CanSig: discovery of shared transcriptional states across cancer patients from single-cell RNA sequencing data

Josephine Yates[1,2,3†], Florian Barkmann[1†], Pawel Czyz[2,3,4†], Agnieszka Kraft[1,3], Marc Glettig[1,5], Frederieke Lohmann[1], Elia Saquand[1], Richard von der Horst[1], Nicolas Volken[1], Niko Beerenwinkel[3,4] and Valentina Boeva[1,2,3,6*]

1: Institute for Machine Learning, Department of Computer Science, ETH Zürich, Zurich Switzerland

2: ETH AI Center, ETH Zürich, Zurich Switzerland

3: Swiss Institute for Bioinformatics (SIB), Lausanne, Switzerland.

4: Department of Biosystems Science and Engineering, ETH Zürich, Switzerland

5: Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

6: Cochin Institute, Inserm U1016, CNRS UMR 8104, Paris Descartes University UMR-S1016, Paris 75014, France

†: these authors contributed equally

*Corresponding author:

Valentina Boeva

Professor of biomedical informatics

ETH Zurich, Department of Computer Science, Institute for Machine Learning, Universitätstrasse 6, 8092 Zurich, Switzerland

+41.77.525.73.05

valentina.boeva@inf.ethz.ch

Keywords: cancer; heterogeneity; single-cell RNA sequencing; transcriptional states; machine learning; transcriptomics;

Running title: Discovery of shared transcriptional states in cancer

Financial support: ETH AI Center fellowship program to P.C. Swiss National Science Foundation (SNSF) to F.B. and J.Y. (grant numbers 205321_207931 and CRSII5_209524).

Conflict of interest: No conflicts of interest to disclose.

Word count: 5221

Number of tables and/or figures: 4

Abstract

Multiple cancer types have been shown to exhibit heterogeneity in the transcriptional states of malignant cells across patients and within the same tumor. The intra-tumor transcriptional heterogeneity has been linked to resistance to therapy and cancer relapse, representing a significant obstacle to successful personalized cancer treatment. However, today there is no easy-to-use computational method to identify heterogeneous transcriptional cell states that are shared across patients from single-cell RNA sequencing (scRNA-seq) data.

To discover shared transcriptional states of cancer cells, we propose a novel computational tool called CanSig. CanSig automatically preprocesses, integrates, and analyzes cancer scRNA-seq data from multiple patients to provide novel signatures of shared transcriptional states and associates these states with known biological pathways. CanSig jointly analyzes cells from multiple cancer patients while correcting for batch effects and differences in gene expressions caused by genetic heterogeneity.

In our benchmarks, CanSig reliably re-discovers known transcriptional signatures on three previously published cancer scRNA-seq datasets, including four main cellular states of glioblastoma cells previously reported. We further illustrate CanSig's investigative potential by uncovering signatures of novel transcriptional states in four additional cancer datasets. Some of the novel signatures are linked to cell migration and proliferation and to specific genomic aberrations and are enriched in more advanced tumors.

In conclusion, CanSig detects transcriptional states that are common across different tumors. It facilitates the analysis and interpretation of scRNA-seq cancer data and efficiently identifies transcriptional signatures linked to known biological pathways. The CanSig method is available as a documented Python package at https://github.com/BoevaLab/CanSig.

Statement of significance

CanSig is an intuitive computational approach to detect shared transcriptional states across tumors and facilitate exploratory analysis of single-cell RNA sequencing data.

Introduction

For about a decade, single-cell RNA sequencing (scRNA-seq) has been a valuable tool to investigate the extensive inter- and intra-patient heterogeneity exhibited by malignant tissues (1,2). Although much research has focused on the heterogeneity of the tumor microenvironment, there has recently been increased interest in the intra- and inter-patient diversity of malignant cells, leading to a number of findings of gene signatures describing unique cell programs or states (3–10). Cell states, including rare ones such as cancer stem cells, are pivotal to the study of cancer as they can influence tumor maintenance, progression, and resistance to treatment (2,11).

In recent years, the discovery of gene signatures of shared transcriptional states in cancer has been often inconsistent across studies, resulting in signatures that were non-reproducible or non-comparable across datasets. Potential signature-discovery methods can be broadly categorized as early or late integration. Studies that used late integration identified signatures of differential states for individual patients and then grouped them into shared signatures (3,4,10). Studies that used early integration combined data from all patients before detecting shared signatures on the integrated set (12,13); this increased the statistical power to discover rare transcriptional states with low presence in individual tumors, highlighting the potential utility of early integration. Inspired by these works, we aimed to develop a fully automated intuitive toolset for early integration and discovery of shared transcriptional states in cancer.

Integrating heterogeneous populations of malignant cells poses specific challenges, including accounting for differences in tumor genetic backgrounds (2,5) and confounding factors such as batch effects (14), depth of sequencing (14), and cell cycle phase (5) that can obscure the biological signal of interest. While several methods exist to address some of these issues when integrating scRNA-seq data from non-cancerous cells (14–16), a fully automated computational technique specifically designed to address the challenges inherent to the across-patient integration of malignant cell datasets is yet to be proposed. Furthermore, current early integration methods do not directly account for all sources of unwanted variation (12,13,17–19) and do not correct for potential artifacts introduced by the integration step via finding overlapping signatures discovered across different ranges of hyperparameters.

We introduce here a computational approach implemented in an easy-to-use tool, CanSig, to discover *de novo* shared transcriptional states in cancerous cells. CanSig corrects for inter-patient differences driven by genomic copy number aberrations and batch effects while preserving the underlying biological signal. Gene expression-based clustering of cells and differential gene expression analysis between clusters are then applied to the latent space, and stable meta-signatures are identified. We validated CanSig on 12 simulated and three experimental datasets and showed that it successfully uncovers ground-truth transcriptional signatures. In addition, we used CanSig to identify and explore *de novo* transcriptional signatures and to reveal a link between genetic and transcriptional heterogeneity in seven experimental cancer datasets. Specifically, CanSig discovered a novel, presumably aggressive signature in esophageal squamous cell carcinoma, illustrating its full potential for exploring transcriptional states in cancer.

