ANS: Adjusted Neighborhood Scoring to improve assessment of gene signatures in single-cell RNA-seq data

Laure Ciernik[1,6], Agnieszka Kraft[1,4,5], Florian Barkmann[1], Josephine Yates[1,2,4], and Valentina Boeva[1,2,3,4,*]

[1]ETH Zurich, Department of Computer Science, Institute for Machine Learning, 8092 Zurich, Switzerland;

[2]ETH AI Center, ETH Zürich, Zurich Switzerland;

[3]INSERM, U1016, Cochin Institute, CNRS UMR8104, Paris Descartes University, Paris, France;

[4]Swiss Institute of Bioinformatics (SIB), Zurich, Switzerland;

[5]University Hospital Zurich, Clinic of Thoracic Surgery, 8092 Zurich;

[6]Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany.

[*]Correspondence should be addressed to Valentina Boeva, valentina.boeva@inf.ethz.ch

Abstract

In the field of single-cell RNA sequencing (scRNA-seq), gene signature scoring is integral for pinpointing and characterizing distinct cell populations. However, challenges arise in ensuring the robustness and comparability of scores across various gene signatures and across different batches and conditions. Addressing these challenges, we evaluated the stability of established methods such as Scanpy, UCell, and JASMINE in the context of scoring cells of different types and states. Additionally, we introduced a new scoring method, the Adjusted Neighbourhood Scoring (ANS), that builds on the traditional Scanpy method and improves the handling of the control gene sets. We further exemplified the usability of ANS scoring in differentiating between cancer-associated fibroblasts and malignant cells undergoing epithelial-mesenchymal transition (EMT) in four cancer types and evidenced excellent classification performance (AUCPR train: 0.95-0.99, AUCPR test: 0.91-0.99). In summary, our research introduces the ANS as a robust and deterministic scoring approach that enables the comparison of diverse gene signatures. The results of our study contribute to the development of more accurate and reliable methods for analyzing scRNA-seq data.

Introduction

High-throughput single-cell RNA sequencing (scRNA-seq) is a powerful technology to profile the transcriptome at the cellular level[1]. Gene expression from single cells allows for quantifying cell types and states, analyzing inter- and intra-sample heterogeneity, discovering cell differentiation trajectories, and constructing gene regulatory networks[2]. The interpretation of scRNA-seq data can be challenging due to high dimensionality, batch effects, dropout, and transcriptional noise[3]. To enhance the reliability of methodologies based on scRNA-seq measurements, it is vital that the methods address the inherent variability and noise in these data. This is especially important when evaluating cell states and programs through gene signature scoring. Thus, robust scoring tools are essential for the reliable characterization of cell states and programs.

Gene signature scoring on scRNA-seq data measures the activity of biological processes and transcriptional cell states for each cell profiled. Gene signatures are sets of genes associated with a specific biological signal, transcriptional state, cell type, or cell state and are used as a surrogate representation for a biological phenotype[4]. During gene signature scoring, the information in the high-dimensional gene expression data is condensed into a per-cell score. The scores can then be used for downstream analysis, *e.g.*, for correction of biological effects or cell-type annotation, allowing a better understanding of the cell composition of samples studied.

Numerous techniques for gene signature scoring in bulk RNA-seq and scRNA-seq have been developed lately. Recent studies, however, showed that signature scoring methods created for bulk RNA-seq, such as ssGSEA[5] and GSVA[6], were not fit for scRNA-seq data since they were more susceptible to dropouts and in tumors suffered from imbalanced expressions of genes in cancer cells versus non-malignant cells[7]. Therefore, researchers were advised to use single cell-specific methods such as signature scoring methods of Scanpy[8,9] and Seurat[10,11], UCell[12], and JASMINE[7].

The scoring methods implemented in the scRNA-seq analysis packages Scanpy[8,9] and Seurat[10,11] are based on the procedure first described by Tirosh *et al.*[13], which computes scores by averaging the difference in expression of the signature and control genes. Due to slight implementation differences (Methods), we distinguish between the Scanpy and the Seurat methods in this study. Although widely used, these scoring approaches may show some drawbacks. As in Scanpy and Seurat, the control genes are selected by binning genes with comparable mean expression, the scores may be biased by the number of bins and the behavior of the mean expression curve.

Two other popular scoring approaches, UCell[12], which is an extension of AUCell[14], and JASMINE (Jointly Assessing Signature Mean and Inferring Enrichment), use rank statistics to increase the stability of results in case of strong technical variation and batch effects. UCell is solely based on Mann-Whitney U statistics. It is expected to be robust against technical variation as it only uses per-cell gene rank information. JASMINE computes scores by averaging the mean of signature gene ranks and an enrichment value[7]. The enrichment values can correspond to odds ratios or the likelihood, referred to as OR and LH in our benchmarks.

Although Noureen et al.[7] investigated the robustness of several signature scoring methods for bulk

and single-cell RNA-seq, assessing the methods' sensitivity and specificity on *in silico* datasets, the study did not assess the comparability of score ranges and score variability linked to differences in the discriminatory potential of genes constituting a signature. Further, a novel benchmark is needed to interrogate the robustness of gene signature scoring methods under the variation in sample cell composition and batch effects; such evaluation would ensure the methods' trustworthiness and usefulness in downstream analysis.

In this work, we benchmark and analyze the stability of the most popular cell scoring methods, Scanpy, Seurat, UCell, and JASMINE. We present an improvement of Tirosh's scoring method more resistant to single-cell technology biases than the original approach, which we term Adjusted Neighbourhood Scoring (ANS). We show that ANS is robust with regard to most influencing factors and returns comparable scores for multiple signatures. Specifically, although all scoring methods demonstrate resilience against variations in data composition and variability in signature qualities, they do not exhibit, except for ANS, comparable score ranges for gene signatures designed to discriminate similar cell types, such as memory and naive B cells. To showcase the usability of our gene signature scoring method in the challenging scenario of related gene signatures, we build a signature for mesenchymal-like cancer cells applicable across cancer types that discriminates malignant cells with mesenchymal-like properties from tumor-infiltrating fibroblasts.

Results

Adjusted Neighbourhood Scoring (ANS) method

We proposed to improve the widely used gene signature scoring method suggested by Tirosh *et al.*[13], which had been implemented in Seurat[10] and with minor modifications in Scanpy[8]. In brief, the Tirosh method computes the gene signature scores for each cell by subtracting a control expression from the signature genes' expressions. The major drawback of the Tirosh signature scoring method comes from the procedure of choosing control genes using fixed binning of genes according to their average gene expression (default: 25 bins, Suppl. Figure S1). As verified by our benchmarking, this procedure may create artifacts for the estimates of expression gain, specifically for highly expressed genes frequently used in gene signatures. Our proposed method, called Adjusted Neighborhood scoring (ANS), solved this issue by selecting control gene sets separately for each signature gene so that the averaged expression of the control gene set and the signature gene matches (Methods). The ANS method was implemented in Python and R (package https://github.com/lciernik/ANS_signature_scoring).

Design of the benchmark

We benchmarked ANS and other gene signature scoring methods, including the original Tirosh method, assessed the stability of the signature scores, and highlighted the importance of the choice of control genes. We compared eight gene signature scoring methods: JASMINE[7] (using the likelihood or odds-ratio settings, Jasmine_LH and Jasmine_OR, respectively), UCell[12], the original method by Tirosh et al.[13] implemented in Seurat[10] (Seurat), a similar method implemented in Scanpy[8] (Scanpy), its alternative that used all genes as control genes (Seurat_AG), an alternative that used least variable genes as control genes (Seurat_LVG), and our proposed method Adjusted Neighbourhood

Scoring (ANS). Like Seurat and Scanpy, Seurat_AG and Seurat_LVG first compute the average expression of the genes, sort them in increasing order, and divide them into 25 equally sized expression bins. While Seurat_AG selects all genes from an expression bin as control genes, Seurat_LVG chooses genes with the smallest dispersion. Conversely, ANS avoids binning and selects control genes in an average-expression neighborhood closest to the signature genes' average expression values (Methods). We implemented all methods originally developed for the R language in Python (Figure 1a) and showed their identical behavior (Figure 1b, Methods). The Python package complementing this work provides access to all implemented methods.

We benchmarked the methods in six different scenarios (Figure 1a, Table 1). The initial and final scenarios involved a preprocessed dataset of peripheral blood mononuclear cells[15] (PBMC), while the remaining four scenarios were carried out on preprocessed datasets of colorectal carcinoma[16] (CRC) and squamous cell carcinoma of the esophagus[1] (ESCC) (Methods). In the first scenario, we showed the implication of control gene selection on gene signature scoring. In the second and third, we investigated the influence of the data composition and batch effects on scoring. In the fourth and fifth scenarios, we analyzed the impact of the variability in the gene signature length on the resulting scores. In the final scenario, we examined the scores produced by different methods when scoring multiple gene signatures and evaluated the performance of cell score-based classifiers to annotate cell types or cell states.

We also evaluated the time requirements for the eight methods as a function of the increasing number of cells and as a function of the increasing number of genes in the signature (Suppl. Figure S2). All methods exhibited linear scalability with dataset size. For short signatures (up to 10 genes), Seurat, Scanpy, and ANS performed the fastest (<10 sec for 150,000 cells), while Seurat and Scanpy remained the quickest for longer signatures (1,000 genes). In contrast, UCell and Jasmine took approximately seven times longer, while ANS took about three times longer in this case. Nevertheless, these methods consistently operated within seconds to a few minutes, even with large signatures and datasets.

Table 1. Qualitative comparison of signature scoring methods based on their properties, influencing factors, and score comparability. ✓ means a scoring method fulfills property or resists an influencing factor. ✕ means the opposite. NA stands for a property or an influencing factor that cannot be evaluated for a scoring method. These results are discussed in detail in Results.

