A novel method to accurately estimate pathway activity in single cells for clustering and differential analysis

Daniel Davis, Avishai Wizel, Yotam Drier

The Lautenberg Center for Immunology and Cancer Research, IMRIC, Faculty of Medicine, Hebrew University of Jerusalem, Israel.

Abstract

Inferring which and how biological pathways and gene sets are changing is a key question in many studies that utilize single-cell RNA sequencing. While typically these questions are addressed by quantifying the enrichment of known gene sets in lists of genes derived from global analysis, there are many benefits to first inferring pathway status of each pathway in each cell. Here we report on a new tool that allows such inference, apply it on several datasets, and demonstrate the improvement over current state of the art methods. We also demonstrate how it can be used to better classify cells based on gene set scores instead of single genes and the benefits of this approach.

Introduction

Single-cell RNA sequencing (scRNA-seq) has been in widespread use for several years, enabling us to deepen our understanding of tissue characteristics and heterogeneity in health and disease (Chi & Deng, 2020). Pathway analysis has proven to be a useful tool to interpret gene expression data in general and scRNA-seq in particular. This is typically accomplished by identifying a set of interesting genes (such as differentially expressed genes) and then determining which pathways these genes are over-represented in. While such approaches are very useful, there are many potential benefits to estimate pathway activity in each cell first and then utilize this information for downstream analysis. Such an approach allows for both overcoming limitation of single-cell data, such as inaccurate estimation of the expression of a single gene in a single cell, and utilizing pathway activity for unsupervised analysis such as clustering. We and others have developed approaches for pathway analysis in single sample for bulk RNA-seq (Drier et al., 2013; Hänzelmann et al., 2013; H. Wang et al., 2016), however methods developed for bulk RNA-seq often do not work well for scRNA-seq data, due to its special characteristics (Noureen et al., 2022).

The most prominent and popular method that does estimate pathway activity in single cells is AUCell (Aibar et al., 2017). AUCell was originally developed as part of an effort to reconstruct gene regulatory networks, but it allows to score gene sets. To do so, it ranks all the genes expressed by a cell based on their transcription level and uses the area under the curve (AUC) measure to represent the fraction of the top 5% genes which belongs to the relevant gene set. While AUCell works well and helped in the analysis of multiple scRNA-seq studies, it suffers from several limitations.

Here we present a novel, highly sensitive method for analyzing the activity of a gene set, or biological pathway, in single cells, which we term Single Pathway analysis in Single Cells (SiPSiC). To demonstrate the robustness of SiPSiC across data types, we applied both SiPSiC and AUCell to datasets of COVID-19 infected lungs and of glioblastoma, which are two contexts where scRNA-seq has been extensively used to study disease pathophysiology.

For the COVID-19 analysis, we used a single-nuclei RNA-seq dataset based on lung resections or biopsies from recently deceased patients (Melms et al., 2021) and a scRNA-seq dataset of resected lungs of African green monkeys inoculated with either active or deactivated SARS-CoV-2 (Speranza et al., 2021). The original analysis of the human data included gene-based clustering from which cluster markers were inferred to elucidate cellular response to Severe Acute Respiratory Syndrome

Coronavirus 2 (SARS-CoV-2) infection. The monkey data was originally analyzed by principal component analysis (PCA), clustering and Fast Gene Set Enrichment Analysis (Korotkevich et al., 2021). We show that SiPSiC can point out the biological differences between control and SARS-CoV-2 infected lung cells, detecting pathways that drive the response to the virus and are not reported in the original papers or missed by AUCell.

While for COVID-19 we compared infected and non-infected tissues, in the glioblastoma analysis we used SiPSiC and AUCell to interrogate the different cellular programs recapitulated by tumor cells, and demonstrate SiPSiC's ability to accurately demarcate both non-malignant and malignant cell subpopulations. Four malignant meta-modules (cellular states) were found in glioblastoma in previous work (Neftel et al., 2019), each characterized by typical biological functions. These are the Oligodendrocyte-progenitor-like (OPC-like), neural-progenitor-like (NPC-like), astrocyte-like (AC-like) and mesenchymal-like (MES-like) states. Neftel et al. used conventional gene-based hierarchical clustering of malignant cells and further processing steps to reveal these cellular states. However, conventional gene-based clustering suffers from significant patient bias, particularly for malignant cells, which hinders the detection of shared malignant programs across patients. We demonstrate that SiPSiC pathway scores allow for more accurate detection of existing malignant and non-malignant programs, as well as clustering together cells with shared biological programs across different patients. Finally, we show that in this dataset as well SiPSiC successfully detects meaningful findings missed by AUCell.

Methods

<u>The SiPSiC algorithm</u>

Taking an scRNA-seq gene expression matrix *X* in TPM or CPM, and a given gene-set, SiPSiC performs the following steps to calculate the score for all the cells in the data:

(1) **Calculate a normalization factor for each gene** : Calculate the median of the 5% cells with highest expression. If it is positive use as normalization factor, if zero use the maximum value as the normalization factor.

The reason behind this step is that scRNA-seq data are normally sparse (Chi & Deng, 2020), namely, the fraction of zeros in the data is large. Thus, by selecting the median of the top 5% cells there is a high likelihood that for most genes the value will be greater than zero, while on the other hand it will also not be an outlier, which may perturb further processing steps.

(2) **Calculate gene scores:** Rank the genes in the gene-set by their total expression (TPM or CPM values) across all cells, and divide the rank by the number of genes. Multiply by the expression and divide by the normalization factor from step 1. Therefore the gene score of gene *i* in cell *j* is $S_{i,j} = X_{i,j} * rank_i \big/ NF_i * n$, where $rank_i$ is the rank of $\sum_j X_{i,j}$, n is the number of genes in the gene set, and NF_i is the normalization factor for gene *i* calculated in step 1.

(3) **Pathway scoring**: For each cell, the gene-set score is the average gene score across all genes in the gene-set. $P_j = \sum_{i=1}^{n_P} S_{i,j} \big/ n_P$ where n_P is the number of genes in gene-set P.

SiPSiC is available at Bioconductor and GitHub (https://github.com/DanielDavis12/SiPSiC).

<u>scRNA-seq datasets and processing</u>

We downloaded published datasets from three papers (Melms et al., 2021; Neftel et al., 2019; Speranza et al., 2021). For the Human COVID dataset, cell identity annotations were downloaded as well. For the glioblastoma smart-seq2 dataset we used the malignant cell meta-module assignment that was published. To identify cell identities for the monkey COVID-19 dataset, we first identified

biomarkers for alveolar, activated B, and CD8+ T cells. First, we used the human COVID-19 dataset to extract marker genes of each cell type compared to all other cells, using the FindMarkers function of Seurat (Butler et al., 2018), with logfc.threshold = 1 and min.diff.pct = 0.5, following the parameters used by Speranza et al. In addition to these markers we also selected canonical marker genes of alveolar, B and T cells from the literature: CAV1, PDPN ($T1\alpha$), SFTPB, SFTPC and SFTPD were used for alveolar cells (Chen et al., 2004), CD20, CD27, CD28 for B cells (Sanz et al., 2019) and CD2, CD8 and CXCR3 (CD138) for T cells (Groom & Luster, 2011; Tzankov et al., 2005). Next, we clustered the monkey dataset using Seurat's FindClusters function with a 0.5 resolution, and all clusters were annotated as a given cell type if and only if they displayed high expression of both the canonical markers and the markers identified in the human dataset. This resulted in 1412, 1061 and 17365 cells identified as alveolar, B, and CD8+ T cells, respectively, before filtering was applied. After filtering, these numbers dropped to 1004 (86 control, 217 3-Days, and 701 10-Days) alveolar cells, 113 (29 control, 71 3-Days, and 13 10-Days) B cells, and 2224 (360 control, 1162 3-Days, and 702 10-Days) T cells.