Materials and Methods

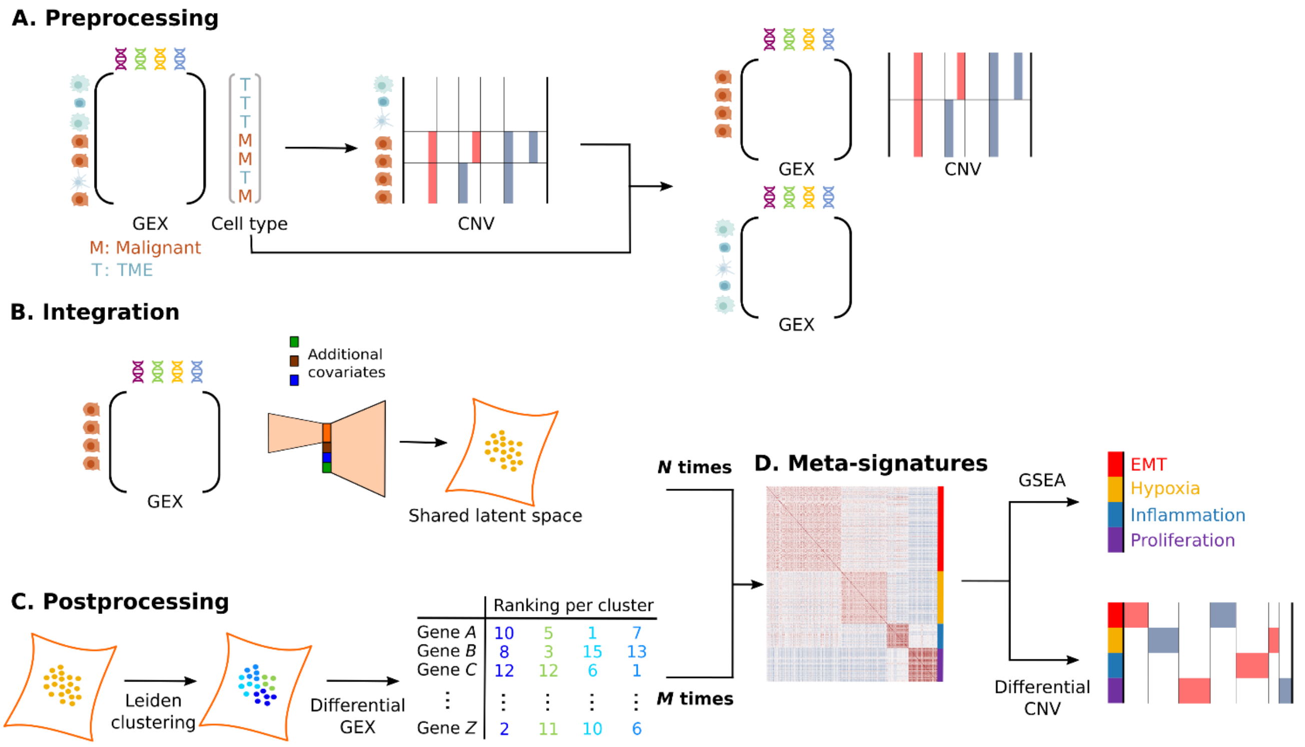
Transcriptional states versus transcriptional programs

We developed a computational approach, CanSig, for the *de novo* discovery of transcriptional states and corresponding gene signatures shared across cancer patients. Below, we define the terms of a transcriptional state and a transcriptional program to make their use disambiguous in the context of this study.

We define a cell program as a set of genes jointly up- or downregulated and linked to the activation of a molecular pathway. By definition, programs are not mutually exclusive: multiple pathways can be simultaneously activated within a single cell. By contrast, we define a cell state as a mutually exclusive combination of programs, meaning that a cell can only be in one state at a time. A state can thus be described by a set of genes jointly regulated in cells that have similar cellular behavior and function at the molecular level.

Pre-processing: removing misannotated cells through the inference of copy number variation

The first module in CanSig corresponds to data pre-processing (Fig. 1A). The output is a gene expression matrix for all high-quality cells with cell labeling as malignant and non-malignant. The preprocessing step involves identifying copy number variations (CNVs) in order to remove cells that are wrongly annotated as malignant (*i.e.*, they do not have CNVs) and cells that are wrongly annotated as non-malignant (*i.e.*, they do have CNVs), using CNVs inferred with *infercnv* (https://github.com/icbi-lab/infercnvpy) (Suppl. Methods).



	Ranking per cluster			
Gene *A*	10	5	1	7
Gene *B*	8	3	15	13
Gene *C*	12	12	6	1
⋮	⋮	⋮	⋮	⋮
Gene *Z*	2	11	10	6

Figure 1.

The CanSig modular approach consists of (A) pre-processing, (B) integration, (C) postprocessing, and (D) meta-signature discovery. **A,** Patient-level raw read count matrices with associated cell type annotations are provided as input. CanSig's pre-processing module infers the copy number variation (CNV) profile using *infercnvpy* and the cells are partitioned into malignant, non-malignant, and undetermined using both provided cell type annotations and the inferred CNV profiles. **B,** CanSig integration module consists of a conditional variational autoencoder (CVAE) that models the gene expression measurements of malignant cells conditioning on user-provided sources of unwanted variation such as batch ID, cell cycle score, and total read count and finds a latent space that preserves biologically meaningful signals. The latent space can be visualized using the Uniform Manifold Approximation and Projection (UMAP). **C,** The malignant cell raw count matrix and the latent space coordinates computed with the integration module are provided as input. CanSig's postprocessing module first performs clustering in the latent space using Leiden clustering. Then it conducts differential gene expression analysis for each cluster versus the rest of the cells. **D,** CanSig groups similar signatures across several integration and postprocessing runs to identify stable meta-signatures over multiple hyperparameter configurations. At this stage, outlier and patient-specific signatures are removed. CanSig then performs differential CNV analysis for each meta-signature versus the rest of the cells to uncover genetic variations linked to signatures. The meta-signature genes are used as input for the Gene Set Enrichment Analysis (GSEA) algorithm. GEX: gene expression; TME: Tumor microenvironment; CNV: copy number variation; GSEA: gene set enrichment analysis; EMT: epithelial-to-mesenchymal transition.

Integration: removing sources of unwanted variation

The second module of the CanSig is the dataset integration module (Fig. 1B). Its goal is to integrate cells into a biologically meaningful latent space that groups cells by their transcriptional state while removing sources of unwanted variation, such as batch effect, cell cycle, percentage counts in mitochondrial genes, differences in the genetic background, and differences in read counts.

In CanSig, we implemented the cell integration using a deep probabilistic generative model called scVI (14). Briefly, scVI models the observed gene expression in each malignant cell with a zero-inflated negative binomial (ZINB) distribution conditioned on the latent space encoding of the malignant cell and its covariates (*e.g.,* batch annotation) (Suppl. Methods).