Scoring method	Determi-nistic	Optimal control genes selection	Independent of total cell number and cell type proportions	Robust to batch effect when scoring all samples together	Robust when scoring for small signatures	Robust to unrelated genes in the signature	Provides comparable scores for multiple cell type signatures
ANS: Adjusted Neighborhood Scoring	✓	✓	✕	✓	✓	✓	✓
Scanpy[8] : based on Tirosh *et al.*[13]	✕	✕	✕	✓	✓	✓	✕
Seurat[10]: based on Tirosh *et al.*[13]	✕	✕	✕	✓	✓	✓	✕

Scoring method	Determi-nistic	Optimal control genes selection	Independent of total cell number and cell type proportions	Robust to batch effect when scoring all samples together	Robust when scoring for small signatures	Robust to unrelated genes in the signature	Provides comparable scores for multiple cell type signatures
Seurat_AG: Seurat scoring with all genes as control	✓	×	×	✓	✓	✓	×
Seurat_LVG: Seurat scoring with least variable genes as control	✓	×	×	✓	✓	×	×
Jasmine_LH: JASMINE with likelihood computation[7]	✓	**NA**	×	✓	✓	✓	×
Jasmine_OR: JASMINE with odds-ratio computation[7]	✓	**NA**	×	✓	×	✓	×
UCell[12]	✓	**NA**	✓	✓	✓	✓	×

Control genes selection bias in gene signature scoring

All scoring methods based on Tirosh *et al.*[13], *i.e.*, all methods except JASMINE and UCell, use control genes for gene signature scoring. Typically, genes whose average expressions are similar to the average expression of the signature gene are used as control genes. Consequently, these control genes exhibit a mean average expression, referred to as the mean control value, that is the closest approximation to the average expression of the signature gene itself. The selection of an optimal control set is crucial, as inappropriate selection can introduce a bias and lead to unreliable scores for all cells. We compared strategies for the control gene selection of the methods ANS, Seurat, Seurat_AG, and Seurat_LVG across diverse preprocessed subsets of cell subtypes within the PBMC dataset for the top expressed genes where the effects of the control gene selection were expected to be the most pronounced; we chose the 8% of genes with the highest average expression values corresponding to the last two bins of the Tirosh-based methods (Figure 1c, Suppl. Figure S3, Methods). We scored each gene in the last two expression bins (top ~900 genes) for each scoring method. On homogeneous datasets, scores of single-gene signatures are expected to be distributed around zero. While we observed minimal score biases in the last but one expression bin, the last bin showed large biases in the score values for all published methods due to the significant variance of the genes' average expressions. While Seurat, Seurat_AG, and Scanpy first underestimated two-thirds and then overestimated the gene scores, Seurat_LVG overestimated almost all in the second expression bin. All methods showed notable score biases for the top fifty expressed genes due to their expression variability, typically excluded in our ANS analyses. ANS, by construction, showed overall the smallest induced bias and thus used the optimal control sets (Figure 1c).

The influence of cell type proportions and batch effects on gene signature scoring

The mean control values and, thus, the choice of control genes depend on the dataset composition of the tissue or sample analyzed. Consequently, when utilizing methods that employ control genes, we anticipated greater variability in gene signature scoring values when scoring was conducted on a per-sample basis compared to a scenario when all cells from all samples were assessed collectively.

To highlight this effect, we calculated scores of 100-gene signatures associated with the malignant cell phenotype of CRC and ESCC (Methods). We compared the distributions of the mean scores per sample and cell type under two scenarios: when scoring was conducted on a per-sample basis and on all samples together (Figure 1d, Suppl. Figure S4). We observed a reduction of score variance for all scoring methods and both CRC and ESCC datasets, except for Seurat_LVG and UCell in the ESCC dataset, when scoring was performed on cells of all the samples together (Suppl. Figure S5), leading to more comparable score ranges over all samples. Additionally, we assessed how scoring on all samples together contributed to diminishing the batch effects for the Tirosh-based scoring methods. For this, we evaluated the differences in score distributions across two batches representing different sequencing chemistry, called SC3Pv2 and SC3Pv3 (Figure 1d). Batch effects significantly affected the score values of all Tirosh-based scoring methods when the scoring procedure was performed on a per-sample basis. In contrast, we observed decreased distribution shifts between chemistry-related batches when scoring on all samples. Therefore, our results suggest that scoring should be performed on all cells in the dataset simultaneously for all methods using control gene sets, including ANS.

The influence of the variability in the gene signature length on gene signature scoring

We evaluated the effects of the signature length on gene signature scoring. We assessed the robustness of scoring methods to small signatures on CRC and ESCC. We compared the scoring methods based on the number of required signature genes associated with the malignant cell phenotype to perfectly classify, in terms of area under the receiver–operator curve (AUCROC), malignant and non-malignant cells based on scores (Methods). Classifiers based on the scores provided by all methods achieved perfect classification accuracy for ESCC using maximally twelve of the most differentially expressed genes (lowest adjusted p-value and highest $\log_2$FC) (Suppl. Figure S4). Similar results were observed on CRC, where all but the Jasmine_OR method achieved perfect classification performance with at most 13 genes (Figure 1e). A classifier based on the Jasmine_OR scores required about 11 times more genes in the gene signature to achieve the perfect classification accuracy. Overall, our results show that starting at 15 genes, the gene signature length shows limited impact on the gene signature score, except for Jasmine.

Robustness to noise in gene expression signatures

Gene signature scoring methods can be affected by the presence of irrelevant genes in the signature. To quantify the effect of introduced noise in a controlled setting, we started with a “pure” 100-gene signature associated with the malignant cell phenotype and iteratively replaced signature genes with

random genes (Methods). All scoring methods, except Seurat_LVG, tolerated up to 85% noise in the signature in CRC, requiring only approximately 15 informative genes in the 100-gene signature while achieving an AUCROC above 0.9 (Figure 1f, Suppl. Figure S4). Seurat_LVG required at least twice as many correct genes in a signature. Similar results are observed for ESCC, where Seurat_LVG also performed the worst. In summary, we observed high robustness to noise, *i.e.*, to the presence of irrelevant genes in the signature, for all scoring methods except Seurat_LVG.

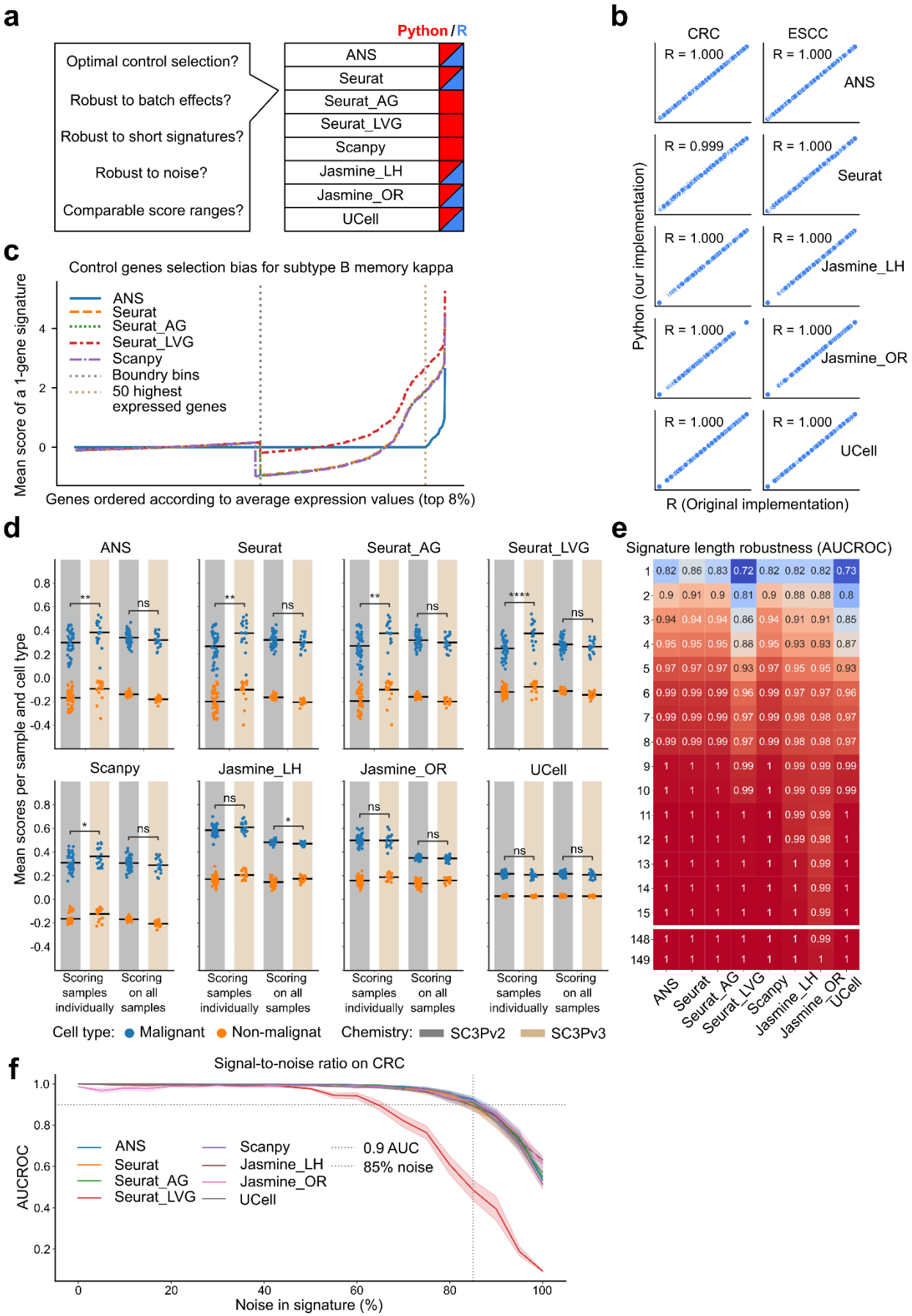
a
Python / R
Optimal control selection?
Robust to batch effects?
Robust to short signatures?
Robust to noise?
Comparable score ranges?
ANS
Seurat
Seurat_AG
Seurat_LVG
Scanpy
Jasmine_LH
Jasmine_OR
UCell
b
CRC
ESCC
Python (our implementation)
R (Original implementation)
c
Control genes selection bias for subtype B memory kappa
Mean score of a 1-gene signature
Boundry bins
50 highest expressed genes
Genes ordered according to average expression values (top 8%)
d
Mean scores per sample and cell type
Scoring samples individually
Scoring on all samples
Cell type: Malignant Non-malignat
Chemistry: SC3Pv2 SC3Pv3
e
Signature length robustness (AUCROC)
f
Signal-to-noise ratio on CRC
AUCROC
0.9 AUC
85% noise
Noise in signature (%)