The same filtering steps were applied to the two COVID datasets and the glioblastoma smart-seq2 dataset. We first removed cells with less than 1,000 expressed genes, and then excluded genes expressed in less than 10% of the remaining cells.

For the smart-seq2 glioblastoma dataset, we used TPM / 10 provided in the original paper (multiplying the data by a constant has no effect on SiPSiC's output) for all malignant cells with reported meta-module. For the two COVID-19 datasets, we first split the data to three subsets by cell type annotation (Alveolar, B cells and CD8+ T cells) and then applied filtering, conversion to CPM and all downstream processing separately for each of these subsets.

For the glioblastoma 10X dataset, to test if SiPSiC can be used to better cluster noisy data, we used all cells and genes reported by Neftel et al. after their filtering.

Glioblastoma pathway analysis

A pathway was considered up- or down-regulated in one group compared to another if the difference of medians was > 0, or < 0, respectively, and it passed a T test with FDR < 0.01. Since the PPR and NEU clusters were enriched in both NPC- and OPC-like cells, when a specific pathway was reported as upregulated in one of these clusters, we considered our result consistent with the findings of Garofano et al. even when only one of the respective cell states (either OPC- or NPC-like) showed upregulation of that pathway. Such "semi consistent" pathways account for 5 of the 31 consistent pathways mentioned in the main text.

Clustering and cluster composition analysis

Clustering was applied to the 10X dataset published by Neftel et al. and not the smart-seq2 dataset, as the 10X dataset is much larger and has a higher proportion of non-malignant cells (9,870 malignant cells out of 16,201 cells total for the 10X dataset, whereas the smart-seq2 dataset has 6,864 malignant cells out of 7,930 total). We used R's Seurat package to calculate the clusters using the FindClusters function based on all genes (in the gene-based clustering only) or pathways. Cluster biomarkers (differential pathways) were inferred in the pathway-based clustering only, by using Seurat's FindAllMarkers function.

We used the marker genes of Neftel et al. both for the malignant and the non-malignant cell types to annotate the cells. In both the gene-based and pathway-based analyses, cell lineage scores were calculated using SiPSiC to ensure that the same annotations are used. For the NPC-like meta-module, Neftel et al. provided two different sets of marker genes named NPClike1 and NPClike2 (referred to in the results section as the first and second NPC-like marker gene sets, respectively), reflecting the two sub-types of this meta-module. For the malignant cell types, cells with scores higher than 0.13 were annotated as belonging to the lineage for which they got this score, whereas in the case of T cells and macrophages a threshold of 0.05 was used.

Results

COVID-19 single cell RNA-seq data interrogation reveals differential activity of hallmark pathways

To test SiPSiC's ability to detect changes in activity of biological pathways after SARS-CoV-2 infection, we applied it to two distinct datasets: single nuclei RNA-seq of recently deceased COVID-19 patients or control lung tissues sampled before the outbreak of COVID-19 (Melms et al., 2021; Supplementary Table S1) and scRNA-seq of African green monkeys shortly after infection with SARS-CoV-2 or inactivated virus (Speranza et al., 2021; Supplementary Table S2). We calculated pathway scores per cell for each of the 50 MSigDB hallmark gene sets (Liberzon et al., 2015). We then compared pathway scores between SARS-CoV-2 positive and negative controls and calculated the difference between the groups' median scores, for each pathway. A pathway was considered up- or down-regulated in one group compared to another if in their comparison we got a false discovery rate (FDR) < 0.01 and a positive or negative median difference, respectively.

First, we applied SiPSiC to human alveolar cells (of both type 1 and 2; 4575 cells from COVID-19 patients vs 4303 control cells). 12 out of 50 pathways were downregulated in COVID-19 patients, and 31 upregulated (t-test, FDR < 0.01, see Supplementary Table S1). Many of the upregulated pathways are consistent with expected biological responses to COVID-19 in alveolar cells, but while the interferon response has also been reported as upregulated in alveolar type 2 (AT2) cells by Melms et al., other pathways were not reported by them. Among them are the mitotic spindle, which is consistent with demonstrated hyperplasia of pneumocytes in the lungs of monkeys after COVID-19 infection (Speranza et al., 2021), WNT/β catenin signaling, consistent with previous reports of wnt5a upregulation in severe cases of COVID-19 (Choi et al., 2020), and complement pathways, as was shown to be hyperactivated in severe SARS-CoV-2 infections (Afzali et al., 2021).

We repeated the pathway analysis for human activated B cells (55 cells of COVID-19 patients, 48 control cells), and CD8+ T cells (103 COVID-19 cells, 6 control). Despite the relatively small cell populations, TGF Beta signaling was found upregulated in activated B cells of COVID-19 patients, consistent with previous reports (Ferreira-Gomes et al., 2021; Melms et al., 2021). In addition, the G2M checkpoint pathway was upregulated in the CD8+ T cell group of COVID-19 patients, another finding missed by the conventional analysis in Melms et al. but consistent with the expected increase in T cell proliferation in the lungs, as was also demonstrated in severe SARS-CoV-2 infections (Liao et al., 2020). These findings demonstrate the high sensitivity of SiPSiC, allowing detecting pathway activity in rare cells or small datasets.

In the African green monkey experiment, two monkeys were inoculated with inactivated SARS-CoV-2 and eight with the active virus (Speranza et al., 2021). To better study the dynamics of viral infection, these eight monkeys were then split into two groups of four monkeys each and euthanized three (first group) or ten (second) days post inoculation, where the monkeys in the 10 days group had already recovered by the 10th day. Hence, we compared three groups of monkey lung cells (control, 3-Days, 10-Days) after calculating the pathway scores per cell type (Methods), following the original experiment. Checking for pathway enrichment in the alveolar cells, we found that here too the cells in the active infection group (3-Days) were enriched in the interferon alpha and gamma response and the complement pathways relative to both other groups (Supplementary Table S2), supporting our findings from the analysis of human alveolar cells. Interestingly, the mTORC1 and PI3K/AKT/mTOR signaling pathways were also enriched in this group, suggesting that similar to other cases of viral-induced pneumonia, SARS-CoV-2 also manipulates mTOR signaling to facilitate viral activity and replication (Karam et al., 2021). Indeed, inhibition of the mTOR-PI3K-AKT pathway shows potential to block SARS-CoV-2 replication (Garcia et al., 2021). Furthermore, the reactive oxygen species pathway was enriched in this group as well, consistent with a previous review which presented a potential role for increased oxidative stress in the pathophysiology of COVID-19 (Nasi et al., 2020). None of these pathways were reported in the paper by Speranza et al., further highlighting SiPSiC's high sensitivity.

Analysis of the immune cells that were found in the monkeys' lungs revealed that in both the B and CD8+ T cells the interferon alpha and gamma response pathways were enriched in the 3-days group, a finding well correlated with prior evidence of elevated interferon levels in the plasma of COVID-19 patients and the resulting effect on immune cells (Schultheiß et al., 2020). In addition, the G2M checkpoint was upregulated in both COVID groups of CD8+ T cells, once again demonstrating consistency with the human dataset analysis and previous evidence. Moreover, the inflammatory response pathway was also enriched in the 3-days group of CD8+ T cells, reflecting the anti-viral activity. The ability of SiPSiC analysis to uncover these novel changes in pathway activity demonstrates the added value of SiPSiC over existing approaches.