Postprocessing: finding gene signatures describing potential shared states for each integration run

The third module of CanSig consists of clustering of cells in the latent space using Leiden clustering (20) followed by differential gene expression analysis (Fig. 1C). By default, the differential gene expression analysis includes computing the z-score for each gene for every cluster and ranking genes according to their z-scores. Statistical testing can be done with a t-test (default), a Wilcoxon rank-sum test, or a logistic regression. For every cluster determined by a set of hyperparameters (random seed, dimensions of the latent space, and the number of clusters), we obtain a ranked gene list referred to as the signature of the cluster.

Discovering meta-signatures: finding shared transcriptional states recurrently discovered across runs with different values of hyperparameters

The fourth and last module of CanSig consists of aggregating results from multiple runs of the integration and postprocessing module to produce stable, shared states (Fig. 1D). Indeed, states found independently of the choice of hyperparameters are more likely to represent a true biological signal rather than a data processing artifact stemming from a particular set of hyperparameters.

We characterize states using meta-signatures, defined as an aggregate of gene signatures discovered across runs with different hyperparameter values and random seeds. The number of states in the dataset is automatically inferred by iteratively clustering signatures until at least two meta-signatures are too correlated with one another in the discovery dataset (by default, with Pearson correlation coefficient between the two vectors of signature scores in cells over 0.55). At this stage, we eliminate meta-signatures that are specific to individual patients and meta-signatures that are composed of too few gene signatures or are mostly produced by a single run, as they are more likely to occur by chance rather than being indicative of a consistent transcriptional state (Suppl. Methods).

Characterizing meta-signatures through GSEA and differential CNV analysis

Once we have obtained meta-signatures associated with shared transcriptional states and assigned cells to each state, we characterize the shared transcriptional states according to the molecular pathways activated and the underlying genetic variation (CNV profiles).

To identify molecular pathways potentially activated in the states, CanSig provides the option to automatically perform gene set enrichment analysis (GSEA) using a pathway database chosen by the user. For each meta-signature, the associated ranked list of genes is used as input for the prerank algorithm of GSEA (21) applied to a molecular pathway database in the GMT format, *e.g.,* any MSigDB database (http://www.gsea-msigdb.org/gsea/msigdb/; default, Hallmarks of Cancer).

To help the user discover CNVs linked to the uncovered shared transcriptional states, CanSig performs differential CNV analysis (Suppl. Methods). This module is based on the CNV profiles calculated on scRNA-seq data by *infercnv* or provided by the user and is fully automated.

Experimental datasets

We used seven experimental scRNA-seq datasets for our analysis. Experimental datasets included (*i*) a high-grade glioma (HGG) dataset published by Yuan et al. (6), (*ii*) a glioblastoma dataset (GBM) published by Neftel et al. (7), (*iii*) a cutaneous squamous cell carcinoma (SCC) dataset published by Ji et al. (9), (*iv*) a colorectal cancer dataset (CRC) published by Pelka et al. (3), (*v*) a colorectal cancer dataset (CRC/iCMS) published by Joanito et al. (8), (*vi*) an esophageal squamous cell carcinoma (ESCC) dataset published by Zhang et al. (4), and (*vii*) a breast cancer (BRCA) dataset published by Wu et al. (10).

The known signatures for shared transcriptional states were extracted from the papers of origin when available, or from Neftel et al. (7) in the case of HGG. A large part of the analysis was conducted using the Scanpy package (22).

Simulated datasets

To investigate whether CanSig is capable of uncovering true biological signals present within the data while correcting for technical biases, we applied CanSig to twelve simulated datasets created with an adapted version of Splatter (23). Specifically, we simulated scRNA-seq datasets for malignant cells being in one of three shared transcriptional states; technical batch effects, patient-specific batch effects linked to copy number variation in malignant cells, inter-patient heterogeneity and library size effects were modeled with different intensities in the twelve datasets (Suppl. Methods). The python code used for the simulation is available at https://github.com/BoevaLab/SplatterSim.

Benchmarking integration methods on simulated datasets

To choose the best integration method for cancer single cells, we benchmarked six integration methods on our simulated datasets. Indeed, although integration method benchmarks already exist (24,25), none focused on malignant cells. We thus used the workflow described in scIB (24) to compare scVI (14), Scanorama (26), BBKNN (27), Harmony (15), Dhaka (28), and ComBat (16) in terms of their ability to integrate the data while preserving the biological signal of the shared transcriptional states.

The batch mixing across different methods was evaluated using k-nearest-neighbor batch estimation (kBET) (29). The preservation of biological variance in the latent space was quantified using three metrics: average silhouette width (ASW), adjusted rand index (ARI), and normalized mutual information (NMI) (*scikit-learn* implementations). ARI and NMI measurements were based on repetitive clustering of the latent space into two to five clusters and averaging the scores (Suppl. Methods).

We ran the CanSig tool with 4, 6, and 8 dimensions for the latent space in the integration module, and 6, 8, 10, and 12 clusters for clustering in the postprocessing module, using 2 random seeds per dimension and clustering, and the meta-signature discovery with a maximum correlation threshold of 0.3.

Applying CanSig to experimental datasets

To assess the performance of CanSig on real-world scRNA-seq data and to verify that CanSig can successfully rediscover previously reported states, we ran the tool on datasets from two cancer types with previously described gene signatures associated with shared transcriptional states (GBM, SCC) and on one dataset of the same cancer type (HGG) for which we used the previously known states. Additionally, we re-analyzed four datasets for which no shared states had yet been described (CRC, CRC/iCSM, ESCC, BRCA). Indeed, gene signatures previously reported for ESCC, CRC, and BRCA were not mutually exclusive and we argue they represent transcriptional programs rather than transcriptional states (see "Transcriptional states versus transcriptional programs"). For CRC/iCSM, the signatures reported in the original paper were patient-specific and thus not shared. We ran the CanSig tool with the same parameters as for the simulated data, except for the signature correlation threshold set to 0.55 (Suppl. Methods).

To assess if the uncovered meta-signatures representing shared transcriptional states corresponded to gene signatures of previously reported transcriptional states or how stable the uncovered states were under variation of the underlying data, we scored cells with the *scanpy* scoring function. For CanSig gene signature scoring we used the 50 top-ranked genes for each meta-signature. For the reported signatures, we used all genes from the lists provided by the authors. We then computed the Pearson correlation between the scores. We considered there was a correspondence between two states if the Pearson correlation between the signatures exceeded 0.65. To evaluate the stability of uncovered transcriptional states, we compared the meta-signatures reported by CanSig on datasets from similar cancer types; we used two colorectal cancer datasets CRC and CRC/iCMS, and two brain cancer datasets HGG/GBM. We compared the performance of CanSig to that of *scalop*, a late integration method used to discover shared cell states in GBM (7). Additionally, we separated the ESCC dataset into two parts by randomly splitting patient IDs and applied CanSig to the created patient subsets to quantify how similar the uncovered signatures in the two splits and the full dataset were.