Figure 1. Performance and robustness analysis of gene signature scoring methods. **a,** Benchmark questions for the gene signature scoring methods and availability of the methods in Python (blue) and R (red). **b,** The scores obtained from R and Python implementations are compared for three cancer datasets (CRC, ESCC, LUAD). Discrepancies in Seurat scores between R (AddModuleScore of Seurat) and Python are attributed to randomization in the scoring method. **c,** Control gene selection bias of the scoring methods for the top 8% of highly expressed genes in B memory kappa cells within the PBMC dataset. The x-axis illustrates genes sorted by their average expression levels, while the y-axis depicts the mean score across all cells for a single-gene signature. Vertical dashed lines indicate the expression bin boundary and the top 50 highly expressed genes. The bias of a scoring method is indicated by how far the mean score of a gene deviates from zero. **d,** The influence of dataset composition and batch effect on scoring CRC cells using a 100-gene signature associated with malignant cells. Each dot represents the mean score for all cells within a sample, grouped by cell type (malignant in blue or non-malignant in orange), sequencing chemistry type (SC3Pv2 in grey or SC3Pv3 in beige), and scoring mode (scoring all the samples together or individually). The black horizontal bar represents the mean value of all dots within each group. **e,** The minimum required signature length per scoring method for perfect classification (AUCROC of 1) when discriminating between malignant and non-malignant cell scores in CRC. All scoring methods, except Jasmine_OR, need at most 13 genes for perfect discrimination. **f,** The robustness to noise in a signature when discriminating between malignant and non-malignant cells in CRC. Starting with a 100-gene signature, we performed iterative replacements of genes with random genes that exhibited a $|\log_2 FC| < 0.5$ and an adjusted p-value > 0.1 during differential gene expression (DGEX) analysis between malignant and non-malignant cells. The standard variation was calculated from 20 simulation runs. P-value annotation: ns: $p > 0.05$, *: $0.01 < p <= 0.05$, **: $1e-3 < p <= 0.01$, ***: $1e-4 < p <= 1e-3$, ****: $p <= 1e-4$.

Comparability of score ranges when scoring for multiple signatures

One of the applications of gene signature scoring is to measure the activation of different transcriptional states and use the scores to make statements about the association of cells to these states. However, interpreting cell scores and/or assignment of cells to states requires scores with high information quantity and comparable score ranges for signatures describing exclusive cell states or programs. To assess the information quantity and the comparability of score ranges produced by ANS and other scoring methods, we investigated the usage of scores for two classification tasks on the PBMC dataset[15]. The first task, termed the *easy* task, aimed to score cells with non-overlapping gene signatures associated with three clearly distinguishable cell types, B cells, monocytes, and NK cells, and differentiate between these cell types based on the signature scores. The second task, termed the *hard* task, sought to distinguish between B-cell subtypes using subtype-specific non-overlapping gene signatures (Figure 2a-b). For each cell type in the *easy* and *hard* task, we selected the corresponding gene expression signature described in Methods. We evaluated the information quantity of scores using cross-validated supervised logistic regression for each scoring method while also addressing the scale imbalance by maximizing the difference in classification performance between scaled and original scores (Methods).

For the *easy* task, for the true cell types, most of the scoring methods produced scores in comparable ranges, *i.e.*, the highest scores were observed for cells of matched cell types (Figure 2a). JASMINE, the sole method demonstrating suboptimal performance for this task, exhibited overlapping score ranges for the B-cell and two other signatures when scored on the B cells. Upon hard labeling based on the scores, *i.e.*, assigning each cell a label associated with the maximum score value (Methods),

four out of eight scoring methods achieved classification metrics close to one, including our ANS method (Figure 2c, Suppl. Table S1).

In the *hard* task, all methods except for ANS produced scores in non-comparable ranges for the matched cell types (Figure 2b). The Tirosh-based scoring methods (Seurat, Seurat_AG, Seurat_LVG, and Scanpy) showed the highest scores for the B memory cell signature independent of the true cell type. Application of UCell to the B intermediate cells showed higher scores for the B memory cell and slightly higher for the B naive cell signatures than for the matched B-intermediate-cell signature. On the contrary, JASMINE scored the highest for the B-intermediate-cell signature for B intermediate cells and failed to show distinguishable score ranges in the remaining cells. When we applied the hard labeling strategy based on the gene signature scores, ANS performed the best among all scoring methods tested, followed by UCell (Figure 2d, Suppl. Table S1).

We observed high information quantity of scores produced by Seurat-based methods (Figure 2e); however, classification based on the original scores failed due to large scale imbalance (Figure 2f, Suppl. Table S1). Jasmine's scores, on the other hand, exhibited the lowest information quantity, as this method showed the lowest performance in the supervised labeling setting**.** We also observed that cell state assignment via clustering based on the gene signature scores in the unsupervised setting was more accurate when gene signature scores were calculated with methods based on Tirosh *et al.*, than with Jasmine or UCell; we refer to the metric as rediscovery score (Suppl. Table S1, Supplementary methods). We obtained similar results highlighting the advantages of ANS compared to other methods when considering signatures without the exclusion of overlapping genes (Suppl. Figure S6, Suppl. Table S2). In summary, while all methods excel in discrimination based on gene signature scores of distinct cell types, the task becomes significantly more demanding when it comes to discriminating signals from different cell states.

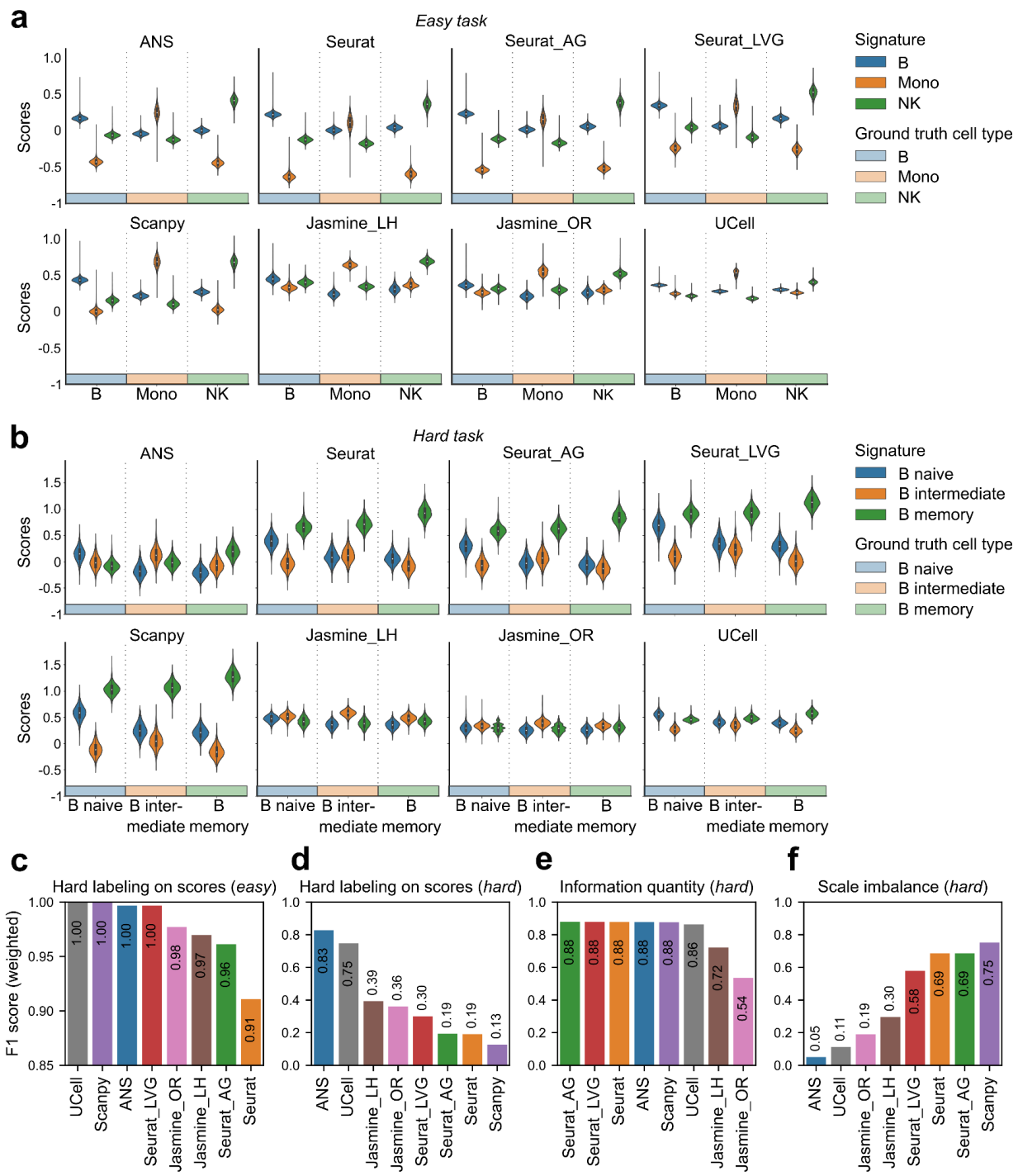

Figure 2. Most scoring methods yield scores in the ranges useful for the cell type annotation via direct score comparison in the *easy* setting; in the *hard* setting of differentiating B-cell subtypes, only the ANS method provides scores in the ranges useful for the cell subtype annotation. **a,** Scores for non-overlapping cell-type-specific signatures (B cells, monocytes, and NK cells) calculated by scoring method on each experimental cell type. **b**, The score distributions for non-overlapping B-cell subtype signatures are calculated by each scoring method on each cell subtype. **c**, Accuracy of the cell-type assignments based on the original scores produced by the methods in the *easy* scenario. **d**, Accuracy of the cell-type assignments based on the original scores produced by the methods in the *hard* scenario. **e**, Information quantity of signatures in the *hard* scenario, *i.e.*, cell-type classification accuracy in the supervised cross-validated setting. **f**, Score scale imbalance in the *hard* setting measured through the performance differences between classification accuracy values based on the scaled and original scores.

Devising cancer cell-specific EMT signature: a case study for the ANS application

As we demonstrated the robustness and performance of the ANS method in the challenging scenario of gene signature scoring on related cell states, we chose to showcase the ability of ANS to help extract biologically relevant gene signatures in a similarly difficult setting. Specifically, we applied ANS to address the problem of the lack of malignant cell-specific EMT signature applicable across cancer types. The difficulty in constructing such a gene signature comes from the high similarity of transcriptional states of malignant EMT-exhibiting cells with malignant non-EMT cells, on the one hand, and with non-malignant mesenchymal cells, such as cancer-associated fibroblasts (CAFs), on the other hand. Building a cancer cell-specific signature for EMT would allow scoring bulk RNA-seq tumor datasets to quantify the degree of EMT transformation in human tumors while minimizing the bias induced by the presence of non-malignant cells of mesenchymal origin in the tumors.