Glioblastoma cell subpopulations pathway enrichment is consistent with prior knowledge

To validate the applicability of SiPSiC to different data types, we analyzed scRNA-seq data of glioblastoma tumors (Neftel et al., 2019). We calculated pathway scores per cell (n = 6,863) for each of the same 50 hallmark pathways (Supplementary Table S3). The cells were then split into four groups based on their cell state assignments (2,102 NPC-, 1,211 OPC-, 1,931 AC- and 1,619 MES-like cells), and comparisons were made between each pair of groups. We found that the G2M checkpoint pathway was upregulated in the OPC- and NPC-like groups compared to the AC- and MES-like groups, indicative of a higher proportion of proliferating cells in these cell states. In addition, the hypoxia pathway was enriched in the MES-like group compared to all other three groups. These findings are consistent with the findings reported in Neftel et al. Furthermore, SiPSiC analysis suggests that the MES-like group was enriched in the inflammatory response pathway, in concordance with prior evidence that the mesenchymal subtype of GBM tumors is enriched in inflammatory response associated genes (Engler et al., 2012).

Garofano et al. classified the cells into four distinct cell clusters: glycolytic/plurimetabolic (GPM), mitochondrial (MTC), neural (NEU) and proliferative/progenitor (PPR) (Garofano et al., 2021). They found that the GPM and MTC clusters were enriched in cells of the MES- and AC-like cell states, respectively, while the PPR and NEU clusters were enriched in cells of both the OPC- and NPC-like states. Gene set enrichment of these clusters is reported for 40 hallmark pathways, and we compared their results to ours. 31 out of the 40 pathways (78%) were consistent, in the sense that if one of these pathways was upregulated in a specific cell cluster (in the reported results), it was also upregulated in our analysis in the same cell state in which this cell cluster is enriched. For instance, Garofano et al. reported that the DNA repair pathway was upregulated in the PPR cell cluster, which is enriched in both OPC- and NPC-like cells. Indeed, in our analysis we found that this pathway was upregulated in both the OPC- and NPC-like cell states. Furthermore, out of the 9 other pathways eight were found by Garofano et al. to be enriched in the MTC cluster, hence we expected them to be upregulated in the AC-like cell state in our analysis. 7 of these pathways were found to be "near-consistent" in the sense that they were significantly upregulated in the AC-like cell state compared to two of the other three cell states and insignificantly either up- (5 pathways) or down-regulated (2 pathways) in it compared to the third other cell state, making 38 of the 40 (95%) pathways reported by Garofano et al. either completely or near consistent in our SiPSiC analysis.

The two remaining pathways are the hallmark notch and mTORC1 signaling pathways. Interestingly, while Garofano et al. found that the MTC cluster was enriched in the hallmark notch signaling pathway (reported FDR < $1.23*10^{-3}$), SiPSiC analysis found that it was upregulated in the OPC-like cell state (FDR < 0.0095). Although Neftel et al. did not report upregulation of notch in any of the cell states, AUCell supported SiPSiC's result finding notch upregulation in the OPC-like cell state (Supplementary Table S3). Importantly, prior research has shown that notch signaling activation inhibits the maturation and may enhance the proliferation of oligodendrocyte progenitor cells (John et al., 2002; C. Wang et al., 2017; S. Wang et al., 1998), suggesting that notch upregulation in OPC-like glioblastoma cells may play similar roles in the pathophysiology of the disease. Similarly, while Garofano et al. found that the mTORC1 signaling pathway was enriched in the PPR cluster (reported FDR < $1.26*10^{-3}$) and Neftel et

al. did not report its upregulation in any of the cell states, SiPSiC found that it was upregulated in the MES-like cell state (FDR < $3.2*10^{-75}$), again supported by AUCell which also found the same (Supplementary Table S3). Prior research has found that in epithelial cells, TGFβ-induced epithelial to mesenchymal transition (EMT) both promotes and requires the activation of the PI3K/AKT/mTOR pathway, and more specifically activates the mTORC1 complex (Lamouille et al., 2014). Importantly, the same has also been suggested for an EMT-like process in glioblastoma (Iser et al., 2017; Zhang et al., 2014). Intriguingly, SiPSiC found upregulation of the TGFβ signaling, PI3K/AKT/mTOR signaling and EMT pathways in the MES-like cell state (with all three findings also supported by AUCell, see Supplementary Table S3), which may well account for upregulation of the mTORC1 signaling in this cell state, supporting our finding over what Garofano et al. reported for this pathway and demonstrating SiPSiC's high sensitivity in detecting true biological processes which are missed in some other analyses.

Unsupervised clustering by pathway scores reveals benign and malignant glioblastoma clones

Unsupervised clustering of scRNA-seq datasets allows to demarcate different coexisting cell lineages and subpopulations, and has thus been extensively used to interrogate tissue heterogeneity in many contexts. To test if clustering based on SiPSiC pathway scores can improve the gold standard of clustering by gene expression, we calculated cell clusters and Uniform Manifold Approximation and Projection (UMAP) of 16201 glioblastoma cells (Neftel et al., 2019) both by gene expression, and by the cells' SiPSiC scores (Methods).

Gene-based clustering produced 21 clusters. Although the dataset was sampled from 9 different patients, 4 clusters consisted only of a single patient's cells, and in 7 others 95% of the cells or more were of a single patient, reflecting the strong patient-bias of the results. In contrast, using SiPSiC's scores the same analysis produced 11 clusters, with 9 of them consisting of cells from all the patients and two much smaller clusters (274 and 77 cells out of 16,201 total) where 84% and 96% of the cells were of a single patient, respectively (Figure 1A-B). These results suggest that SiPSiC based clustering was largely based on similar biological traits of the cells rather than patient identity. To validate this, we identified differential pathways in each cluster, inferred putative lineages of the different clusters, and finally tested marker genes of each cell type across the clusters for confirmation (Methods).

Cluster 3 was enriched in five immune-related gene sets: TNFA signaling via NFKB, inflammatory response, interferon alpha response, IL-6/JAK/STAT3 signaling and interferon gamma response. As the immune cells present in the dataset were macrophages and T cells, we expected this cluster to be enriched in at least one of these cell types. Testing it for enrichment in both the macrophage and T cell markers, we found that out of 2118 cells assigned to this cluster 2024 (95%) were immune cells (1960 macrophages and 64 T cells), highlighting SiPSiC's ability to detect the major biological function of cells when applied with a broad set of pathways. Two adjacent clusters on the UMAP projection (1, 6) contain 3217 additional macrophages, forming a well isolated group of 91% of the macrophages (Figure 1C). In addition, out of 150 T cells that were found in the data, 72 were assigned to the adjacent cluster 1 and formed a clearly distinct group on the UMAP (Figure 1D).

Cluster 2 demonstrated enrichment of Wnt-beta-catenin signaling and hedgehog signaling. As both pathways were found by SiPSiC to be highly enriched in the NPC-like cells in the analysis of the first glioblastoma dataset (FDR < $1.5*10^{-20}$ for the Wnt pathway and FDR < $3.5*10^{-31}$ for the hedgehog), we tested the cluster for enrichment in two different marker gene sets of this meta-module (see Methods). Out of 2380 cells assigned to this cluster 2093 (88%) and 1533 (64%) were NPC-Like cells, and these are 80% and 84% of all the NPC-Like cells in the data, according to the first and second NPC-Like markers, respectively (Figure 1E-F). These results again demonstrate SiPSiC's high accuracy.

