before entering the dropout zone, the signature score was identical between the no-dropout cell and the 60%-dropout cell. However, when the floating gene was between ranks 4000 and 7000, the signature was scored higher in the dropout cell; when moving further toward ranks 7000–10,000, the signature was scored lower in the dropout cell. Thus, the scores teetered around the tie rank of dropout genes, that is, 7000. This simple experiment demonstrates that though tedious, dropouts are sufficient to change ssGSEA scores in unintended ways.

Discussion

In this study, we benchmarked five signature-scoring methods, including two that were developed for bulk sample RNAseq analysis (ssGSEA and GSVA), and three that were developed for scRNA-seq analysis (AUCell, SCSE, and JASMINE). Because ssGSEA and GSVA generate highly correlated scores and GSVA is slower, we focused on ssGSEA and the other three methods. Unlike methods that test for signature enrichment between sample groups, these signature-scoring methods quantify expression activity of a gene set in a sample independent of other samples. Outputs from these methods can be conveniently tested for association with complex sample characteristics. We show that single-cell-based methods are more robust at identifying *bona fide* up- and down regulated gene signatures from scRNAseq data. In contrast, bulk-sample-based ssGSEA and GSVA are more susceptible to dropouts. These disparities stem from their distinct mechanism to score signatures. AUCell uses Area Under the Curve (AUC) to test the enrichment of a gene set among top expressed genes in a cell. SCSE measures a signature using normalized total expression of the signature genes. AUCell and SCSE intuitively counter dropout effects by using only expressed genes. JASMINE considers the enrichment of signature genes in expressed genes to counter dropout effects, and meanwhile, evaluates the average expression level of the expressed signature genes. Bulk-sample-based methods were not designed to deal with excessive dropouts, which may mislead analysis especially if cell biology underpins different dropout rates. We further show that among single-cell methods, JASMINE and SCSE showed better concordance, and JASMINE outperforms SCSE in down sampling tests.

Dropouts were initially thought to result from low amounts of input RNA and transcriptional stochasticity. However, more studies showed that dropouts are associated with cell states and identities (**Gulati et al., 2020**; **Qiu, 2020**). Thus, in cellular contexts where cells may have systematic differences in dropout rates, any tool that fails to account for dropouts may generate misleading even erroneous data. In conclusion, our results caution against using bulk-sample-based signature-scoring methods to score single cells. Our study also suggests that it is important to consider cellular contexts when benchmarking methods for scRNA-seq data analysis, particularly when bulk-sample-based methods are involved.

Materials and methods

Single cell data sets and signature scoring

We used single cell data sets from 10 published studies (**Bi et al., 2021**; **Chung et al., 2017**; **Darmanis et al., 2017**; **He et al., 2021**; **Lee et al., 2020**; **Ma et al., 2019**; **Neftel et al., 2019**; **Puram et al., 2017**; **Tirosh et al., 2016**; **Venteicher et al., 2017**) for the evaluation of number of expressed genes in tumor versus normal cells to identify significant heterogenous patterns among the two phenotypes. Annotations of cell identity were also downloaded from each publication. We filtered all the data sets by removing non-expressed genes and then applied regularized negative binomial regression implemented in Seurat for normalization. We used C2 (n = 6226), C3 (n = 3556) and Hallmarks (n = 50) modules from MSigDB (**Subramanian et al., 2005**) v.7.2, to calculate the ratio of signature genes across all data sets and further signature scoring. We tested five tools for signature score calculations, including SCSE (**Pont et al., 2019**), AUCell (**Aibar et al., 2017**), ssGSEA, GSVA (**Hänzelmann et al., 2013**), and JASMINE. GSVA was included in tumor-normal comparisons but was dropped in gold standard tests and down sampling experiments due to slow running speed and highly correlated outputs with ssGSEA. We used GSVA and ssGSEA methods implemented in the GSVA Bioconductor (**Hänzelmann et al., 2013**) and AUCell method from AUCell Bioconductor packages (**Aibar et al., 2017**) with default parameters. We implemented SCSE in the R environment (v4.0) according to the equation reported in their paper (**Pont et al., 2019**). The output scores were used as is in tumor/normal cell comparisons and simulation analyses.

Jointly Assessing Signature Mean and Inferring Enrichment

For each signature, JASMINE calculates the approximate mean using gene ranks among expressed genes and the enrichment of the signature in expressed genes. The two are then scaled to 0–1 and averaged to result in the final JASMINE score.