Indeed, when we explored the behavior of previously published EMT signatures using the ANS scoring method on single cells from the ESCC[1], lung adenocarcinoma[17] (LUAD), CRC[16] and breast carcinoma[18] tumors, we observed that the pan-cancer EMT signatures, such as the Hallmark EMT[19], the pEMT signature[20], and five other EMT signatures[21–25], resulted in extremely high scores in CAFs in addition to the mesenchymal-like (MES-like) malignant cells (Figure 3a and Suppl. Figure S7). This bias was likely due to the high expression in CAFs of such mesenchymal markers as *ACTA2*, *CDH2*, *FN1*, *SNAI2*, and *VIM* for the Hallmark EMT signature[26], and *PDPN*, *TNC*, and *VIM* for the pEMT signature[27,28]. The other cancer EMT gene signatures produced high scores also for other cell types, *e.g.*, the signatures of Foroutan *et al.*[24], Gröger *et al.*[21], Mak *et al.*[23], and Tan *et al.*[22] resulted in high scores for pericytes in CRC and ESCC, and the former also scored high in endothelial cells for all datasets (Suppl. Figure S7).

To build a cancer cell-specific EMT signature, we leveraged two public scRNA-seq datasets, the ESCC[1] and LUAD[17], for training and then validated the signature on the single cells of CRC[16] and breast carcinoma[18]. To devise a cancer-specific EMT signature that scores high for MES-like malignant cells but low for non-malignant CAFs and malignant cells not undergoing EMT, we extracted genes with the abovementioned properties in ESCC and LUAD single-cell datasets (Methods). The devised cancer cell-specific EMT signature included many known EMT-related genes. We found 8 of the 135 signature genes intersecting with the Hallmark EMT and the pEMT signatures (Figure 3b). Moreover, 43 out of 135 genes appeared to be in at least one of the previously considered pan-cancer EMT signatures (Suppl. Table S3). By construction, the devised signature scored highly for the MES-like malignant cells in these two cancer types used for training (Figure 3a). We observed its strong discriminative power on the validation datasets CRC and BREAST (Figure 3a, Suppl. Table S4). The accuracy of our gene signature discriminating MES-like malignant cells versus all remaining cells measured with the area under the precision-recall curve (AUCPRC) on CRC and breast carcinoma datasets was 0.701 and 0.511, respectively; it was slightly outperformed on CRC only by the gene signature proposed by Hollern *et al.*[25] (AUCPRC 0.707) (Suppl. Table S4). However, this signature performed sub-optimally in our ESCC and LUAD training datasets (AUCPRC 0.028 versus 0.592 of our gene signature in ESCC, and 0.244 versus 0.513 in LUAD).

To investigate the clinical significance of the proposed cancer cell-specific EMT signature, we used bulk RNA-seq and clinical information from The Cancer Genome Atlas (TCGA) datasets. Specifically, we were interested in analyzing the link between the extent of cancer cells expressing EMT and the molecular subtypes of analyzed samples. Calculation of cancer cell-specific phenotypes using bulk RNA-seq data poses a significant advantage in cases when corresponding single-cell data is not available for a number of reasons. First, since EMT plays a crucial role as one of the main drivers of tumor plasticity and invasiveness, accurately measuring its activity in both TME and cancer cells is essential for designing successful targeted therapies. Second, an association between the cancer cell-specific EMT activity and histological subtypes may highlight the role of the cancer cell EMT phenotype in promoting, likely more invasive, molecular subtypes, and expand our understanding thereof. We selected 24 TCGA datasets for which more than one histological subtype was available, scored the cancer cell-specific EMT signature, and associated the scores with histological subtypes (independently in each cancer type). We found our signature to be significantly associated with the histological subtype of 10 cancer datasets (Figure 3c, Supp. Table S5). These results highlight the biological relevance of the proposed signature and its potential utility in further exploration of cancer molecular subtypes.

This example demonstrated how gene signature scoring on single-cell data could be applied to design meaningful gene signatures and annotate cells according to their transcriptional states for further biological exploration.

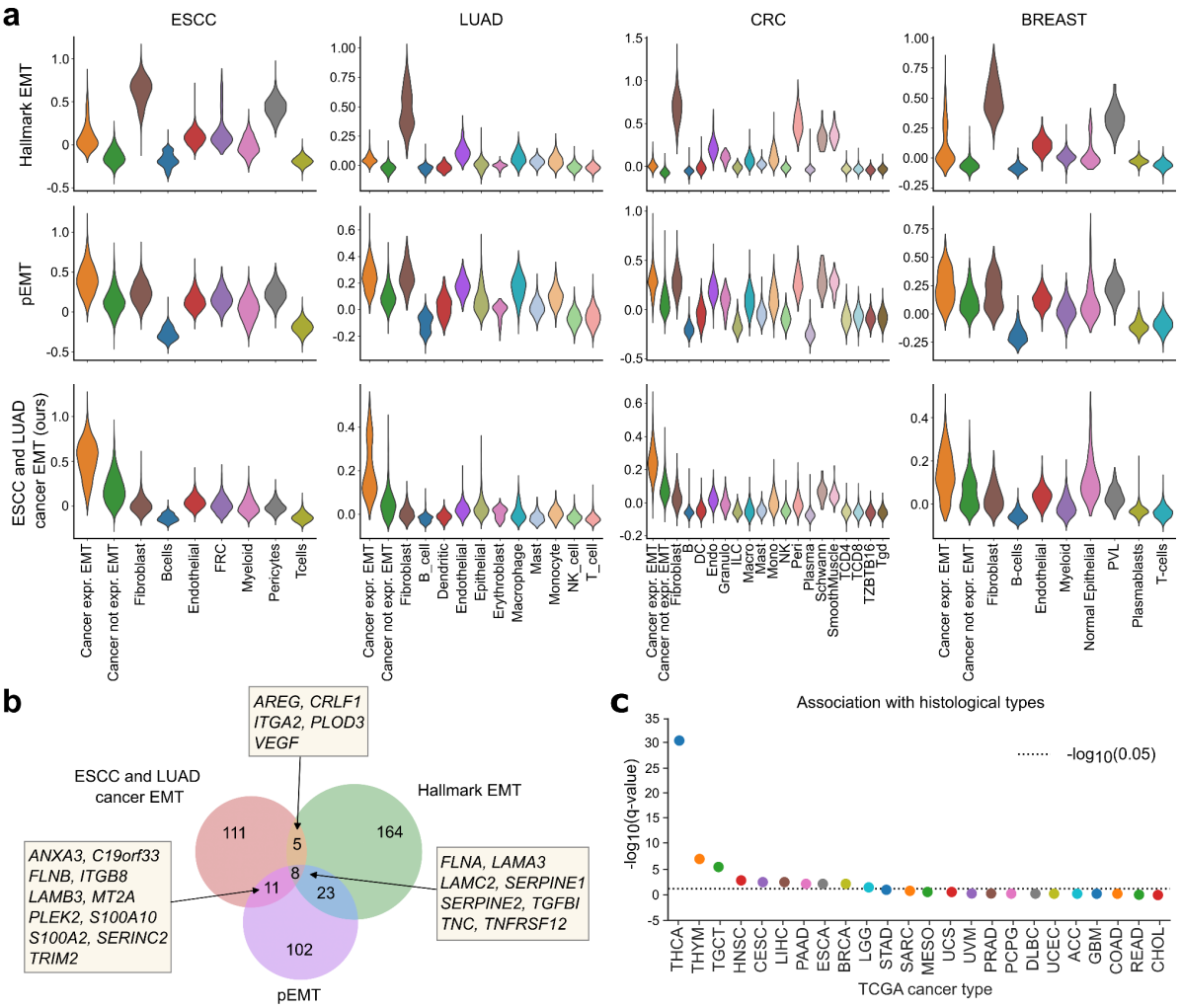

Figure 3. Application of ANS to devise EMT signature specific to malignant cells. **a**, Score distributions of the Hallmark EMT, the pEMT, and our proposed ESCC- and LUAD-specific cancer EMT gene signatures in different cell types of ESCC, LUAD, CRC, and breast carcinoma. **b**, Venn diagram for the overlap in gene lists for the Hallmark EMT, pEMT signatures, and the signature we designed. **c**, Association of ESCC- and LUAD-specific cancer EMT signature scores and histological subtypes in TCGA (only cancer types with at least 1 histotype are included). Significant associations (qvalue < 0.05) are represented by dots above the dotted line. The FDR-adjusted p-values from the Kruskal-Wallis test were -log_10 transformed and sorted in decreasing value, with a higher y-value indicating higher significance.

Discussion

In the present work, we developed a deterministic and robust gene signature scoring method, Adjusted Neighborhood Scoring (ANS), which improves the approach introduced by Tirosh *et al.*[13]. In contrast to the Tirosh approach, ANS avoids binning genes by mean gene expression values and, by construction, chooses the optimal control gene sets to normalize gene expression values for technical effects. We demonstrated that the issue of sub-optimal stability of ANS and all methods based on the Tirosh approach, such as Scanpy and Seurat when scoring gene signatures on samples with variable proportions of different cell types, can be solved by scoring all cells simultaneously from all samples in the dataset.

We compared ANS with other scoring methods developed for single-cell data to demonstrate the robustness and stability of the proposed scoring procedure. All scoring methods were shown to be stable upon scoring small signatures and to be robust against noise, allowing up to 85% of noise in the signature if the remaining genes belong to the most differentially expressed gene discriminating the cell types tested. ANS was the top-performing approach in terms of providing high information quantity and comparable ranges of calculated scores for matched gene signatures and cell types. The remaining Tirosh-based scoring approaches showed high information quantity in their scores. However, they suffered from large scale imbalances, possibly influenced by the biased selection of the control sets and lack of score normalization. We also showed that scale imbalance could be avoided in Tirosh-based scoring methods through unsupervised clustering, converting signature scores to cluster assignments correlating with the matched cell state. Nonetheless, it is worth noting that the underlying assumption of exclusive state activation within individual cells could be limiting in certain situations. Furthermore, we have included all single-cell RNA-seq signature scoring techniques within a signature scoring package, enabling the utilization of methods originally implemented in R in Python.