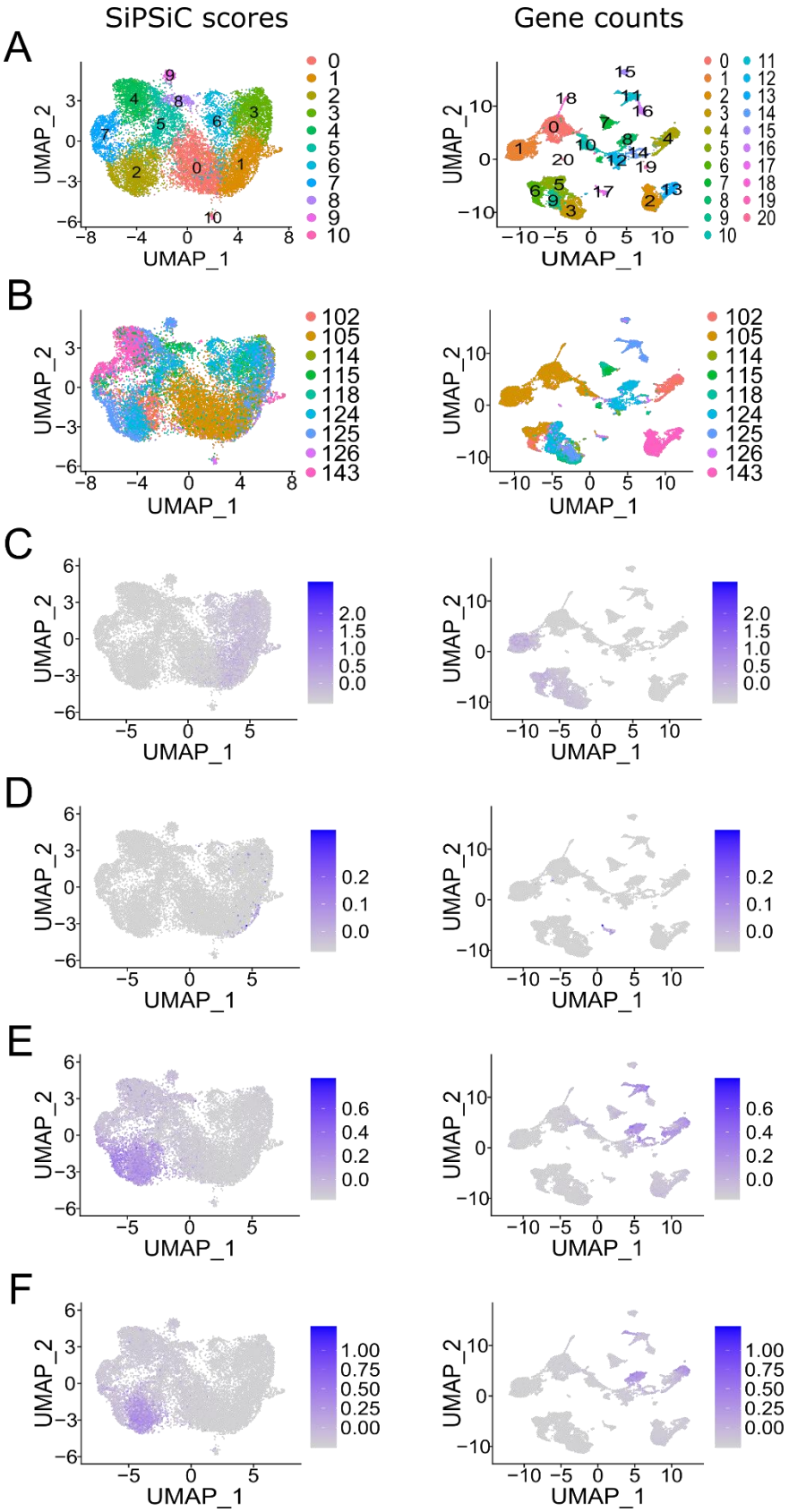

Figure 1. UMAP projections based on hallmark pathway scores by SiPSiC (left) and gene counts (right). (**A**) Cells were clustered by Louvain algorithm either according to SiPSiC scores or to gene counts. UMAPs show cells colored by cluster. (**B**) Cells colored by patient identity. (**C-F**) Marker gene set scores were calculated by SiPSiC and cells were colored according to the scores of macrophages (C), T cells (D), first (E) and second (F) NPC-like cells.

Testing the gene-based clusters for enrichment in the same marker gene sets, we found that T cells are well clustered here as well (97% of T cells were in cluster 17). However, the macrophages were split into two main groups: One group consists of clusters 3, 5, 6 and 9, containing 3978 out of a total of 5667 macrophages (70%). 88.5% - 100% of the cells in each cluster are identified as macrophages. However, cluster 1 contains 1579 additional macrophages, which are approximately 28% of all the macrophages in the data. Notably, nearly all the cells in this cluster belong to a single patient (1717 out of 1718 cells), again showing strong patient bias that hinders the ability to cluster the cells properly, even for drastic differences such as the difference between glioblastoma cells and macrophages (Figure 1C).

Moreover, NPC-like glioblastoma cells are even more scattered across different clusters. Here, no more than 41% of the NPC-like cells in the data were assigned to a single cluster (according to either of the two marker gene sets). While about 90% of NPC-like cells (2343/2605 for NPC1, 1714/1822 for NPC2) fall in four distinct clusters (0, 4, 8 and 11), the clusters are determined by patient identity - 99.8%, 97.7%, 99.6% and 100% of the cells in clusters 0, 4, 8 and 11 each belong to a single patient, respectively (Figure 1 panels B, E, F). In particular 98% of cells in cluster 0 are not NPC-like, demonstrating the strong patient bias.

Gene expression based clustering of malignant cells is known to be driven by patient of origin, as was also observed in glioblastoma (Neftel et al., 2019). This phenomenon may help differentiate malignant from normal cells, but hinders sub-clustering of malignant cells. Here we show that SiPSiC based clustering can capture shared biological processes and be less affected by patient identity, improving the ability of clustering to robustly capture various non-malignant cell types, and importantly also shared features of different malignant cells across patients.

Pathway scoring methods benchmarking reveals higher accuracy of SiPSiC for different data types

AUCell is a common, well-established method for calculating pathway scores for each cell from scRNA-seq data. To compare SiPSiC and AUCell we used AUCell to calculate scores for the hallmark gene sets using the same three datasets above. To guarantee integrity of the results, the same preprocessing steps were applied and cell assignments to the different groups were kept.

In the analysis of the monkey COVID dataset (see Supplementary Table S2), SiPSiC found 20 pathways to be upregulated and 3 pathways to be downregulated in the active infection group (3-days) of pneumocytes compared to control. In comparison, AUCell found 9 pathways to be upregulated and 10 pathways to be downregulated in this group compared to control. Out of these pathways, 9 were upregulated and 1 was downregulated according to both methods. Similarly, in the B cell analysis SiPSiC found that 2 pathways were upregulated and 7 pathways were downregulated in the 3-days group compared to control, whereas AUCell produced a significant result for 1 pathway only, which was downregulated in the active infection group and did not overlap with SiPSiC's findings. In addition, for the CD8+ T cells SiPSiC found that 20 pathways were upregulated and 4 were downregulated in the 3-days group compared to control. In contrast, AUCell found 11 and 4 pathways to be up and downregulated in the active infection group, respectively, where 10 pathways were upregulated but none were downregulated according to both methods.