Assume R_g represents the rank of an expressed signature gene g among genes with expression value >0 in a cell, then a mean rank vector V_{mean} is calculated as follows:

$$V_{mean} = \sum_{i=1}^{m} R_{g,i} / \left(m \times N\right)$$

Where *m* represents the total number of expressed signature genes and *N* represents the total number of expressed genes in the cell. Because the mean is based on ranks, it is robust to scale normalization methods. It assesses the expression level of the signature only using genes detected in one cell, thus minimizing the effect of dropouts.

To assess signature enrichment in the expressed genes, we calculate either the Odds Ratio (OR) using four variables: a = signature genes expressed, b = signature genes not expressed, c = non-signature genes expressed, and d = non-signature genes not expressed. The signature enrichment using OR is calculated as follows:

$$OR = \left(a * d\right) / \left(b * c\right)$$

In practice, *c* is unlikely zero. For smaller signatures, b can be occasionally 0. In that case, we replace it with 1. In our testing, we observed that 99% of the gene sets did not encounter this issue in any of the cells we analyzed. Despite this rare incidence, we provide Likelihood Ratio (LR) as an alternative to OR in the JASMINE function. LR is calculated as follows:

$$LR = a * \left(c + d\right) / \left(c * \left(a + b\right)\right)$$

OR and *LR* generate highly similar results according to our tests. *OR* (or *LR*) assesses the distribution of signature genes against the dropout background. Finally, V_{mean} and *OR* are linearly scaled to [0–1] and are averaged to generate JASMINE scores.

Effect size for gene sets scores

We first filtered the 9835 gene sets to 7503 by requiring a minimum size of 20 genes. We scored these gene sets (C2, C3, and hallmarks) among tumor and normal phenotypes using Cohen’s d (***Cohen, 1988***). We utilized an R package ‘effectsize’ (***Ben-Shachar et al., 2020***) to calculate this metric. We used Cohen’s d ≥ one as a threshold for positive or up cases and d ≤ –1 as negative or down cases with respect to tumor versus normal cells.

Gene signature simulation

We first identified differentially expressed genes for each dataset using MAST (***Finak et al., 2015***), a method that explicitly accounts for dropout rates. We then randomly drew *N* genes from upregulated genes to generate an up gene set of size *N*, and similarly for down gene sets. Because in practice up and down regulated gene sets contain genes that have no expression change, or even changes at opposite directions, we added noises to the simulated gene sets. For instance, when setting noise level to 20%, an up gene set of size *N* would have 20% of its genes drawn from the remainders other than the upregulated genes. Following this procedure, we generated gene sets at the sizes of 50, 100, 150, 200, and 300 genes. For each size, we set the noise levels to 0, 20, 40, 60, and 80%. For each noise-size combination, we randomly generated 200 gene sets. In total, 5000 gene sets were generated per data set for each direction.

Down sampling

We subset 100 tumor cells per data set for down sampling. The R package ‘scuttle’ (***McCarthy et al., 2017***) was used to down sample each cell to 50% total coverage. The down sampled data sets were then scaled back to ensure equal total coverage before comparison. Down sampling was also performed at various levels (20, 40, 60, and 80%) to examine its impact upon scoring of different methods especially ssGSEA.

Dropouts impact on ssGSEA

We generated a toy example to demonstrate the impact of dropouts on ssGSEA scores. We used a signature comprised of n + 1 genes. We randomly selected 99 genes from top 1000 genes of the dummy expression matrix. These 99 genes are fixed. Then starting from 1001 onward, we added one gene to the signature each time, starting from 1001st gene index throughout the remaining list. This allowed us to investigate how signature score changes as the additional gene gradually moves from high expression to dropouts. We simulated the scenario with 0% (original scores), and 60% dropout rates. This simulation clearly reflects how the dropout rates significantly affect the scoring by ssGSEA.

Data availability

Single cell data sets used in this study including their downloading sources were listed in ***Supplementary file 1***. Gene sets were downloaded from MSigDB v.7.2 (http://www.gsea-msigdb.org/gsea/msigdb/index.jsp) for C2, C3 and Hallmark modules. JASMINE and ssGSEA testing codes are available on Github (https://github.com/NNoureen/JASMINE, copy archived at swh:1:rev:ba00996ad165ff471c6fada83e6cf76af50acdfa; ***Noureen, 2021b***).