We showcased the use of the ANS scoring method to derive a gene signature specific to MES-like malignant cells that we validated in CRC and breast carcinoma. Building a cancer cell-specific signature for EMT could allow scoring bulk RNA-seq tumor datasets to quantify the degree of EMT transformation in human tumors while minimizing the bias induced by the presence of non-malignant cells of mesenchymal origin in the tumors. Our ESCC- and LUAD-specific cancer EMT signature outperformed others in distinguishing MES-like malignant cells from CAFs across all datasets. When comparing MES-like malignant cells with other malignant cells, our signature ranked among the top four in terms of AUCPR for ESCC, LUAD, and CRC but placed sixth for breast carcinoma (Suppl. Table S3). Notably, the Hallmark EMT[19] signature displayed the highest discriminatory power in this context. The difficulty in discrimination may arise from the signature's reliance on epithelial markers, lacking strong EMT-related markers, which results in closer scores between malignant epithelial cells. Our cancer MES-like signature included a transcriptional factor, which so far has not been considered in any of the used pan-cancer EMT signatures: *PITX1*. *PITX1* encodes a member of the RIEG/PITX homeobox family and plays a crucial role in organ development and left-right asymmetry[29,30]. Previous research on *PITX1* in cancer has revealed cancer-type-specific effects. While upregulation of *PITX1* has been shown to be associated with tumor progression in lung adenocarcinoma[31], kidney renal clear cell carcinoma[32], breast[33], epithelial ovarian[34], and prostate[35] cancer, interestingly, in other cancer types, such as melanoma[36], ESCC [37], osteosarcoma[38] and head and neck squamous cell carcinoma[39], down-regulation of *PITX1* has been found to enhance tumor progression. Other EMT-related genes exclusively present in our signature included two long-non-coding RNA genes: *ABHD11-AS1* and *BCYRN1*. *ABHD11-AS1* has previously been suggested as a prognostic biomarker for pancreatic cancer[40], while *BCYRN1* was shown to promote cell migration and invasion in lung and colorectal cancer[41–44]. Additionally, the protein-coding genes *FAM83A*, *ITGA3*, *ITGB4*, *L1CAM*, *MUC16*, and *SAA1* have been reported to promote cancer progression in CRC[45,46], lung cancer[47,48], head-neck squamous cell carcinoma[49,50], breast[51], ovarian[52,53] and endometrial cancer[54]. Lastly, our analysis has revealed that the identified cancer cell-specific EMT signature corresponds to histological subtypes in several TCGA datasets, highlighting the potential role of EMT-expressing cancer cells in their molecular characteristics. This finding highlights

the biological significance of the proposed signature; however, future work is needed to verify its clinical applicability.

We acknowledge the limitations inherent in our study. The evaluation of scoring methods in the *easy* task scenario was conducted based on cell-type classifications. However, due to the exclusive assignment of each cell to a specific cell type and the absence of annotations for intermediate or transitioning states, our benchmarking did not encompass the assessment of performance in scenarios involving cells undergoing such intermediate states. Also, when applying our method to establish the cancer-specific EMT signature in the three datasets, LUAD, ESCC, and CRC, we relied on the selection of the cancer cells expressing previously published EMT-related genes. In contrast, in the basal-like breast cancer dataset, we relied on the original annotation of cell states and used the GM3 state strongly associated with EMT to label the EMT-presenting malignant cells. The selection procedure might have biased the downstream analysis.

In summary, this work presents an improved version of the scoring method proposed by Tirosh[13], called Adjusted Neighborhood Scoring (ANS). ANS is a deterministic and robust scoring method that outputs comparable scores for multiple gene expression signatures. The ANS method is implemented in a user-friendly way to be integrated into R and Python single-cell analysis pipelines. By employing this technique, one can confidently evaluate cellular programs and develop highly specific cell state signatures, as exemplified by the cancer EMT signature.

Methods

Dataset and Preprocessing

We used five different datasets in this work: four cancer datasets and one peripheral blood mononuclear cell (PBMC) dataset with cell type annotation supported by multimodal analysis[15]. The scRNA-seq datasets for cancer included colorectal carcinoma[16] (CRC) with 60 samples, squamous cell carcinoma of the esophagus[1] (ESCC) with 54 samples, lung adenocarcinoma[17] (LUAD) with 19 samples and breast carcinoma[18] with 7 basal-like samples (Suppl. Figure S8). The PBMC datasets consisted of data from three time points for each of the eight patients, *i.e.*, 24 samples.

Preprocessing scRNA-seq data is crucial to address inherent biases and enable downstream analysis. Cancer datasets and the PBMC dataset were preprocessed differently. For the cancer datasets, we used the preprocessing module of `CanSig`[55] with the default parameter as the first preprocessing step (https://github.com/BoevaLab/CanSig). In brief, we applied standard quality control metrics and removed potential misannotated cells using CNV inference; the preprocessing steps are extensively described in the original paper. Additionally, we filtered genes based on a minimum expression in 1% of cells, and normalized cell reads using a shift logarithm $\log_2(x+1)$ with the mean total raw count per cell as the target sum, as described and implemented by Ahlmann-Eltze and Huber[56]. Regarding the PBMC dataset, we created several sub-datasets based on cell type annotations. The first sub-dataset consisted of B-cells, monocytes, and natural killer cells, while the second subdataset contained B-cell subtypes (B-naive, B-intermediate, and B-memory cells). The remaining four were single-cell-type subsets with the following cell subtypes: B memory kappa, CD8 TM2, CD14 Mono, and NK 3. All sub-datasets were preprocessed as follows: Initially, we filtered low-quality reads according to the guidelines provided in chapter 6.3 of the book "Single-cell Best Practices" by Heumos *et al.*[57]. A threshold of 5 median absolute deviations (MADs) was selected for filtering covariates, including total counts, number of genes with positive counts in a cell, and the percentage of mitochondrial counts. Similar to the cancer datasets, we filtered genes based on a minimum expression in 1% of cells and normalized the cell reads using the shift logarithm with the mean total raw count per cell as the target sum.

Control genes selection in Tirosh-based gene signature scoring and ANS

The methods implemented in the scRNA-seq analysis packages `Scanpy`[8] (in Python) and `Seurat`[10] (in R) were built upon the procedure first described by Tirosh *et al.*[13] Here, we recall this procedure in a formal manner. Let $\mathbf{X} \in \mathbb{R}_{\geq 0}^{n \times p}$ be the preprocessed expression matrix for n cells and p genes, generally representing log-transformed normalized read counts. Let $S = \{s_1, \cdots, s_m\}$ be a given gene signature, *i.e.*, a set of m genes. Let $\bar{\mathbf{g}} = (\bar{g}_i)_{i=1:p} \in \mathbb{R}_{\geq 0}^{p}$ be the average expression vector for each gene i over all cells ordered by increasing average expression (i.e., $\bar{g}_i \leq \bar{g}_{i+1}$ for any i).

In Tirosh-based gene signature scoring methods, $\bar{\mathbf{g}}$ is split into 25 equally sized bins, called expression bins. For each signature gene s, we sample c (control size) genes from s's expression bin. Let C_s be the set of control genes for the signature gene s. The score for a cell j is then computed as follows:

$$\text{score}_j = \frac{1}{m}\sum_{s\in S}\left(\mathbf{X}_{j,s} - \frac{1}{c}\sum_{k\in C_s}\mathbf{X}_{j,k}\right)$$

In this work, we kept the main idea of the scoring approach proposed by Tirosh *et al*. but suggested and validated three alternatives to select control genes.

When implementing the first two alternatives, we preserved the gene binning approach. In the "all genes as control genes" alternative (Seurat_AG), we selected all non-signature genes from an expression bin as control genes. Note that Seurat_AG was not using the control genes size parameter c. In the Least Variable control Genes alternative (Seurat_LVG), we first computed the least variable genes for each expression bin. The least variable genes were calculated with Scanpy's `highly_variable_genes`[58] method (flavor 'seurat') and selected by taking the c genes with the smallest dispersion.

In the third alternative, called Adjusted Neighborhood scoring (ANS), the control genes for a signature gene are selected based on the average expression neighborhood. Let $\tilde{\mathbf{g}}$ be equal to $\bar{\mathbf{g}}$, but excludes the signature genes S. Let $C = \{C_k\}_{k=1}^{(p-m)-c+1} = \{\{\tilde{g}_k, \cdots, \tilde{g}_{k+c}\}\,|1 \leq k \leq (p-m-c+1)\}$ be the set of windows of size c (control size) of $\tilde{\mathbf{g}}$. For each signature gene s, we computed the control set as follows:

$$C_s = \arg\min_{C_k\in C}\left|\bar{g}_s - \frac{1}{c}\sum_{\tilde{g}_l\in C_k}\tilde{g}_l\right|.$$

In other words, for each signature gene, we selected c control genes around this gene whose average mean expression closely matched the mean expression of the signature gene. Of note, ANS excluded signature genes within the $\frac{c}{2}$ genes with the highest average expression to avoid invalid control gene selection.

Evaluation of scoring methods

To compare and evaluate gene signature scoring methods, we utilized cell type classification/annotation with scores. We calculated scores for cell-type-specific signatures and evaluated score distributions on matched and unmatched cell types using provided annotations. Cancer datasets are scored for malignant cell-associated signatures (binary classification) (Supplementary Methods), while PBMC datasets score multiple signatures linked to specific cell types (multiclass classification). For validation, matched cell types serve as the positive class (higher scores) and the remaining cells as the negative class (lower scores). Classification metrics evaluated include AUCROC for balanced datasets and AUCPRC for unbalanced datasets. We compute the F1-score weighted, Jaccard-score weighted, and balanced accuracy for hard labeling of the cells. We used the implementation of `scikit-learn`[59] (version 1.2.0.).