Looking for major differences between the results of the two methods for this dataset, we found that while SiPSiC showed clear upregulation of the complement pathway in alveolar cells undergoing active SARS-CoV-2 infection (the 3-days group, FDR < $3.7*10^{-7}$ relative to control), AUCell failed to achieve a statistically significant result for this pathway (Figure 2A). Another compelling example is the interferon pathways in the analysis of B cells, where SiPSiC showed upregulation of both the interferon alpha and gamma pathways in the 3-days group compared to the two other groups (FDR < $8.3*10^{-5}$ and FDR < 0.0009 compared to control for the interferon alpha and gamma pathways, respectively), while AUCell's analyses of these pathways failed to show a significant difference between the active infection group (3-Days) and the control group (Figure 2B-C). Testing for differences in the analysis of

CD8+ T cells, MYC targets V2 pathway came up, where SiPSiC found upregulation in the 3-days group compared to the control group, while AUCell produced a median score of zero for all groups (Figure 2D). We indeed expect to see upregulation of MYC targets in T cells, since MYC is known to be involved in T cell proliferation (Gnanaprakasam & Wang, 2017), and increased T cell proliferation occurs in COVID-19 patients' lungs (Liao et al., 2020). To ascertain this finding, we compared MYC expression in the control and active infection groups. Indeed, average MYC expression 3 days after infection was 89% higher than control ($p < 0.0032$, unpaired Wilcoxon test), strongly supporting SiPSiC's finding.

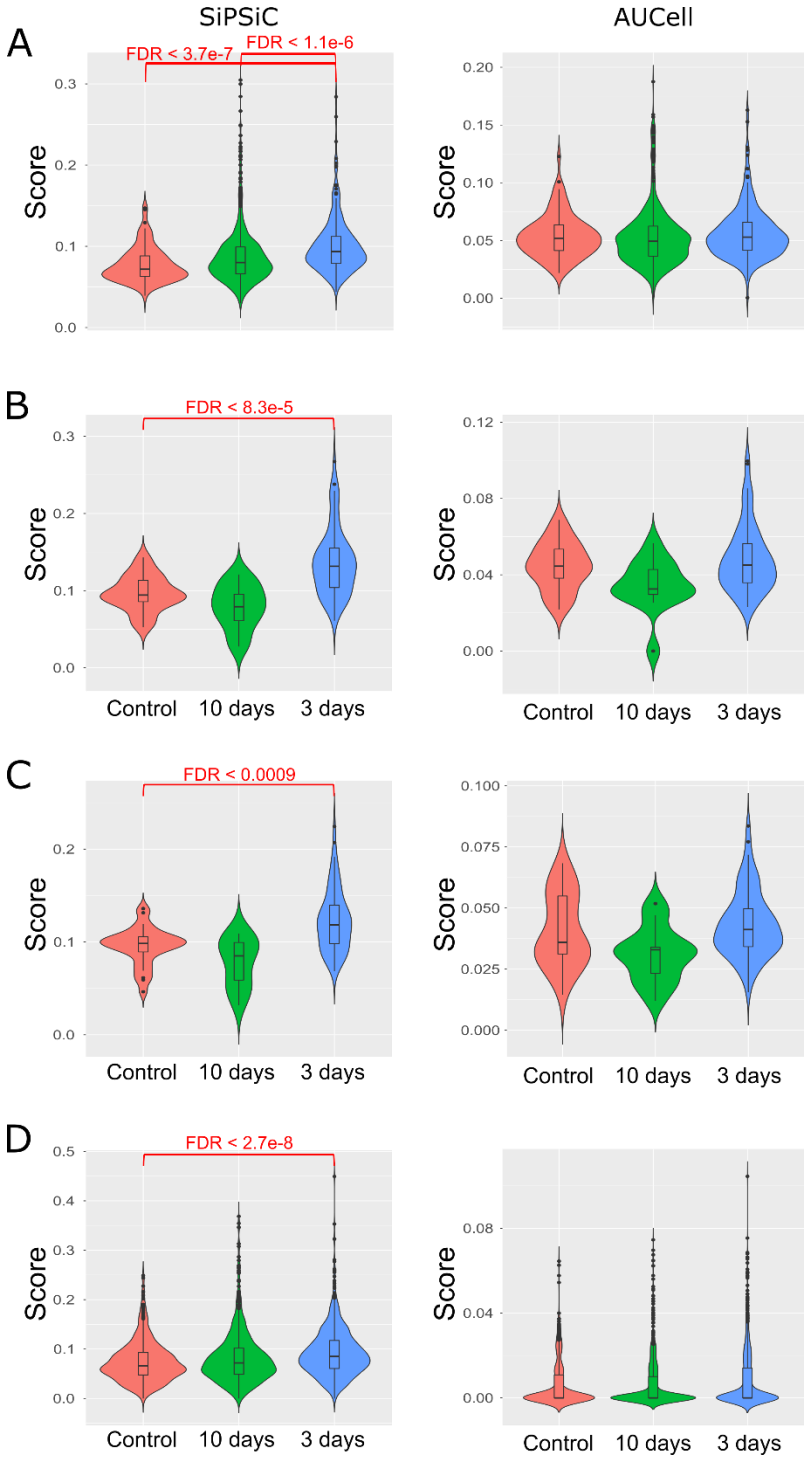

Figure 2. Violin plots presenting the score distribution of selected hallmark pathways for the African green monkey COVID dataset, calculated by SiPSiC (left) and AUCell (right). FDRs of T tests between significantly differential distributions are shown in red. (**A**) The complement pathway in alveolar cells. (**B**) Interferon alpha in B cells. (**C**) Interferon gamma in B cells. (**D**) MYC targets V2 in CD8+ T cells.

As stated above, SiPSiC found 31 and 12 pathways to be up- and down-regulated in the alveolar cells of the human COVID dataset, respectively. AUCell, in comparison, found 29 and 9 pathways to be up- and down-regulated in those cells, respectively (Supplementary Table S1). Out of these pathways, 26 and 8 pathways were upregulated or downregulated according to both methods, respectively. SiPSiC was again more accurate than AUCell in scoring the complement pathway for these cells- While SiPSiC showed upregulation of the complement system in the COVID cells group (FDR < $7.3*10^{-16}$), AUCell predicts higher complement pathway score in the control cells, although with borderline statistical significance (FDR < 0.035; Figure 3A). Furthermore, as mentioned above, SiPSiC showed upregulation of TGF beta in activated B cells and the G2M checkpoint pathway in CD8+ T cells. However, AUCell skipped completely 19 and 24 pathways out of 50 for the activated B and CD8+ T cells, respectively, while failing to achieve statistical significance for these two pathways (Figure 3B-C) or any of the remaining pathways.