To investigate the novel signatures uncovered in SCC, we additionally scored the meta-signatures on the spatial transcriptomics (ST) data provided in the original paper (patients 2, 4, 6, and 10) using the *scanpy* scoring function.

To investigate whether the states uncovered by CanSig were biologically meaningful, we correlated the meta-signatures with clinical characteristics, known subtypes, survival, and CNVs (for ESCC) in external validation cohorts from *The Cancer Genome Atlas* (TCGA) (Suppl. Methods).

All code to reproduce the analyses and figures presented in the manuscript can be found at https://github.com/BoevaLab/CanSig-Supplementary-Information.

Results

Benchmarking integration strategies on simulated data

Many approaches have been developed to integrate scRNA-seq data from non-cancerous tissues to determine groups of non-malignant cells with similar functions while removing batch effects and other experimental artifacts. To benchmark the efficacy of available data integration models in the cancer cell setting, we created several simulated scRNA-seq datasets mimicking data coming from a cancer patient cohort (Methods). We generated simulated single-cell RNA-sequencing datasets in which cells were assigned to one of three transcriptional states (Fig. 2A). To mimic real-world scenarios, we included various sources of technical and biological variability in the simulation such as batch effects, patient-specific effects linked to copy number variations in malignant cells, inter-patient heterogeneity, and library size effects (Suppl. Methods).

Using the twelve generated datasets, we evaluated the performance of six existing models for the correction of sample-specific effects and determined the optimal method to include in the CanSig integration module (Fig. 2B, Suppl. Table 1). Performance was measured along two axes: conservation of the biological signal of shared transcriptional state in the integrated data and batch integration. The conservation of the biological signal was assessed with the average of NMI, ARI, and ASW, measuring the consistently between cluster labels and cell state annotations; the success of batch integration was measured with the kBET index, consistent with the benchmarking protocol established in scIB (24) (Methods). In this experiment, we found that the scVI model (14) performed exceptionally well in maintaining biological signals when integrating different datasets (ranking first among all methods tested) and in the mixing of cells from different batches (ranking third) (Fig. 2B). Overall, scVI was able to balance the need for integrating patient data with preserving biological variance. Given the importance of preserving underlying biology for accurate marker gene identification and the fact that possible patient-specific signatures get filtered out at the meta-signature step, we have chosen scVI as the dataset integration method in CanSig.

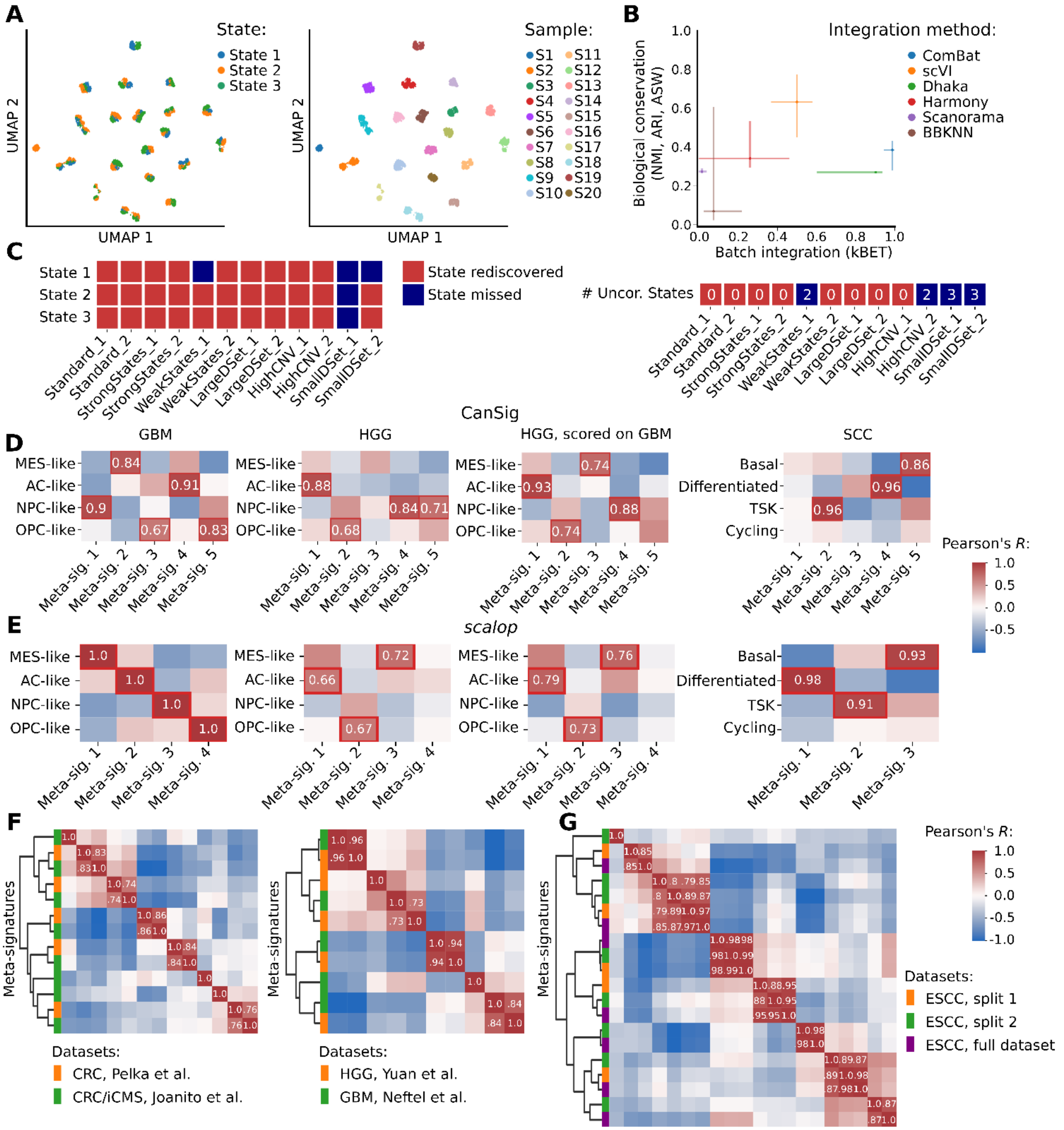


Figure 2.