Scoring Methods in R and Python

We implemented the R methods UCell (https://github.com/carmonalab/UCell) and JASMINE (https://github.com/NNoureen/JASMINE), as well as the original method proposed by Tirosh et al.[13],

in Python. Additionally, we provided the ANS method in both Python and R. We showed that the scoring methods yield identical scores for both Python and R. We selected a 100-gene signature for malignant cells based on the genes with the lowest adjusted p-value and the highest $\log_2$FC (Supplementary Methods). We then scored this signature on preprocessed datasets of two cancer types: CRC and ESCC. To ensure consistency, we converted these datasets into `SingleCellExperiments`[60] and employed the original R versions of the scoring methods. Notably, the R implementation of the Seurat method based on the original version of Trisoh *et al.*, known as `AddModuleScore`[61], is included in the scRNA-seq package `Seurat`[15]. To evaluate the concordance between the R and Python scores, we computed the Pearson correlation coefficient using the `scipy` package[62] (version 1.9.3) in Python (Figure 1b). The implementations can be found in our software package (https://github.com/lciernik/ANS_signature_scoring).

Benchmark experiments

In the first part of the study, we compared the properties and resistance to influencing factors of eight gene signature scoring methods: JASMINE[7] using the likelihood or odds-ratio (Jasmine_LH and Jasmine_OR), UCell[12], the original method by Tirosh et al.[13] implemented in Seurat[10,11] (Seurat), the method implemented in Scanpy[8] (Scanpy), Seurat using all genes as control genes (Seurat_AG), Seurat using the least variable genes as control genes (Seurat_LVG), and our proposed method Adjusted Neighbourhood Scoring (ANS). Unless stated differently, we chose signatures of desired lengths associated with the malignant cell phenotype (Supplementary Methods) for the cancer datasets CRC and ESCC. This selection was based on identifying genes with the smallest adjusted p-value and the largest $\log_2$FC.

Control gene selection bias on Tirosh-based gene signature scoring

The Tirosh-based scoring methods rely on the selection of control genes. However, Seurat, Seurat_AG, Seurat_LVG, and Scanpy can introduce biases in scoring when there is substantial variability in gene expression averages within expression bins, which is especially visible in the top 8% of the highest expressed genes (Suppl. Figure S1). The ANS method was motivated to reduce the issue of bias induced by control genes. We compared the strategies for the control gene selection of the methods ANS, Seurat, Seurat_AG, and Seurat_LVG across the four preprocessed single-cell-type PBMC sub-datasets: B memory kappa, CD8 TM2, CD14 Mono, and NK 3 (Dataset preprocessing in Methods). For each dataset, we scored each gene in the top 8% of the highest expressed genes and, thus, converted the expression matrix to a score matrix. On homogeneous datasets, scores of a gene are expected to be distributed around zero. We visualized the mean scores and standard variation for each gene in increasing average expression order with the `lineplot` function of the plotting package `searborn`[63] (Figure 1c, Suppl. Figure S3).

The influence of the cell type proportions and batch effects on gene signature scoring

The two following experiments explored the influence of dataset composition on gene signature scoring. The first experiment investigated the influence of different cell type proportions. The second experiment tested the influence of batch effects on scoring methods. Both experiments used the preprocessed CRC dataset and a 100-gene signature for malignant cells (Supplementary Methods) with the smallest adjusted p-values and the highest $\log_2$FCs. The CRC dataset contains samples

sequenced with two different sequencing chemistries. We used the different sequencing chemistries as surrogates for batch effects. We scored each sample individually and the entire dataset. For each scoring method, we averaged the cell scores for each sample, cell type (malignant or non-malignant), scoring mode (scoring samples individually or all together), and sequencing chemistry (SC3Pv2 or SC3Pv3). We compared the variance between the sample score averages for each scoring method and scoring mode in the first experiment and observed reduced or equal variance when scoring on the entire dataset. In order to evaluate the impact of batch effects, we utilized Mann-Whitney U testing to compare the sample score averages of malignant cells between the two sequencing chemistries for each scoring method and scoring mode. We used the Python package `statannotations`[64] (version 0.4.4.) to compute the Mann-Whitney U tests and add statistical annotations in Figure 1d. Scoring samples individually showed significant differences in score ranges between chemistries. We performed identical experiments on the preprocessed ESCC dataset, excluding the batch effect component because of using uniform sequencing chemistries for all samples in those datasets (Suppl. Figure S4).

The influence of signature length on gene signature scoring

The next experiment aimed to assess the robustness of scoring methods to small signatures. Specifically, we sought to determine the minimum number of genes necessary to achieve perfect discrimination between malignant and non-malignant genes using scores, thereby enabling binary classification. For each scoring method, we followed a systematic approach. We began by selecting a base signature, as described in Supplementary Methods. Subsequently, we removed genes for which a valid control set could not be constructed, specifically those belonging to the $\frac{c}{2}$ smallest or highest expressed genes. The remaining genes were then sorted based on ascending adjusted p-value and descending $\log_2$FC. We iterated over the signature to ascertain the number of genes required, progressively expanding the range of elements considered from the initial subset containing only the first element until the final iteration encompassing the entire list. Utilizing the methodology outlined in the Method section *Evaluation of scoring methods*, we computed the area under the receiver operating characteristic curve (AUCROC) for each signature length, using both the scores and malignancy annotations. We ceased computing as soon as an AUCROC value of 1 was achieved. We conducted the analysis for the preprocessed CRC and ESCC datasets (Figure 1e, Suppl. Figure S4).

Robustness to noise in gene expression signature

To compare the robustness of the gene signature scoring methods to noise, we utilized a 100-gene base signature, which achieved an AUCROC of 1 for all scoring methods and iteratively replaced genes by noise. We selected noise genes based on their adjusted p-value > 0.01 and their |log2FC| <= 0.5 during DGEX (Supplementary methods), indicating that these genes lacked statistical significance in distinguishing between malignant and non-malignant cells. During each simulation run, we iteratively and randomly substituted the signature genes with noise genes, starting from the pure signature and progressing until the entire signature consisted of noise genes. We conducted 20 simulation runs for each gene signature scoring method to ensure comprehensive results. The experiment was performed on preprocessed CRC and ESCC datasets (Figure 1f, Suppl. Figure S4).

Comparable score ranges

To assess the information quantity and ranges of scores when scoring for multiple signatures, we considered two tasks on the PBMC dataset[15]. The dataset included cell type annotations at three granularity levels, with level one being the broadest cell type level and level three the one with the most subtypes. We used the published table[65] of differentially expressed genes per subcell type level 3 to get the cell-type specific signatures. For the *easy* task, we selected B cells, Monocytes, and NK cells (cell types on level 1) and preprocessed them as described in the first section of Methods. We chose the cell type-specific signatures for the cell types on level 1 of the hierarchy by unioning the sets of differentially expressed genes belonging to all subtypes of finer granularity. For the *hard* task, we repeated the procedure on B cell subtypes (cell subtypes on level 2). We selected B intermediate, memory, and naive cells and preprocessed them as described in the first section of Methods. We again picked the subcell type-specific signatures by unioning the differentially expressed genes belonging to any subtype of each B-cell subtype. To enhance discrimination power, we removed genes occurring in multiple signatures, as their inclusion could introduce bias by assigning higher scores to all cells in discrimination tasks. We removed signature genes with the highest average gene expression from the top c/2 genes to allow valid control gene selection for ANS. The classification metrics F1-score weighted, Balanced accuracy, and Jaccard-score weighted from metrics `scikit-learn` (version 1.2.0) were used. We compared the classification performance in four scenarios: classification based on the original scores, measurement of the scores' information quantity based on cross-validated supervised logistic regression, measurement of the scale imbalance through the performance difference between classification based on the scaled and original scores, and the measurement of the rediscovery score through unsupervised clustering (Suppl. Table S1). During hard labeling based on scores, we assigned the cell type to each cell based on the highest-scoring signature associated with that cell type. To evaluate the information contained within the scores, we employed a cross-validated supervised logistic regression approach. This allowed us to estimate the classification performance by assessing the predictive power of the scores. We utilized stratified k-fold cross-validation (`StratifiedKFold`, k=10, random state=42) to estimate the classification performance, employing the aforementioned classification metrics. For hyperparameter tuning of the regularization parameter C, we performed another round of stratified k-fold cross-validation (`StratifiedKFold`, k=10, random state=42)). The regularization parameter C was tested across a range of values (0.001, 0.01, 0.1, 1, 10) using the `LogisticRegressionCV` function from `scikit-learn`. Key parameters included the metric (f1-score weighted), max_iter (10,000), and class_weight ("balanced"). We assessed the scale imbalance within the signature scores by comparing the classification performance between a scaled version of the scores and the original ones. The methodology for achieving classification performance on the scaled scores was as follows. We sought to determine the optimal scaling values (a_1, a_2, a_3) for each signature (S_1, S_2, S_3) that maximized the classification metrics, e.g., $F1\text{-score}_{\text{scaled}} = F1\text{-score}\ (a_1 + \text{scores}\ S_1, a_2 + \text{scores}\ S_2, a_3 + \text{scores}\ S_3)$, where $a_i \in [\min(\text{scores}\ S_1, \text{scores}\ S_2, \text{scores}\ S_3), \max(\text{scores}\ S_1, \text{scores}\ S_2, \text{scores}\ S_3)]$ with a stride of 0.05 for the *easy* and 0.025 for the *hard* task. Finally, we estimated the rediscovery score through unsupervised clustering. We applied a post-processing step based on Gaussian mixture models[66] (GMMs) to transform scores into probabilities representing cell activities in each cell (Supplementary Methods). Thus, the cell type for each cell was assigned based on the highest

probability signature. We also considered the equivalent experiment with overlapping signatures in Suppl. Table S2.

Case study: selection of cancer EMT cells and cancer-specific EMT signature establishment

For the entire case study, we used ANS for evaluation with control sets of size c=100. The generation of the cancer-specific EMT signature consisted of two main steps. In the first step, we classified the malignant cells as EMT-expressing or non-expressing. Based on this classification, we explored the genetic differences between these two groups and between EMT cancer cells and cancer-associated fibroblasts.