Acknowledgements

This work was supported by CPRIT (RR170055 to ZS). NN was supported by a CPRIT postdoctoral fellowship award (RP170345).

Additional information

Funding

Funder	Grant reference number	Author
Cancer Prevention and Research Institute of Texas	RR170055	Siyuan Zheng
Cancer Prevention and Research Institute of Texas	RP170345	Nighat Noureen

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Nighat Noureen, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Software, Visualization, Writing - original draft; Zhenqing Ye, Investigation, Visualization; Yidong Chen, Investigation, Resources; Xiaojing Wang, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing - original draft; Siyuan Zheng, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Visualization, Writing - original draft

Author ORCIDs

Nighat Noureen http://orcid.org/0000-0001-7495-8201
Siyuan Zheng http://orcid.org/0000-0002-1031-9424

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.71994.sa1
Author response https://doi.org/10.7554/eLife.71994.sa2

Additional files

Supplementary files

- Supplementary file 1. Single cell RNAseq datasets. The table lists all the ten datasets used in this study. Cell annotations were downloaded from the original publications.
- Transparent reporting form

Data availability

The current manuscript is a computational study, so no data have been generated for this manuscript. Single cell data sets used in this study including their downloading sources were listed in Supplementary file 1. Gene sets were downloaded from MSigDB v.7.2. JASMINE source code is available on Github (https://github.com/NNoureen/JASMINE, copy archived at swh:1:rev:ba00996ad165ff471c6fada83e6cf76af50acdfa). Source Data contain the numerical data used to generate the figures.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Noureen N	2021	Signature-scoring methods developed for bulk samples are not adequate for cancer single-cell RNA sequencing data	https://github.com/NNoureen/JASMINE	GitHub, GitHub

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Tirosh I, Izar B, Prakadan SM	2016	Single-cell expression of metastatic melanoma	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72056	NCBI Gene Expression Omnibus, GSE72056
Puram SV, Tirosh I, Parikh AS, Patel AP	2017	Single-cell expression of head and neck cancer	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103322	NCBI Gene Expression Omnibus, GSE103322
Bi K, Mx He, Bakouny Z, Kanodia A	2021	Single-cell expression of advanced renal cell carcinoma	https://singlecell.broadinstitute.org/single_cell/study/SCP1288/tumor-and-immune-reprogramming-during-immunotherapy-in-advanced-renal-cell-carcinoma	Broad Single Cell portal, SCP1288
Lee HO, Hong Y, Etlioglu HE, Cho YB	2020	Single-cell expression of colorectal cancer	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE144735	NCBI Gene Expression Omnibus, GSE144735
Ma L, Hernandez MO, Zhao Y, Mehta M	2019	Single-cell expression of liver cancer	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE125449	NCBI Gene Expression Omnibus, GSE125449
Neftel C, Laffy J, Filbin MG, Hara T	2019	Single-cell expression of IDH wildtype glioblastoma	https://singlecell.broadinstitute.org/single_cell/study/SCP393/single-cell-rna-seq-of-adult-and-pediatric-glioblastoma	Broad Single Cell portal, SCP393
Venteicher AS, Tirosh I, Hebert C, Yizhak K	2017	Single-cell expression of IDH mutant astrocytoma	https://singlecell.broadinstitute.org/single_cell/study/SCP50/single-cell-rna-seq-analysis-of-astrocytoma	Broad Single Cell portal, SCP50
Darmanis S, Sloan SA, Croote D, Mignardi M	2017	Single-cell expression of glioblastoma	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84465	NCBI Gene Expression Omnibus, GSE84465
Chung W, Eum HH, Lee HO, Lee KM	2017	Single-cell expression of breast cancer	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75688	NCBI Gene Expression Omnibus, GSE75688

Continued on next page

Continued

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Mx He, Cuoco MS, Crowdis J, Bosma-Moody A, Zhang Z, Bi K, Kanodia A	2021	Single-cell expression of prostate cancer	https://singlecell.broadinstitute.org/single_cell/study/SCP1244/transcriptional-mediators-of-treatment-resistance-in-lethal-prostate-cancer	Broad Single Cell portal, SCP1244

References

Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, Rambow F, Marine J-C, Geurts P, Aerts J, van den Oord J, Atak ZK, Wouters J, Aerts S. 2017. SCENIC: single-cell regulatory network inference and clustering. *Nature Methods* **14**:1083–1086. DOI: https://doi.org/10.1038/nmeth.4463, PMID: 28991892