CanSig stably rediscovers simulated and previously reported states in *in-silico* and experimental datasets. **A,** UMAP representation of one of the simulated datasets (“Standard_1”) in gene expression space, colored according to the three simulated transcriptional states (left) and the sample of origin (right). **B,** Benchmark of integration methods on simulated malignant data. The six integration methods, ComBat (16), scVI (14), Dhaka (28), Harmony (15), Scanorama (26), and BBKNN (27), were run on the twelve simulated datasets. The performance in terms of biological conservation (mean of ARI, NMI and ASW) and batch integration (kBET) is reported: median performance across sets and hyperparameters with the interquartile range [25-75]. **C,** Results of CanSig meta-signatures discovery on the simulated datasets. We score the uncovered meta-signatures and the true simulated states on the dataset and compute the Pearson correlation between cells. We consider a state to be rediscovered if Pearson correlation is above 0.65 (red square). States missed by CanSig are shown in blue. We report the number of uncovered meta-signatures that do not correlate with any of the simulated states (*i.e.,* false positives). **D**, **E,** Heatmap of Pearson correlation between previously reported signatures in GBM, HGG (using signatures reported in GBM), and SCC and meta-signatures uncovered by CanSig (D) and *scalop* (E). The meta-signatures and previously reported signatures are scored on the dataset of origin (with HGG additionally scored on GBM). Only correlation above 0.65 is annotated. The correspondence between meta-signatures and previously reported signatures is highlighted in red. The ground-truth GBM cell states were discovered with *scalop*, hence the perfect correlation is reported. **F,** Comparison of CanSig meta-signatures across two colorectal cancer datasets (CRC and CRC/iCMS) and two glioma datasets (HGG and GBM) and **G,** across random splits and the full dataset of ESCC. CanSig meta-signatures are scored on CRC/iCMS, GBM or ESCC respectively and the Pearson correlation is computed between the cell scores. Agglomerative clustering with average linkage was performed to group meta-signatures. Only correlation above 0.65 is annotated. ARI: Adjusted Rand Index; NMI: Normalized Mutual Information; ASW: Average Silhouette Width, kBET: k-nearest neighbor batch effect test; CRC: colorectal cancer from Pelka et al. (3); CRC/iCMS: colorectal cancer from Joanito et al. (8); HGG: high-grade glioma from Yuan et al. (6); GBM: glioblastoma from Neftel et al. (7); ESCC: esophageal squamous cell carcinoma from Zhang et al. (4); SCC: cutaneous squamous cell carcinoma from Ji et al. (9).

CanSig rediscovers known signatures in simulated and experimental datasets

We evaluated the ability of CanSig to rediscover ground-truth transcriptional states without predicting false states by comparing the gene signatures obtained from CanSig on twelve simulated datasets with the simulated ground truth. CanSig successfully rediscovered all three simulated states in nine out of the twelve datasets missing ground-truth states in about 14% of cases across datasets of different complexity (Fig. 2C). In eight out of twelve datasets, CanSig did not uncover any meta-signatures that did not correspond to a ground truth state. Performance of CanSig was found to be better on datasets containing at least 20 patients, *i.e.* on all datasets excluding “SmallDSet_1” and “SmallDSet_2”, with 10 patients only. There, CanSig rediscovered ground-truth states in 97% of cases (29/30) and reported only 4 unrelated signatures out of 33.

After demonstrating the ability of CanSig to rediscover simulated ground-truth states in *in-silico* data, we evaluated the ability of our approach to rediscover previously reported states using experimental datasets. Three experimental datasets with previously reported states were analyzed: high-grade glioma (HGG) (6), glioblastoma (GBM) (7), and cutaneous squamous cell carcinoma (SCC) (9). We found that CanSig successfully rediscovered all previously reported states in all datasets (Fig. 2D). Notably, the meta-signatures uncovered on the HGG dataset also recapitulated the previously reported states of GBM when scored on the GBM dataset, indicating the stability of these signatures across datasets. Specifically, meta-signature 5 identified in the HGG dataset appeared to reflect a transitioning state between NPC-like and OPC-like states, as evidenced by its correlation with these states. Furthermore, CanSig discovered two additional shared transcriptional states in the SCC dataset, discussed in more detail in the following sections.

Finally, we compared the performance of CanSig implementing early-stage integration to a late-stage integration method, *scalop*, used for cell state discovery in GBM (7) (Fig. 2E). We applied *scalop* to the two experimental datasets it had not previously been applied to, HGG and SCC, and assessed the correspondence of the uncovered states. In SCC, *scalop* rediscovered signatures of all previously reported transcriptional states. However, in HGG, *scalop* only rediscovered 3 of the 4 previously reported states, both when scored on HGG and on GBM, whereas the CanSig method was able to rediscover all four states.

CanSig discovers similar gene meta-signatures across similar datasets

To demonstrate the robustness of CanSig in uncovering meta-signatures across variations in the input data, we performed a comparison of CanSig results on pairs of independent datasets of colorectal cancer (3,8) and glioma (6,7), as well as on a random split on patients of a dataset of esophageal squamous cell carcinoma (ESCC) (4).

Our analysis of CanSig's meta-signatures uncovered from the colorectal cancer data, CRC (3) and CRC/iCMS (colorectal cancer dataset with intrinsic-consensus molecular subtypes) (8), revealed that CanSig was able to stably discover similar transcriptional states across the two datasets. Specifically, CanSig identified five states in the CRC dataset and eight states in the CRC/iCMS dataset; all five states discovered in the CRC dataset were also identified in the CRC/iCMS dataset with three CRC/iCMS states being specific to the second dataset (Fig. 2F). Similarly, in our analysis of the CanSig results of the glioma datasets, HGG (6) and GBM (7), CanSig identified four similar states and one state specific to each dataset. The alternative, late integration approach *scalop* did not perform significantly better than CanSig in this experiment (Suppl. Fig. 1).

Additionally, the application of CanSig to the ESCC dataset showed that our approach was able to stably discover similar signatures across different splits of a dataset on the patient level. In particular, CanSig uncovered eight states for the first subset of patients, five for the second subset, and seven for the full dataset. Of the seven states identified in the full dataset, four were also identified in both splits, while the remaining three states were specific to one split, possibly owing to differences in patient composition (Fig. 2G). These results supported the robustness of CanSig in uncovering similar sets of transcriptional states across variations in the input data.