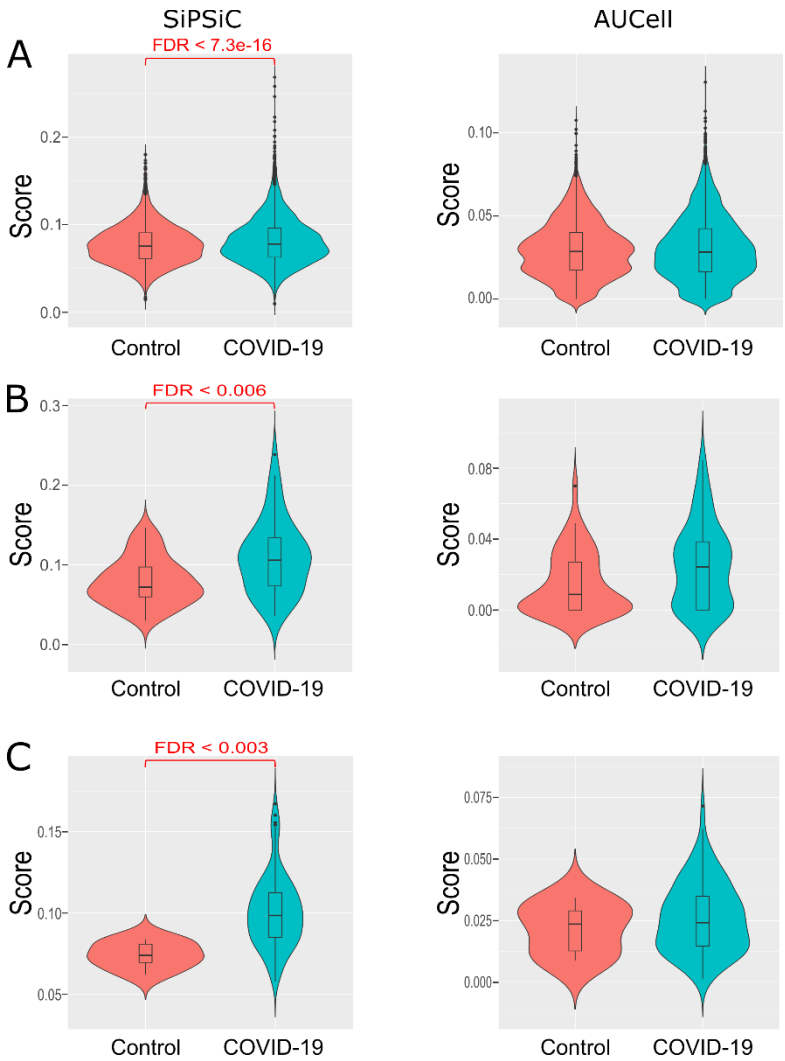

Figure 3. Violin plots presenting the score distribution of selected hallmark pathways for the human COVID dataset, calculated by SiPSiC (left) and AUCell (right). FDRs of T tests between significantly differential distributions are shown in red. (**A**) The complement pathway in alveolar cells. (**B**) TGF beta signaling in B cells. (**C**) The G2M checkpoint pathway in CD8+ T cells.

An intriguing finding is that the inflammatory response, TGFβ signaling and TNFα signaling via NFκB were all downregulated pathways in the COVID group of the human alveolar cells compared to control, both according to SiPSiC and AUCell, suggesting perhaps a counterintuitive reaction of infected epithelial cells suppressing immune response.

Comparing the results the two methods produced for the glioblastoma data, we again see higher sensitivity of SiPSiC in capturing the complement pathway. Garofano et al. showed upregulation of the hallmark complement pathway in the GPM cluster of cells, which was enriched in MES-like cells. SiPSiC agrees with this finding, showing upregulation of the complement pathway in the MES-like cell state compared to all other cell states. However, AUCell found downregulation of this pathway in the MES-like state compared to the AC-like state (Figure 4A). SiPSiC also identified significant upregulation of the inflammatory response pathway (FDR < $1.2*10^{-34}$) in MES-like cells, where AUCell did not reach statistical significance comparing the MES-and AC-like cell states (p = 0.38, FDR = 0.42; Figure 4B). Of note, these results of SiPSiC are also supported by evidence showing that the mesenchymal subtype of glioblastoma is correlated with high levels of both immune markers and infiltration of immune cells (Verhaak et al., 2010; Q. Wang et al., 2017). Furthermore, the hallmark unfolded protein response pathway is upregulated in the PPR cell cluster, which is enriched in NPC and OPC like cells. AUCell placed the MES-like cell state significantly above all other states and the NPC and AC-like states last, however SiPSiC placed the OPC-like state first, a statistically significant result, and the NPC-like state above the AC-like state, showing better concordance with the findings of Garofano et al (Figure 4C).

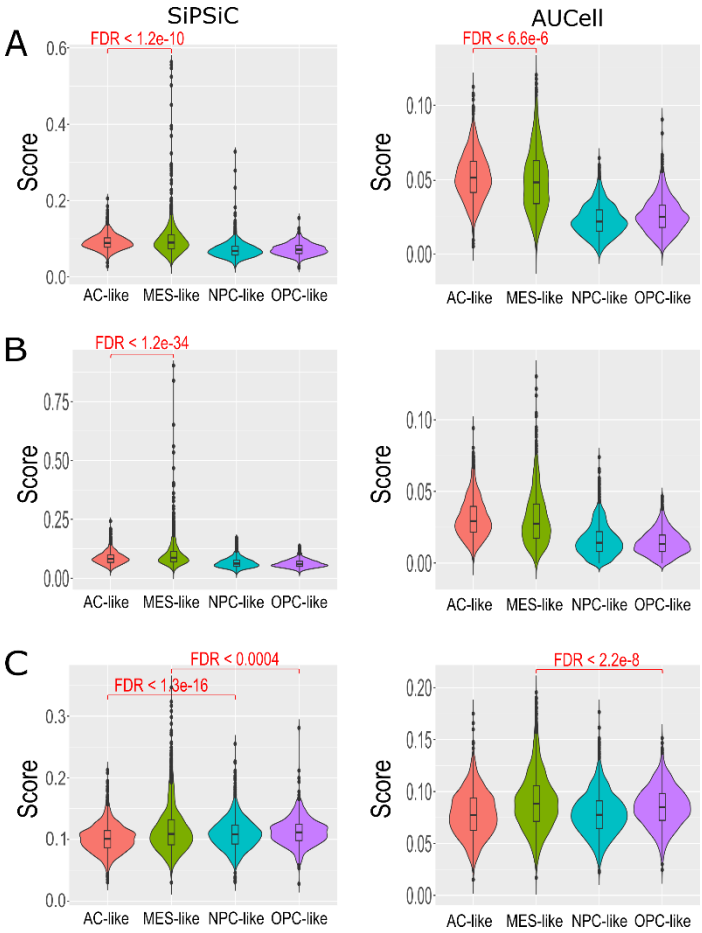

Figure 4. Score distribution of selected hallmark pathways for the glioblastoma dataset, calculated by SiPSiC (left) and AUCell (right). FDRs of T tests between significantly differential distributions are shown in red. (**A**) The complement pathway. While in SiPSiC the MES-like meta module (cell state) was higher than all other states, in AUCell it was significantly lower than the AC-like state. (**B**) The inflammatory response pathway. (**C**) The unfolded protein response pathway. AUCell placed the MES-like state above the OPC-like state, while SiPSiC placed the OPC-like state first.