Ben-Shachar M, Lüdecke D, Makowski D. 2020. effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software* **5**:2815. DOI: https://doi.org/10.21105/joss.02815

Bi K, He MX, Bakouny Z, Kanodia A, Napolitano S, Wu J, Grimaldi G, Braun DA, Cuoco MS, Mayorga A, DelloStritto L, Bouchard G, Steinharter J, Tewari AK, Vokes NI, Shannon E, Sun M, Park J, Chang SL, McGregor BA, et al. 2021. Tumor and immune reprogramming during immunotherapy in advanced renal cell carcinoma. *Cancer Cell* **39**:649–661. DOI: https://doi.org/10.1016/j.ccell.2021.02.015, PMID: 33711272

Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**:411–420. DOI: https://doi.org/10.1038/nbt.4096, PMID: 29608179

Chung W, Eum HH, Lee HO, Lee KM, Lee HB, Kim KT, Ryu HS, Kim S, Lee JE, Park YH, Kan Z, Han W, Park WY. 2017. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature Communications* **8**:15081. DOI: https://doi.org/10.1038/ncomms15081, PMID: 28474673

Cohen J. 1988. Statistical Power Analysis for the Behavioral Sciences. Routledge.

Darmanis S, Sloan SA, Croote D, Mignardi M, Chernikova S, Samghababi P, Zhang Y, Neff N, Kowarsky M, Caneda C, Li G, Chang SD, Connolly ID, Li Y, Barres BA, Gephart MH, Quake SR. 2017. Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. *Cell Reports* **21**:1399–1410. DOI: https://doi.org/10.1016/j.celrep.2017.10.030, PMID: 29091775

Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, Linsley PS, Gottardo R. 2015. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16**:278. DOI: https://doi.org/10.1186/s13059-015-0844-5, PMID: 26653891

Gulati GS, Sikandar SS, Wesche DJ, Manjunath A, Bharadwaj A, Berger MJ, Ilagan F, Kuo AH, Hsieh RW, Cai S, Zabala M, Scheeren FA, Lobo NA, Qian D, Yu FB, Dirbas FM, Clarke MF, Newman AM. 2020. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science (New York, N.Y.)* **367**:405–411. DOI: https://doi.org/10.1126/science.aax0249, PMID: 31974247

Hänzelmann S, Castelo R, Guinney J. 2013. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**:7. DOI: https://doi.org/10.1186/1471-2105-14-7, PMID: 23323831

He MX, Cuoco MS, Crowdis J, Bosma-Moody A, Zhang Z, Bi K, Kanodia A, Su MJ, Ku SY, Garcia MM, Sweet AR, Rodman C, DelloStritto L, Silver R, Steinharter J, Shah P, Izar B, Walk NC, Burke KP, Bakouny Z, et al. 2021. Transcriptional mediators of treatment resistance in lethal prostate cancer. *Nature Medicine* **27**:426–433. DOI: https://doi.org/10.1038/s41591-021-01244-6, PMID: 33664492

Hicks SC, Townes FW, Teng M, Irizarry RA. 2018. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics (Oxford, England)* **19**:562–578. DOI: https://doi.org/10.1093/biostatistics/kxx053, PMID: 29121214

Holland CH, Tanevski J, Perales-Patón J, Gleixner J, Kumar MP, Mereu E, Joughin BA, Stegle O, Lauffenburger DA, Heyn H, Szalai B, Saez-Rodriguez J. 2020. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biology* **21**:36. DOI: https://doi.org/10.1186/s13059-020-1949-z, PMID: 32051003

Lee H-O, Hong Y, Etlioglu HE, Cho YB, Pomella V, Van den Bosch B, Vanhecke J, Verbandt S, Hong H, Min J-W, Kim N, Eum HH, Qian J, Boeckx B, Lambrechts D, Tsantoulis P, De Hertogh G, Chung W, Lee T, An M, et al. 2020. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nature Genetics* **52**:594–603. DOI: https://doi.org/10.1038/s41588-020-0636-z, PMID: 32451460

Ma L, Hernandez MO, Zhao Y, Mehta M, Tran B, Kelly M, Rae Z, Hernandez JM, Davis JL, Martin SP, Kleiner DE, Hewitt SM, Ylaya K, Wood BJ, Greten TF, Wang XW. 2019. Tumor Cell Biodiversity Drives Microenvironmental Reprogramming in Liver Cancer. *Cancer Cell* **36**:418–430. DOI: https://doi.org/10.1016/j.ccell.2019.08.007, PMID: 31588021

Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Kamińska B, Huelsken J, Omberg L, Gevaert O, Colaprico A, Czerwińska P, Mazurek S, Mishra L, Heyn H, Krasnitz A, Godwin AK, Lazar AJ, Cancer Genome Atlas Research Network, Stuart JM, et al. 2018. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell* **173**:338-354.. DOI: https://doi.org/10.1016/j.cell.2018.03.034, PMID: 29625051

McCarthy DJ, Campbell KR, Lun ATL, Wills QF. 2017. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics (Oxford, England*) **33**:1179–1186. DOI: https://doi.org/10.1093/bioinformatics/btw777, PMID: 28088763

Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, Richman AR, Silverbush D, Shaw ML, Hebert CM, Dewitt J, Gritsch S, Perez EM, Gonzalez Castro LN, Lan X, Druck N, Rodman C, Dionne D, Kaplan A, Bertalan MS, et al. 2019. An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* **178**:835-849.. DOI: https://doi.org/10.1016/j.cell.2019.06.024, PMID: 31327527

Noureen N, Wu S, Lv Y, Yang J, Alfred Yung WK, Gelfond J, Wang X, Koul D, Ludlow A, Zheng S. 2021a. Integrated analysis of telomerase enzymatic activity unravels an association with cancer stemness and proliferation *Nature Communications* **12**:139. DOI: https://doi.org/10.1038/s41467-020-20474-9, PMID: 33420056

Noureen N. 2021b. JASMINE swh:1:rev:ba00996ad165ff471c6fada83e6cf76af50acdfa. GitHub. https://archive.softwareheritage.org/swh:1:dir:f5b092ae4a9b8e9a7314131251bc01391abd8e1f;origin=https://github.com/NNoureen/JASMINE;visit=swh:1:snp:68b3eef22d209a4b966ce1c768b5eded68dae4ff;anchor=swh:1:rev:ba00996ad165ff471c6fada83e6cf76af50acdfa

Pont F, Tosolini M, Fournié JJ. 2019. Single-Cell Signature Explorer for comprehensive visualization of single cell signatures across scRNA-seq datasets. *Nucleic Acids Research* **47**:e133. DOI: https://doi.org/10.1093/nar/gkz601, PMID: 31294801

Puram SV, Tirosh I, Parikh AS, Patel AP, Yizhak K, Gillespie S, Rodman C, Luo CL, Mroz EA, Emerick KS, Deschler DG, Varvares MA, Mylvaganam R, Rozenblatt-Rosen O, Rocco JW, Faquin WC, Lin DT, Regev A, Bernstein BE. 2017. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* **171**:1611-1624.. DOI: https://doi.org/10.1016/j.cell.2017.10.044, PMID: 29198524

Qiu P. 2020. Embracing the dropouts in single-cell RNA-seq analysis. *Nature Communications* **11**:1169. DOI: https://doi.org/10.1038/s41467-020-14976-9, PMID: 32127540

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **102**:15545–15550. DOI: https://doi.org/10.1073/pnas.0506580102, PMID: 16199517

Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, Rotem A, Rodman C, Lian C, Murphy G, Fallahi-Sichani M, Dutton-Regester K, Lin JR, Cohen O, Shah P, Lu D, Genshaft AS, Hughes TK, Ziegler CGK, Kazer SW, et al. 2016. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science (New York, N.Y.)* **352**:189–196. DOI: https://doi.org/10.1126/science.aad0501, PMID: 27124452

Venteicher AS, Tirosh I, Hebert C, Yizhak K, Neftel C, Filbin MG, Hovestadt V, Escalante LE, Shaw ML, Rodman C, Gillespie SM, Dionne D, Luo CC, Ravichandran H, Mylvaganam R, Mount C, Onozato ML, Nahed BV, Wakimoto H, Curry WT, et al. 2017. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science (New York, N.Y.*) **355**:eaai8478. DOI: https://doi.org/10.1126/science.aai8478, PMID: 28360267

Zhang Y, Ma Y, Huang Y, Zhang Y, Jiang Q, Zhou M, Su J. 2020. Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. *Computational and Structural Biotechnology Journal* **18**:2953–2961. DOI: https://doi.org/10.1016/j.csbj.2020.10.007, PMID: 33209207