CanSig discovers novel signatures of shared transcriptional states in experimental datasets linked with genetic variation and spatial organization in tumor tissues

We applied CanSig, using default parameter settings, to seven experimental datasets including breast cancer scRNA-seq data (BRCA) (10) and six datasets previously mentioned in the benchmarking setting (HGG, GBM, SCC, CRC, CRC/iCMS, ESCC).

CanSig identified between 5 and 8 shared transcriptional states per cancer type (Fig. 3A). To gain insight into the underlying molecular pathways and genetic variations associated with the states, we performed gene set enrichment analysis (GSEA) with the MSigDB gene database of cancer hallmarks (http://www.gsea-msigdb.org/gsea/msigdb/human/genesets.jsp?collection=H) and differential analysis for copy number variation (CNV) using corresponding modules implemented in CanSig.

GSEA revealed several molecular pathways recurrently activated in specific shared transcriptional states across multiple cancer types (Fig. 3B). Cell proliferation (E2F targets) and TNFA signaling were reported to be activated for at least one state in each cancer type. Other recurrent pathways, associated with uncovered cell states in six out of seven cancer types, included epithelial-to-mesenchymal transition (EMT), oxidative phosphorylation, and MYC targets v2 (also linked to cell proliferation).

To understand the possible biological meaning of two, relatively rare transcriptional states CanSig uncovered in SCC as *de novo* states (*i.e.*, they did not match the three previously reported states in this cancer type) we analyzed the GSEA results for the two corresponding gene signatures (meta-signatures 1 and 3, linked to 2% and 11% of cells annotated as malignant). Meta-signature 1 was associated with immune-reactive pathways (IL2 STAT5 signaling, allograft rejection) as well as TNFA signaling, suggesting that it may represent an immune-active state. Meta-signature 3 was associated with oxidative phosphorylation and MYC targets, which might suggest it characterizes a proliferative state. To further investigate these findings, we compared the spatial organization of the novel meta-signature 3 to that of meta-signature 4, which was strongly associated with the differentiated signature previously reported (9) and was closest to meta-signature 3 in terms of cell-to-cell distances in the latent space (Fig. 3A). Through scoring these meta-signatures in spatial transcriptomics slides of four

SCC patients (patients 2, 4, 6 and 10), we found that despite gene expression similarity (Pearson correlation: 0.31), they exhibited distinct spatial patterns: meta-signature 3 was more prevalent at the edges of tumors, consistent with its association with high proliferation (30) (Fig. 3C, Suppl. Fig. 2).

Differential CNV analysis identified numerous chromosomal regions with gains or losses significantly associated with uncovered states in all cancer types except for the high-grade glioma (Fig. 3D, Suppl. Fig. 3). A CNV of a region was considered to be associated with a state if there was a significant increase in the number of cells that had gained or lost the region compared to the rest of the states (least 25% difference and significant p-value). In our analysis, genomic gains were more frequently associated with specific transcriptional states than genomic losses. Our results corroborate previously reported observations that copy numbers of certain genes coding for transcription factors and signaling proteins may predispose to specific transcriptional states (7,10).

Discovered shared transcriptional states can be further investigated using external datasets to determine the biological and clinical relevance of the states. To demonstrate the potential for further investigation, we juxtaposed the gene signatures of transcriptional states identified by CanSig from scRNA-seq data with information from datasets of *The Cancer Genome Atlas* (TCGA). For each corresponding cancer type in TCGA, we scored the CanSig meta-signatures in all patients of the TCGA cohort and linked the patient scores with survival, known molecular subtypes, and clinical characteristics. We found five meta-signatures that were significantly associated with survival (Fig. 3E, Suppl. Table 2). The CanSig meta-signatures for shared transcriptional states were also found to be associated with known molecular subtypes. In HGG, GBM, and BRCA, all signatures were associated with a known mRNA subtype (Fig. 3E, Suppl. Table 3). In the remaining cancer types, at least half of the uncovered meta-signatures were associated with known subtypes. Finally, several meta-signatures were associated with clinical characteristics such as gender, clinical stage, and age (Fig. 3E, Suppl. Table 4). Overall, the majority of meta-signatures uncovered by CanSig were significantly associated with some of the reported patients' characteristics and/or survival in an external cohort, illustrating the biological relevance of the uncovered states.