Classification of cancer EMT cells

As a first step in the signature establishment, we selected cancer cells expressing EMT, called cancer EMT cells. We computed scores for ESCC with the dataset-specific mesenchymal signature[1]. We scored the entire dataset and the subset of malignant cells. Cancer cells being within the 10% highest scoring cells on the entire dataset and in the subset were classified as cancer EMT cells. The 3’860 cancer EMT cells out of 41’399 cancer cells (9.32%) and 178’109 total cells (2.17%), stem from 53 out of 54 samples. For LUAD and CRC, no dataset-specific mesenchymal signatures were reported. Therefore, we selected the cancer cells expressing EMT as follows. We scored the datasets for seven pan-cancer EMT signatures: Hallmark EMT signature [19], pEMT gene module [20], and five other EMT signatures [21–25]. The pan-cancer EMT signatures were found on EMTome [67]. For each pan-cancer signature, we ranked the cancer cells by ANS scores and computed for each cell the median rank. Finally, 10% of cancer cells with the smallest median signature ranks were classified as cancer EMT cells. We selected the median to avoid outlier ranks of the pan-cancer signature. For the breast carcinoma dataset, we used the program labels provided by the original paper[18] and excluded cells falling in the “grey zone” of the EMT program expression ("GM3" in the original paper). We considered only samples assigned to the “Basal” subtype and defined the "grey zone" to be cells not labeled with “GM3”, but having a signature score for GM3 above 0.15 (scoring was applied on all samples similar in the original paper). In the LUAD dataset, 822 out of 8’205 cancer (10.02%) and 55’659 total (1.48%) cells were classified as EMT. All 19 samples contained cancer EMT cells. In the CRC dataset, 4’328 out of 43’252 cancer (10.01%) and 140’090 total (3.09%) cells were classified as cancer EMT cells distributed over all 60 samples. In the breast carcinoma dataset, 2031 out of 5115 cancer (39,71%) and 27458 total (7,40%) cells were classified as cancer EMT cells distributed over all 7 samples. Dataset compositions are shown in Suppl. Figure S8.

Establishment of a cancer-specific EMT signature in ESCC and LUAD

The signature establishment consisted of finding differentially expressed genes between cancer cells expressing EMT and CAFs as well as cancer cells not expressing EMT in ESCC and LUAD. The cancer EMT cell classification for all datasets is provided in the supplementary information. We first started by finding a cancer-specific signature for ESCC, then found one for LUAD. Finally, we unioned the ESCC and LUAD-specific cancer EMT signature and refined it again on ESCC. For differential gene expression (DGEX) analysis between two cell groups of interest, we used Wilcoxon rank-sum (Mann-Whitney U) tests [68] with tie correction and Benjamini-Hochberg correction implemented in

Scanpy's `rank_genes_groups`. Tests applied on scores for a single gene during refinement were done as described in the supplementary methods.

Cancer-specific EMT signature for ESCC: First, we created two subsets of the ESCC dataset: cancer EMT cells and CAFs and all cancer cells. Let A be the set of DGEX genes ($logFC \geq 2$ and an adjusted p-value < 0.001) between cancer EMT cells and CAFs. Let B be the set of DGEX genes ($logFC \geq 1.5$ and an adjusted p-value < 0.001) between cancer EMT cells and cancer cells not expressing EMT. We joined sets A and B, called gene set C, and removed all genes from C that are DGEX ($logFC \geq 2$ and an adjusted p-value < 0.001) in T- and Myeloid cells versus cancer EMT cells. The reduced gene set C, called gene set D, scored high for many cancer cells and thus required more refinement to distinguish the cancer EMT cells from the other cancer cells. Let the set X be all cancer cells that scored above 0.2 for the gene set D. We selected all DGEX genes ($logFC \geq 1.5$ and an adjusted p-value < 0.001), called gene set E, in X between cancer EMT cells and cancer cells not expressing EMT. We scored each gene in $E - D$ (n=451) independently and kept genes with significantly higher scores for cancer EMT cells than for CAFs, cancer cells not expressing EMT, and the rest of the cells (Supplementary Methods). We unioned the gene sets D and E, resulting in our ESCC-specific cancer EMT signature. The detailed procedure can be found in the Jupyter Notebook `find_cancer_emt_signature_ESCC.ipynb` in the [Supplementary material GitHub repository].

Cancer-specific EMT signature for LUAD: For the LUAD-specific cancer EMT signature, we followed a similar procedure as for ESCC. First we created two subsets of the LUAD dataset: cancer EMT cells and CAFs and all cancer cells. Let A be the set of DGEX genes ($logFC \geq 2$ and an adjusted p-value < 0.001) between cancer EMT cells and CAFs. Let B be the set of DGEX genes ($logFC \geq 1$ and an adjusted p-value < 0.001) between cancer EMT cells and cancer cells not expressing EMT. We unioned gene sets A and B, called gene set C. For each cell type X not equal to cancer EMT cells, we selected DGEX genes ($logFC \geq 1$ and an adjusted p-value < 0.001) of X versus cancer EMT cells and removed them from gene set C. The resulting gene set corresponded to our LUAD-specific cancer EMT signature. The detailed procedure can be found in the Jupyter Notebook `find_cancer_emt_signature_LUAD.ipynb` in the [supplementary material GitHub repository].

Joining cancer-specific EMT signatures for ESCC and LUAD and refinement on ESCC: To obtain a signature that can distinguish between cancer EMT cells and CAFs, and cancer EMT cells and cancer cells that do not express EMT, we unioned the ESCC- and LUAD-specific cancer EMT signatures. We scored for each gene in the unioned signature in ESCC (n=267) and only kept a gene if it had significantly higher scores for cancer EMT cells than for CAFs, cancer cells not expressing EMT, and the rest of the cells (signature refinement as described in Supplementary Methods). We excluded the insignificant genes and incorporated the LUAD-specific cancer EMT signature genes absent in the preprocessed ESCC dataset and the ESCC-specific cancer EMT signature genes absent in the preprocessed LUAD dataset. Suppl. Table S3 shows the 135 ESCC- and LUAD-specific cancer EMT signature genes. The detailed procedure of signature union and refinement can be found in the Jupyter Notebook `union_ESCC_and_LUAD_specific_EMT_signature_and_refine_on_ESCC.ipynb` in the [supplementary material github repository].

Validation of cancer-specific EMT signature

We applied the ESCC and LUAD-specific cancer EMT signature on the CRC and breast carcinoma datasets for validation. For all datasets, we computed the AUCPRC (*Evaluation of scoring methods* in Methods) for the signature scores between cancer EMT cells and CAFS, cancer cells not expressing EMT, and all other cells (rest) (Figure 3a).

Correlation of cancer-specific EMT signature to histological types

To analyze the link between de novo EMT signature and histological subtype of TCGA datasets, we downloaded pan-cancer normalized RSEM (RNA-Seq by Expectation Maximization) gene expression data from the GDC portal (https://portal.gdc.cancer.gov/). Clinical information on the histological subtypes was retrieved from Liu et al.[69]. The EMT signature was scored in each corresponding cancer type for which more than one histological subtype was available. The signature was scored using the *ssgsea* method from the `GSVA` R package[6]. The association between EMT scores and histological subtypes was evaluated using the Kruskal-Wallis test with FDR correction. The analysis was performed using R Statistical Software (v4.1.2; R Core Team 2021).

Statistical information

The statistical tools, methods, and threshold for each analysis are explicitly described with the results or detailed in the figure legends or Materials and Methods.

Data availability

The preprocessed datasets for CRC, ESCC, LUAD, breast carcinoma, and PBMC and the used signatures can be downloaded: https://drive.google.com/drive/folders/10L2gqapJbyOn_MbrZRHQG--n0Xj7wIyg?usp=sharing. The raw transcript count matrices can be found at GEO: GSE178341 for CRC, GSE176078 for breast cancer, and GSE160269 for ESCC. The raw transcripts for LUAD can be found with the accession number HRA000154 at http://bigd.big.ac.cn/gsa-human. The raw PBMC dataset can be downloaded at https://atlas.fredhutch.org/nygc/multimodal-pbmc/ . TCGA pan-cancer normalized RSEM gene expression data was downloaded from the GDC portal (https://portal.gdc.cancer.gov/). TCGA clinical information was retrieved from Liu et al.[69].

Code availability

The code for the experiments and visualizations can be found in the GitHub repository https://github.com/BoevaLab/ANS_supplementary_information. The Python package published in the GitHub repository https://github.com/lciernik/ANS_signature_scoring includes the implementations of all considered scoring methods. The repository also contains the R implementation of ANS.

Acknowledgments

This work was supported by the Stiftung für angewandte Krebsforschung (SAKF) at the University Hospital Zurich (to A.K.), Swiss National Science Foundation (SNF) (projects CRSII5_209524 and 205321_207931 to F.B. and J.Y.).

Author information

Authors and Affiliations

ETH Zurich, Department of Computer Science, Institute for Machine Learning, 8092 Zurich, Switzerland
Laure Ciernik, Agnieszka Kraft, Florian Barkmann, Josephine Yates, Valentina Boeva

ETH AI Center, ETH Zürich, Zurich Switzerland
Josephine Yates, Valentina Boeva

INSERM, U1016, Cochin Institute, CNRS UMR8104, Paris Descartes University, Paris, France
Valentina Boeva

Swiss Institute of Bioinformatics (SIB), Zurich, Switzerland
Agnieszka Kraft, Florian Barkmann, Josephine Yates, Valentina Boeva

University Hospital Zurich, Clinic of Thoracic Surgery, 8092 Zurich
Agnieszka Kraft

Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany
Laure Ciernik

Contributions

V.B., A.K., and L.C. conceived the study. J.Y. and F.B. preprocessed the cancer datasets and helped in the analysis and interpretation of results. L.C. developed the ANS scoring method and performed the analysis and interpretation of the benchmark and case study with the assistance of V.B. and A.K. A.K. conducted the association study of cancer EMT scores with histopathological types. L.C. wrote the paper with input from all authors.

Corresponding author

Correspondence to Valentina Boeva: valentina.boeva@inf.ethz.ch. ETH Zurich, Universitatstrasse 6, 8092 Zurich, Switzerland

Ethics declarations

Competing interests

The authors declare no competing interests.

References

1. Zhang, X. *et al.* Dissecting esophageal squamous-cell carcinoma ecosystem by single-cell transcriptomic analysis. *Nat. Commun.* **12**, 5291 (2021).
2. Zhao, Z.-H., Wang, X.-Y., Schatten, H. & Sun, Q.-Y. Single cell RNA sequencing techniques and applications in research of ovary development and related diseases. *Reprod. Toxicol.* **107**, 97–103 (2022).
3. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
4. Nevins, J. R. & Potti, A. Mining gene expression profiles: expression signatures as cancer phenotypes. *Nat. Rev. Genet.* **8**, 601–609 (2007).
5. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).
6. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
7. Noureen, N., Ye, Z., Chen, Y., Wang, X. & Zheng, S. Signature-scoring methods developed for bulk samples are not adequate for cancer single-cell RNA sequencing data. *Elife* **11**, e71994 (2022).
8. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
9. scanpy.tl.score_genes — Scanpy 1.10.0.dev documentation. https://scanpy.readthedocs.io/en/latest/generated/scanpy.tl.score_genes.html.
10. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
11. AddModuleScore function - RDocumentation. https://www.rdocumentation.org/packages/Seurat/versions/2.3.4/topics/AddModuleScore.
12. Andreatta, M. & Carmona, S. J. UCell: Robust and scalable single-cell gene signature scoring. *Comput. Struct. Biotechnol. J.* **19**, 3796–3798 (2021).

13. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).

14. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

15. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).

16. Pelka, K. *et al.* Spatially organized multicellular immune hubs in human colorectal cancer. *Cell* **184**, 4734–4752.e20 (2021).

17. Xing, X. *et al.* Decoding the multicellular ecosystem of lung adenocarcinoma manifested as pulmonary subsolid nodules by single-cell RNA sequencing. *Sci Adv* **7**, (2021).

18. Wu, S. Z. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 (2021).

19. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417–425 (2015).

20. Barkley, D. *et al.* Cancer cell states recur across tumor types and form specific interactions with the tumor microenvironment. *Nat. Genet.* **54**, 1192–1201 (2022).

21. Gröger, C. J., Grubinger, M., Waldhör, T., Vierlinger, K. & Mikulits, W. Meta-analysis of gene expression signatures defining the epithelial to mesenchymal transition during cancer progression. *PLoS One* **7**, e51136 (2012).

22. Tan, T. Z. *et al.* Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol. Med.* **6**, 1279–1293 (2014).

23. Mak, M. P. *et al.* A Patient-Derived, Pan-Cancer EMT Signature Identifies Global Molecular Alterations and Immune Target Enrichment Following Epithelial-to-Mesenchymal Transition. *Clin. Cancer Res.* **22**, 609–620 (2016).

24. Foroutan, M., Cursons, J., Hediyeh-Zadeh, S., Thompson, E. W. & Davis, M. J. A Transcriptional Program for Detecting TGFβ-Induced EMT in Cancer. *Mol. Cancer Res.* **15**, 619–631 (2017).

25. Hollern, D. P., Swiatnicki, M. R. & Andrechek, E. R. Histological subtypes of mouse mammary tumors reveal conserved relationships to human cancers. *PLoS Genet.* **14**, e1007135 (2018).

26. Epithelial-Mesenchymal Transition (EMT) markers. *Novus Biologicals* https://www.novusbio.com/antibody-news/antibodies/antibodies-for-epithelial-mesenchymal-transition-emt-marker (2016).

27. Han, C., Liu, T. & Yin, R. Biomarkers for cancer-associated fibroblasts. *Biomark Res* **8**, 64 (2020).

28. Nurmik, M., Ullmann, P., Rodriguez, F., Haan, S. & Letellier, E. In search of definitions: Cancer-associated fibroblasts and their markers. *Int. J. Cancer* **146**, 895–905 (2020).

29. PITX1 paired like homeodomain 1 [Homo sapiens (human)] - Gene - NCBI. https://www.ncbi.nlm.nih.gov/gene/5307.

30. GeneCards Human Gene Database. PITX1 Gene - GeneCards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=PITX1.

31. Song, X. *et al.* High PITX1 expression in lung adenocarcinoma patients is associated with DNA methylation and poor prognosis. *Pathol. Res. Pract.* **214**, 2046–2053 (2018).

32. Zhang, Y., Zhang, Z., Zhang, W., Hu, H. & Bao, G. Upregulated Transcription Factor PITX1 Predicts Poor Prognosis in Kidney Renal Clear Cell Carcinoma-Based Bioinformatic Analysis and Experimental Verification. *Dis. Markers* **2021**, 7694239 (2021).

33. Wang, Q., Zhao, S., Gan, L. & Zhuang, Z. Bioinformatics analysis of prognostic value of PITX1 gene in breast cancer. *Biosci. Rep.* **40**, (2020).

34. Li, Y. *et al.* Differential gene expression identifies a transcriptional regulatory network involving ER-alpha and PITX1 in invasive epithelial ovarian cancer. *BMC Cancer* **21**, 768 (2021).

35. Poos, A. M. *et al.* PITX1 Is a Regulator of TERT Expression in Prostate Cancer with Prognostic Power. *Cancers* **14**, (2022).

36. Ohira, T., Nakagawa, S., Takeshita, J., Aburatani, H. & Kugoh, H. PITX1 inhibits the growth and proliferation of melanoma cells through regulation of SOX family genes. *Sci. Rep.* **11**, 18405 (2021).

37. Otsubo, T. *et al.* DNA hypermethyation and silencing of PITX1 correlated with advanced stage and poor postoperative prognosis of esophageal squamous cell carcinoma. *Oncotarget* **8**, 84434–84448 (2017).

38. Zhang, Y. *et al.* PITX1 suppresses osteosarcoma metastasis through exosomal LINC00662-mediated M2 macrophage polarization. *Clin. Exp. Metastasis* **40**, 79–93 (2023).

39. Jin, Y. & Qin, X. Paired like homeodomain 1 and SAM and SH3 domain-containing 1 in the progression and prognosis of head and neck squamous cell carcinoma. *Int. J. Biochem. Cell Biol.* **127**, 105846 (2020).

40. Qiao, X. *et al.* Long noncoding RNA ABHD11-AS1 predicts the prognosis of pancreatic cancer patients and serves as a promoter by activating the PI3K-AKT pathway. *Eur. Rev. Med. Pharmacol. Sci.* **22**, 8630–8639 (2018).

41. Hu, T. & Lu, Y.-R. BCYRN1, a c-MYC-activated long non-coding RNA, regulates cell metastasis of non-small-cell lung cancer. *Cancer Cell Int.* **15**, 36 (2015).

42. Song, S. *et al.* LncRNA BCYRN1 promotes cell migration and invasion of non-small cell lung cancer via the miR-30b-3p/ROCK1 axis. *Neoplasma* **69**, 583–593 (2022).

43. Gu, L., Lu, L., Zhou, D. & Liu, Z. Long Noncoding RNA BCYRN1 Promotes the Proliferation of Colorectal Cancer Cells via Up-Regulating NPR3 Expression. *Cell. Physiol. Biochem.* **48**, 2337–2349 (2018).

44. Yu, J.-H. & Chen, Y. Clinical significance of lncRNA BCYRN1 in colorectal cancer and its role in cell metastasis. *Eur. Rev. Med. Pharmacol. Sci.* **23**, 9371–9378 (2019).

45. Tian, L., Chen, M., He, Q., Yan, Q. & Zhai, C. MicroRNA-199a-5p suppresses cell proliferation, migration and invasion by targeting ITGA3 in colorectal cancer. *Mol. Med. Rep.* **22**, 2307–2317 (2020).

46. Cheriyamundath, S. & Ben-Ze'ev, A. Wnt/β-Catenin Target Genes in Colon Cancer Metastasis: The Special Case of L1CAM. *Cancers* **12**, (2020).

47. Zheng, Y.-W. *et al.* FAM83A Promotes Lung Cancer Progression by Regulating the Wnt and Hippo

Signaling Pathways and Indicates Poor Prognosis. *Front. Oncol.* **10**, 180 (2020).

48. Lei, Y. *et al.* ERO1L promotes IL6/sIL6R signaling and regulates MUC16 expression to promote CA125 secretion and the metastasis of lung cancer cells. *Cell Death Dis.* **11**, 853 (2020).
49. Ji, H. *et al.* FAM83A promotes proliferation and metastasis via Wnt/β-catenin signaling in head neck squamous cell carcinoma. *J. Transl. Med.* **19**, 423 (2021).
50. Schinke, H. *et al.* A transcriptomic map of EGFR-induced epithelial-to-mesenchymal transition identifies prognostic and therapeutic targets for head and neck cancer. *Mol. Cancer* **21**, 178 (2022).
51. Zhang, H., Cui, X., Cao, A. ’na, Li, X. & Li, L. ITGA3 interacts with VASP to regulate stemness and epithelial-mesenchymal transition of breast cancer cells. *Gene* **734**, 144396 (2020).
52. Zhai, Y. *et al.* MUC16 affects the biological functions of ovarian cancer cells and induces an antitumor immune response by activating dendritic cells. *Ann Transl Med* **8**, 1494 (2020).
53. Li, Z. *et al.* Serum amyloid a, a potential biomarker both in serum and tissue, correlates with ovarian cancer progression. *J. Ovarian Res.* **13**, 67 (2020).
54. Chen, J., Gao, F. & Liu, N. L1CAM promotes epithelial to mesenchymal transition and formation of cancer initiating cells in human endometrial cancer. *Exp. Ther. Med.* **15**, 2792–2797 (2018).
55. Yates, J. *et al.* CanSig: Discovering de novo shared transcriptional programs in single cancer cells. *bioRxiv* 2022.04.14.488324 (2022) doi:10.1101/2022.04.14.488324.
56. Ahlmann-Eltze, C. & Huber, W. Comparison of transformations for single-cell RNA-seq data. *Nat. Methods* (2023) doi:10.1038/s41592-023-01814-1.
57. Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* 1–23 (2023) doi:10.1038/s41576-023-00586-w.
58. scanpy.pp.highly_variable_genes — Scanpy 1.9.1 documentation. https://scanpy.readthedocs.io/en/stable/generated/scanpy.pp.highly_variable_genes.html.
59. Pedregosa, Varoquaux & Gramfort. Scikit-learn: Machine learning in Python. *of machine Learning ….*

60. Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* **17**, 137–145 (2020).

61. Calculate module scores for feature expression programs in single cells. https://satijalab.org/seurat/reference/addmodulescore.

62. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

63. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

64. Charlier, F. *et al. trevismd/statannotations: v0.5*. (2022). doi:10.5281/zenodo.7213391.

65. [No title]. https://atlas.fredhutch.org/nygc/multimodal-pbmc/.

66. Reynolds, D. Gaussian Mixture Models. in *Encyclopedia of Biometrics* (eds. Li, S. Z. & Jain, A.) 659–663 (Springer US, 2009). doi:10.1007/978-0-387-73003-5_196.

67. Vasaikar, S. V. *et al.* EMTome: a resource for pan-cancer analysis of epithelial-mesenchymal transition genes and signatures. *Br. J. Cancer* **124**, 259–269 (2021).

68. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).

69. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400–416.e11 (2018).