Overall, these results demonstrate that SiPSiC is more sensitive than AUCell, and is able to pick real, biologically meaningful results that AUCell misses. One major reason for this is that AUCell scores a pathway in a given cell regardless of a) the transcriptome of other cells in the data; b) the transcript

counts of the different genes (as only the ranking is considered); and c) genes not in the top 5% highly expressed genes. As opposed to that, SiPSiC uses the expression of all the genes normalized by their expression across other cells (see Methods). This allows us both to utilize the information in a larger set of genes (not just the top 5%), and to be more sensitive in detecting more subtle differences (by relying on normalized expression). This is especially important if information on only a small number of genes in the gene set is available, or if many of the genes are not highly expressed, making it harder to determine a robust score for this gene set. For example, when AUCell scores the MYC V2 pathway in T cells of the monkey COVID-19 dataset, the median score of both the control and COVID cell groups is zero, indicating that no genes of this pathway are among the top 5% in most cells. However, using the normalized count values of this pathway's genes in the different cells, SiPSiC managed to unveil the upregulation of the pathway in the active COVID-19 group (3-days), in accordance with prior knowledge as discussed above. A similar case is the DNA repair pathway in the alveolar cells of the human COVID dataset, where the median score of both the COVID and control cell groups is zero when scored by AUCell, whereas SiPSiC found upregulation of this pathway in the COVID group (FDR < 0.004; Supplementary Table S1). This finding of SiPSiC is supported by recent evidence suggesting that similar to other coronaviruses, SARS-CoV-2 can also induce DNA damage (Pánico et al., 2022) and that drugs targeting DNA damage response can block SARS-CoV-2 replication (Garcia et al., 2021), suggesting that SARS-CoV-2 not only increases DNA damage in the infected cells, but also relies on the cells' reaction to it. We reason, then, that meaningful results can also be obtained by using genes that are not among the top transcribed genes, which are ignored by AUCell.

Discussion

scRNA-seq is a powerful tool to interrogate cellular heterogeneity, allowing researchers to comprehensively query changes in biological processes in high resolution. However, computational methods to infer activity of biological processes are still lacking. As scRNA-seq data are constantly being produced and shared, and as their scale and accuracy grow, it is of high value to further optimize existing methods in the field and develop novel, better ones, driving the discovery of unfamiliar biological processes and shedding light on poorly understood phenomena.

In this paper we present SiPSiC, a novel method for inferring pathway scores from scRNA-seq data. We demonstrate its high sensitivity, accuracy, and consistency with existing knowledge across different data types, including findings often missed by the original conventional analyses of these data. We demonstrate that SiPSiC scores can be used to cluster the cells and compute their UMAP projections in a manner that better captures the biological underpinnings of tissue heterogeneity. We also show its superiority over AUCell in dissecting the functional differences between distinct groups of cells, especially when the data are sparse and partial. We reason that this superiority is attributed to the use of actual gene expression, as well as to factoring by the rank of average expression over all cells, both ignored by AUCell. To allow comparable contribution of different genes, we normalize the expression of each gene. However, to take advantage of the higher information content of highly expressed genes, the higher the total expression of a gene is, the more it contributes to the pathway score. Yet the way our algorithm works poses two potential pitfalls. First, since we use gene expression for score normalization, rather than ranks, the scores may change substantially depending on the units in which data are presented. Using logarithms of the counts, for example, is bound to change the effect size (group median differences) and statistics. Thus, based on our analyses, we recommend providing SiPSiC with TPM or CPM values and not logTPM or logCPM. Second, SiPSiC is more susceptible to outliers, again due to the use of actual gene expression values. Since the final pathway score is a weighted average of all the single gene scores, outlier scores in a fraction of the genes may pull the final score towards them, possibly leading to false positives or negatives. We therefore recommend to properly filter both genes and cells before applying SiPSiC.

Acknowledgments

This work was supported by the European Research Council Horizon 2020 grant 949029 (YD).

References

Afzali, B., Noris, M., Lambrecht, B. N., & Kemper, C. (2021). The state of complement in COVID-19. *Nature Reviews Immunology 2021 22:2*, *22*(2), 77–84. https://doi.org/10.1038/s41577-021-00665-1

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z. K., Wouters, J., & Aerts, S. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, *14*(11), 1083–1086. https://doi.org/10.1038/nmeth.4463

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*(5), 411–420. https://doi.org/10.1038/nbt.4096

Chen, Z., Jin, N., Narasaraju, T., Chen, J., McFarland, L. R., Scott, M., & Liu, L. (2004). Identification of two novel markers for alveolar epithelial type I and II cells. *Biochemical and Biophysical Research Communications*, *319*(3), 774–780. https://doi.org/10.1016/j.bbrc.2004.05.048

Chi, W., & Deng, M. (2020). Sparsity-Penalized Stacked Denoising Autoencoders for Imputing Single-Cell RNA-seq Data. *Genes*, *11*(5), 532. https://doi.org/10.3390/genes11050532

Choi, E. Y., Park, H. H., Kim, H., Kim, H. N., Kim, I., Jeon, S., Kim, W., Bae, J.-S., & Lee, W. (2020). Wnt5a and Wnt11 as acute respiratory distress syndrome biomarkers for severe acute respiratory syndrome coronavirus 2 patients. *European Respiratory Journal*, *56*(5), 2001531. https://doi.org/10.1183/13993003.01531-2020

Drier, Y., Sheffer, M., & Domany, E. (2013). Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(16), 6388–6393. https://doi.org/10.1073/pnas.1219651110

Engler, J. R., Robinson, A. E., Smirnov, I., Hodgson, J. G., Berger, M. S., Gupta, N., James, C. D., Molinaro, A., & Phillips, J. J. (2012). Increased Microglia/Macrophage Gene Expression in a Subset of Adult and Pediatric Astrocytomas. *PLoS ONE*, *7*(8), e43339. https://doi.org/10.1371/journal.pone.0043339

Ferreira-Gomes, M., Kruglov, A., Durek, P., Heinrich, F., Tizian, C., Heinz, G. A., Pascual-Reguant, A., Du, W., Mothes, R., Fan, C., Frischbutter, S., Habenicht, K., Budzinski, L., Ninnemann, J., Jani, P. K., Guerra, G. M., Lehmann, K., Matz, M., Ostendorf, L., … Mashreghi, M. F. (2021). SARS-CoV-2 in severe COVID-19 induces a TGF-β-dominated chronic immune response that does not target itself. *Nature Communications 2021 12:1*, *12*(1), 1–14. https://doi.org/10.1038/s41467-021-22210-3

Garcia, G., Sharma, A., Ramaiah, A., Sen, C., Purkayastha, A., Kohn, D. B., Parcells, M. S., Beck, S., Kim, H., Bakowski, M. A., Kirkpatrick, M. G., Riva, L., Wolff, K. C., Han, B., Yuen, C., Ulmert, D., Purbey, P. K., Scumpia, P., Beutler, N., … Arumugaswami, V. (2021). Antiviral drug screen identifies DNA-damage response inhibitor as potent blocker of SARS-CoV-2 replication. *Cell Reports*, *35*(1), 108940. https://doi.org/10.1016/j.celrep.2021.108940

Garofano, L., Migliozzi, S., Oh, Y. T., D'Angelo, F., Najac, R. D., Ko, A., Frangaj, B., Caruso, F. P., Yu, K., Yuan, J., Zhao, W., Luisa Di Stefano, A., Bielle, F., Jiang, T., Sims, P., Suvà, M. L., Tang, F., Su, X. D., Ceccarelli, M., … Iavarone, A. (2021). Pathway-based classification of glioblastoma uncovers

a mitochondrial subtype with therapeutic vulnerabilities. *Nature Cancer*, *2*(2), 141–156. https://doi.org/10.1038/s43018-020-00159-4

Gnanaprakasam, J. N., & Wang, R. (2017). MYC in Regulating Immunity: Metabolism and Beyond. *Genes*, *8*(3), 88. https://doi.org/10.3390/genes8030088

Groom, J. R., & Luster, A. D. (2011). CXCR3 in T cell function. *Experimental Cell Research*, *317*(5), 620–631. https://doi.org/10.1016/j.yexcr.2010.12.017