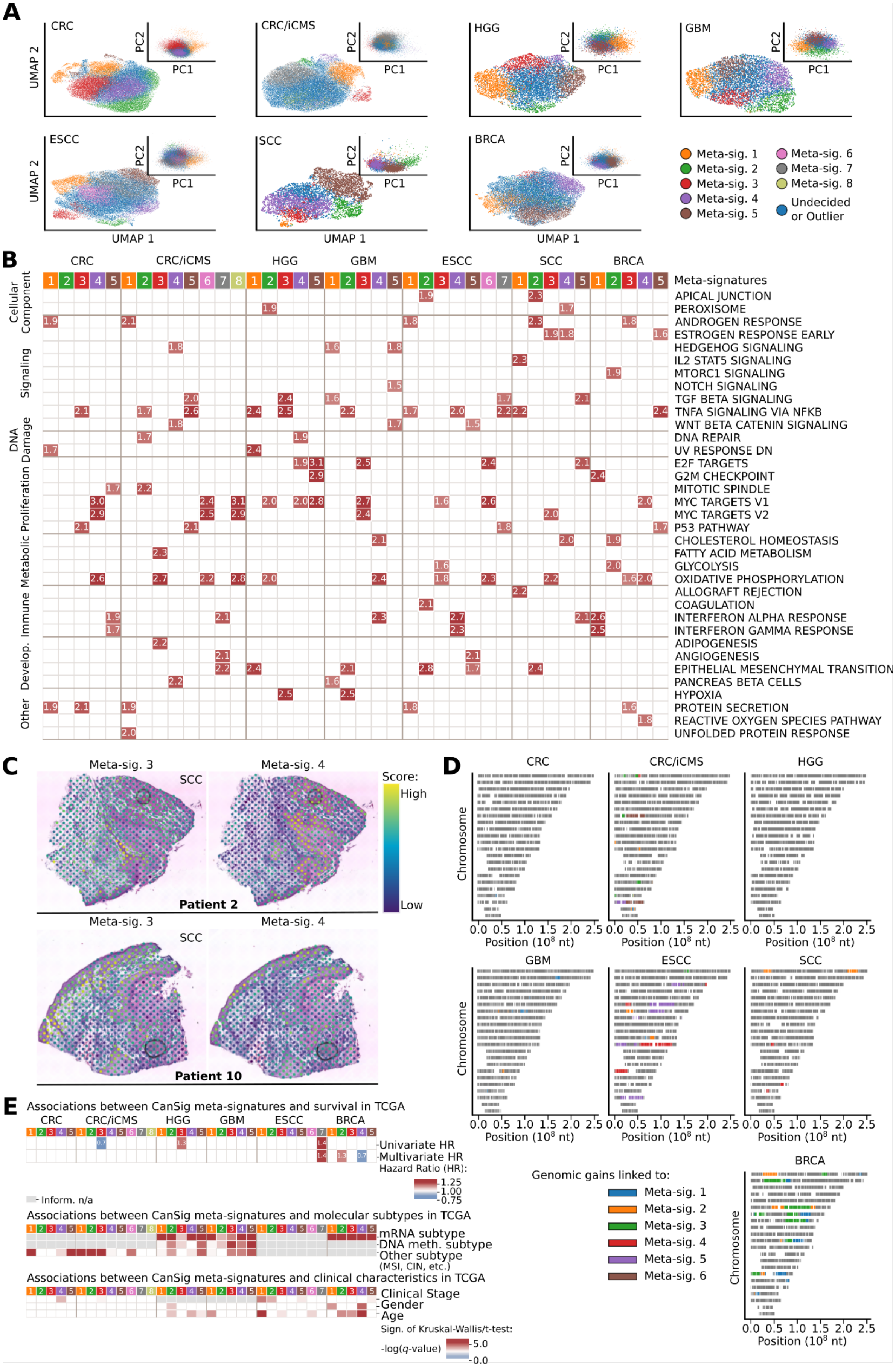

A
CRC
CRC/iCMS
HGG
GBM
ESCC
SCC
BRCA
UMAP 2
UMAP 1
PC2
PC1
Meta-sig. 1
Meta-sig. 2
Meta-sig. 3
Meta-sig. 4
Meta-sig. 5
Meta-sig. 6
Meta-sig. 7
Meta-sig. 8
Undecided or Outlier
B
CRC
CRC/iCMS
HGG
GBM
ESCC
SCC
BRCA
Meta-signatures
Cellular Component
Signaling
DNA Damage
Proliferation
Metabolic
Immune
Develop.
Other
APICAL JUNCTION 1.9 2.3
PEROXISOME 1.9 1.7
ANDROGEN RESPONSE 1.9 2.1 1.8 2.3 1.8
ESTROGEN RESPONSE EARLY 1.9 1.8 1.6
HEDGEHOG SIGNALING 1.8 1.6 1.8
IL2 STAT5 SIGNALING 2.3
MTORC1 SIGNALING 1.9
NOTCH SIGNALING 1.5
TGF BETA SIGNALING 2.0 2.4 1.6 1.7 2.1
TNFA SIGNALING VIA NFKB 2.1 1.7 2.6 2.4 2.5 2.2 1.7 2.0 2.2 2.2 2.4
WNT BETA CATENIN SIGNALING 1.8 1.7 1.5
DNA REPAIR 1.7 1.9
UV RESPONSE DN 1.7 2.4
E2F TARGETS 1.9 3.1 2.5 2.4 2.1
G2M CHECKPOINT 2.9 2.4
MITOTIC SPINDLE 1.7 2.2
MYC TARGETS V1 3.0 2.4 3.1 2.0 2.0 2.8 2.7 1.6 2.6 2.0
MYC TARGETS V2 2.9 2.5 2.9 2.4 2.0
P53 PATHWAY 2.1 2.1 1.8 1.7
CHOLESTEROL HOMEOSTASIS 2.1 2.0 1.9
FATTY ACID METABOLISM 2.3
GLYCOLYSIS 1.6 2.0
OXIDATIVE PHOSPHORYLATION 2.6 2.7 2.2 2.8 2.0 2.4 1.8 2.3 2.2 1.6 2.0
ALLOGRAFT REJECTION 2.2
COAGULATION 2.1
INTERFERON ALPHA RESPONSE 1.9 2.1 2.3 2.7 2.1 2.6
INTERFERON GAMMA RESPONSE 1.7 2.3 2.5
ADIPOGENESIS 2.2
ANGIOGENESIS 2.1 2.1
EPITHELIAL MESENCHYMAL TRANSITION 2.2 2.4 2.1 2.8 1.7 2.4
PANCREAS BETA CELLS 2.2 1.6
HYPOXIA 2.5 2.5
PROTEIN SECRETION 1.9 2.1 1.9 1.8 1.6
REACTIVE OXYGEN SPECIES PATHWAY 1.8
UNFOLDED PROTEIN RESPONSE 2.0
C
Meta-sig. 3
Meta-sig. 4
SCC
Patient 2
Patient 10
Score:
High
Low
D
CRC
CRC/iCMS
HGG
GBM
ESCC
SCC
BRCA
Chromosome
0.0 0.5 1.0 1.5 2.0 2.5
Position (10^8 nt)
Genomic gains linked to:
Meta-sig. 1
Meta-sig. 2
Meta-sig. 3
Meta-sig. 4
Meta-sig. 5
Meta-sig. 6
E
Associations between CanSig meta-signatures and survival in TCGA
CRC
CRC/iCMS
HGG
GBM
ESCC
BRCA
Univariate HR
Multivariate HR
Hazard Ratio (HR):
1.25
1.00
0.75
Inform. n/a
Associations between CanSig meta-signatures and molecular subtypes in TCGA
mRNA subtype
DNA meth. subtype
Other subtype (MSI, CIN, etc.)
Associations between CanSig meta-signatures and clinical characteristics in TCGA
Clinical Stage
Gender
Age
Sign. of Kruskal-Wallis/t-test:
-log(q-value)
5.0
0.0

Figure 3.