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

Iser, I. C., Pereira, M. B., Lenz, G., & Wink, M. R. (2017). The Epithelial-to-Mesenchymal Transition-Like Process in Glioblastoma: An Updated Systematic Review and In Silico Investigation. *Medicinal Research Reviews*, *37*(2), 271–313. https://doi.org/10.1002/med.21408

John, G. R., Shankar, S. L., Shafit-Zagardo, B., Massimi, A., Lee, S. C., Raine, C. S., & Brosnan, C. F. (2002). Multiple sclerosis: Re-expression of a developmental pathway that restricts oligodendrocyte maturation. *Nature Medicine*, *8*(10), 1115–1121. https://doi.org/10.1038/nm781

Karam, B. S., Morris, R. S., Bramante, C. T., Puskarich, M., Zolfaghari, E. J., Lotfi-Emran, S., Ingraham, N. E., Charles, A., Odde, D. J., & Tignanelli, C. J. (2021). mTOR inhibition in COVID-19: A commentary and review of efficacy in RNA viruses. *Journal of Medical Virology*, *93*(4), 1843–1846. https://doi.org/10.1002/jmv.26728

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N., & Sergushichev, A. (2021). Fast gene set enrichment analysis. *BioRxiv*, 060012. https://doi.org/10.1101/060012

Lamouille, S., Xu, J., & Derynck, R. (2014). Molecular mechanisms of epithelial–mesenchymal transition. *Nature Reviews Molecular Cell Biology*, *15*(3), 178–196. https://doi.org/10.1038/nrm3758

Liao, M., Liu, Y., Yuan, J., Wen, Y., Xu, G., Zhao, J., Cheng, L., Li, J., Wang, X., Wang, F., Liu, L., Amit, I., Zhang, S., & Zhang, Z. (2020). Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nature Medicine*, *26*(6), 842–844. https://doi.org/10.1038/s41591-020-0901-9

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, *1*(6), 417–425. https://doi.org/10.1016/j.cels.2015.12.004

Melms, J. C., Biermann, J., Huang, H., Wang, Y., Nair, A., Tagore, S., Katsyv, I., Rendeiro, A. F., Amin, A. D., Schapiro, D., Frangieh, C. J., Luoma, A. M., Filliol, A., Fang, Y., Ravichandran, H., Clausi, M. G., Alba, G. A., Rogava, M., Chen, S. W., … Izar, B. (2021). A molecular single-cell lung atlas of lethal COVID-19. *Nature*, *595*(7865), 114–119. https://doi.org/10.1038/s41586-021-03569-1

Nasi, A., McArdle, S., Gaudernack, G., Westman, G., Melief, C., Rockberg, J., Arens, R., Kouretas, D., Sjölin, J., & Mangsbo, S. (2020). Reactive oxygen species as an initiator of toxic innate immune responses in retort to SARS-CoV-2 in an ageing population, consider N-acetylcysteine as early therapeutic intervention. *Toxicology Reports*, *7*, 768–771. https://doi.org/10.1016/j.toxrep.2020.06.003

Neftel, C., Laffy, J., Filbin, M. G., Hara, T., Shore, M. E., Rahme, G. J., Richman, A. R., Silverbush, D., Shaw, M. L., Hebert, C. M., Dewitt, J., Gritsch, S., Perez, E. M., Gonzalez Castro, L. N., Lan, X., Druck, N., Rodman, C., Dionne, D., Kaplan, A., … Suvà, M. L. (2019). An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell*, *178*(4), 835-849.e21. https://doi.org/10.1016/j.cell.2019.06.024

Noureen, N., Ye, Z., Chen, Y., Wang, X., & Zheng, S. (2022). Signature-scoring methods developed for bulk samples are not adequate for cancer single-cell RNA sequencing data. *ELife* , *11*(e71994). https://doi.org/10.7554/eLife

Pánico, P., Ostrosky-Wegman, P., & Salazar, A. M. (2022). The potential role of COVID-19 in the induction of DNA damage. *Mutation Research/Reviews in Mutation Research*, *789*, 108411. https://doi.org/10.1016/j.mrrev.2022.108411

Sanz, I., Wei, C., Jenks, S. A., Cashman, K. S., Tipton, C., Woodruff, M. C., Hom, J., & Lee, F. E.-H. (2019). Challenges and Opportunities for Consistent Classification of Human B Cell and Plasma Cell Populations. *Frontiers in Immunology*, *10*. https://doi.org/10.3389/fimmu.2019.02458

Speranza, E., Williamson, B. N., Feldmann, F., Sturdevant, G. L., Pérez-Pérez, L., Meade-White, K., Smith, B. J., Lovaglio, J., Martens, C., Munster, V. J., Okumura, A., Shaia, C., Feldmann, H., Best, S. M., & de Wit, E. (2021). Single-cell RNA sequencing reveals SARS-CoV-2 infection dynamics in lungs of African green monkeys. In *Sci. Transl. Med* (Vol. 13). http://stm.sciencemag.org/

Tzankov, A., Bourgau, C., Kaiser, A., Zimpfer, A., Maurer, R., Pileri, S. A., Went, P., & Dirnhofer, S. (2005). Rare expression of T-cell markers in classical Hodgkin's lymphoma. *Modern Pathology*, *18*(12), 1542–1549. https://doi.org/10.1038/modpathol.3800473

Verhaak, R. G. W., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., Alexe, G., Lawrence, M., O'Kelly, M., Tamayo, P., Weir, B. A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., … Hayes, D. N. (2010). An integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR and NF1. *Cancer Cell*, *17*(1), 98. https://doi.org/10.1016/J.CCR.2009.12.020

Wang, C., Zhang, C.-J., Martin, B. N., Bulek, K., Kang, Z., Zhao, J., Bian, G., Carman, J. A., Gao, J., Dongre, A., Xue, H., Miller, S. D., Qian, Y., Hambardzumyan, D., Hamilton, T., Ransohoff, R. M., & Li, X. (2017). IL-17 induced NOTCH1 activation in oligodendrocyte progenitor cells enhances proliferation and inflammatory gene expression. *Nature Communications*, *8*(1), 15508. https://doi.org/10.1038/ncomms15508

Wang, H., Cai, H., Ao, L., Yan, H., Zhao, W., Qi, L., Gu, Y., & Guo, Z. (2016). Individualized identification of disease-associated pathways with disrupted coordination of gene expression. *Briefings in Bioinformatics*, *17*(1), 78–87. https://doi.org/10.1093/bib/bbv030

Wang, Q., Hu, B., Hu, X., Kim, H., Squatrito, M., Scarpace, L., deCarvalho, A. C., Lyu, S., Li, P., Li, Y., Barthel, F., Cho, H. J., Lin, Y. H., Satani, N., Martinez-Ledesma, E., Zheng, S., Chang, E., Sauvé, C. E. G., Olar, A., … Verhaak, R. G. W. (2017). Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell*, *32*(1), 42-56.e6. https://doi.org/10.1016/j.ccell.2017.06.003

Wang, S., Sdrulla, A. D., diSibio, G., Bush, G., Nofziger, D., Hicks, C., Weinmaster, G., & Barres, B. A. (1998). Notch Receptor Activation Inhibits Oligodendrocyte Differentiation. *Neuron*, *21*(1), 63–75. https://doi.org/10.1016/S0896-6273(00)80515-2

Zhang, L., Wang, H., Zhu, J., Ding, K., & Xu, J. (2014). FTY720 reduces migration and invasion of human glioblastoma cell lines via inhibiting the PI3K/AKT/mTOR/p70S6K signaling pathway. *Tumor Biology*, *35*(11), 10707–10714. https://doi.org/10.1007/s13277-014-2386-y