CanSig discovers novel states in experimental datasets linked with molecular pathways, genetic variation and clinical characteristics. **A,** UMAP and principal component analysis (PCA, inset) representations of selected latent spaces *(d=6)* for the seven experimental datasets on which CanSig was applied. Cells are colored according to the uncovered meta-signatures they are assigned to. Cells that are assigned to signatures that were cast as outliers or that cannot be confidently assigned to a meta-signature are annotated as undecided/outliers (blue). **B,** Heatmap representing the results from GSEA on the hallmarks of cancer for all seven experimental datasets. For each meta-signature for each cancer, the three most positively significantly enriched hallmarks (FWER p<0.05) are selected. Only significant associations with NES>1.5 are shown. **C**, Hematoxylin and eosin (H&E) staining of tissue sections and scoring of spatial transcriptomics (ST) spots. Spots are scored using the 50 top-ranked genes of the meta-signatures 3 and 4. **D**, Ideogram representation of copy number gains significantly associated with shared transcriptional states in all seven experimental datasets. Each row corresponds to a chromosome; the x-axis corresponds to the chromosomal position. Chromosomal regions inferred by *infercnvpy* are represented in grey. Regions associated with a meta-signature with FDR p<0.05 and with at least 25% more cells with a gain in the meta-signature are highlighted in the corresponding color. **E,** Associations between CanSig meta-signatures and the clinical characteristics in the corresponding cancer type in TCGA. The meta-signatures are scored on TCGA patients of each cohort by using the average RSEM value over the 50 top-ranked meta-signature genes. Cox's Proportional Hazard model is applied for survival prediction on each meta-signature score in a univariate analysis and then in a multivariate analysis correcting for age, stage, and tumor purity. Significant associations after FDR correction (FDR p<0.1) are represented. Available subtype information is retrieved and scores between every subtype are compared using the Kruksal-Wallis test. Associations might be positive or negative, *i.e.,* the signature might be significantly enriched or depleted in subtypes. The heatmap is colored according to -log(q-value). For clinical characteristics, categorical information is analyzed using the Kruskal-Wallis test; associations with age are analyzed using Pearson correlation. We grayed out boxes for which no information was available.

Meta-sig: meta-signature; GSEA: Gene Set Enrichment Analysis; FWER: family-wise error rate; NES: Normalized Enrichment Score; FDR: false discovery rate; TCGA: The Cancer Genome Atlas; RSEM: RNA-Seq by Expectation Maximization; CRC: colorectal cancer dataset from Pelka et al. (3); CRC/iCMS: colorectal cancer dataset from Joanito et al. (8); HGG: high-grade glioma cancer from Yuan et al. (6); GBM: glioblastoma cancer from Neftel et al. (7); ESCC: esophageal squamous cell carcinoma from Zhang et al. (4); SCC: cutaneous squamous cell carcinoma from Ji et al. (9); BRCA: breast cancer from Wu et al. (10).

CanSig discovers a signature of a novel transcriptional state linked to a higher clinical stage and EMT and WNT-pathway activation in esophageal squamous cell carcinoma (ESCC)

To demonstrate the potential of CanSig to uncover cancer cell transcriptional states that can lead to new biological insights, we performed a detailed investigation of one of the meta-signatures uncovered in the esophageal squamous cell carcinoma (ESCC) dataset. Specifically, we focused on meta-signature 5, which was significantly associated (NES ≥1.5, q-value < 0.05) with angiogenesis, epithelial-to-mesenchymal transition (EMT), and WNT beta-catenin signaling, as determined by GSEA

(Fig. 4A-B). We paid specific attention to this novel gene signature as WNT signaling has been shown to be a major promoter of cancer progression, stemness, and metastasis (31).

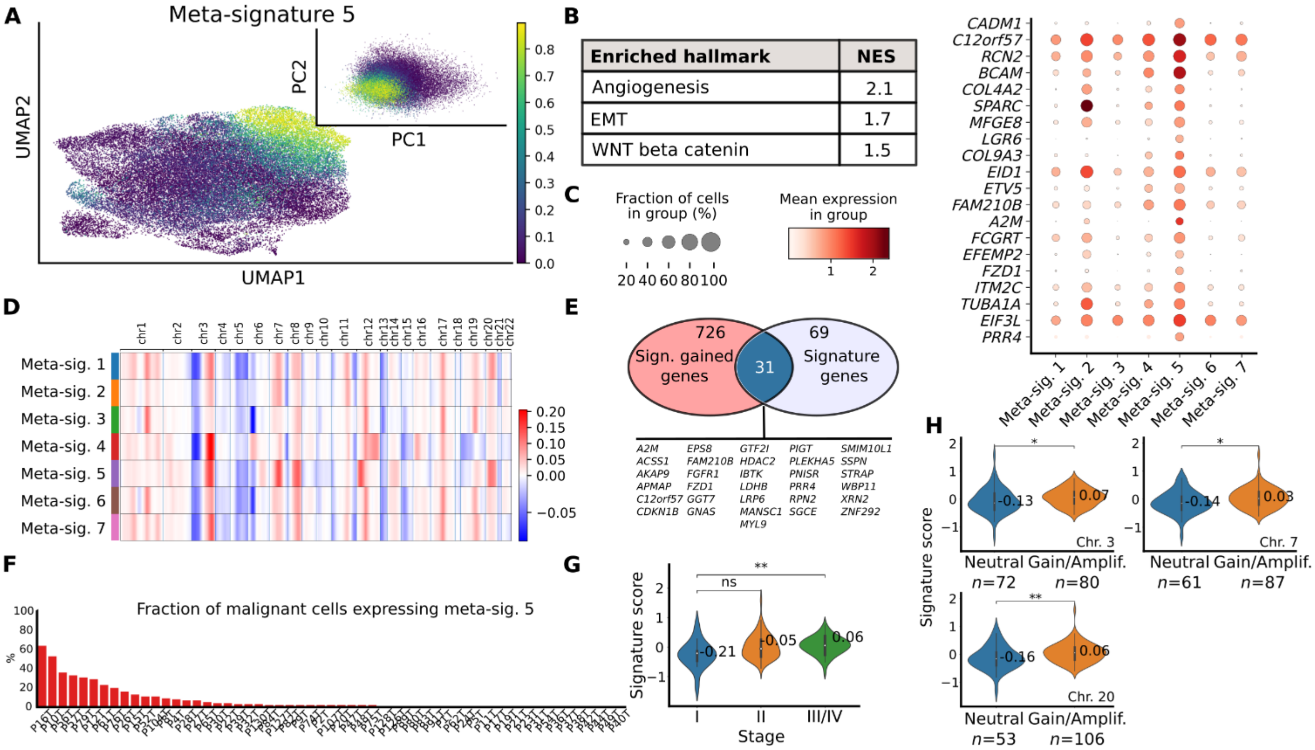
Further analysis of the top signature genes (20 most differentially expressed) revealed a high degree of specificity towards meta-signature 5, including signature-specific genes that had previously been linked to metastasis, stemness properties, and WNT beta-catenin signaling (32–38) (Fig. 4C). Notably, *LGR6* was found to be a marker of stem cells in skin squamous cell carcinoma in mice (37), further strengthening the possibility of stem-like properties of cells within this transcriptional state.

Meta-signature 5 also showed a strong association with genomic gains shared among patients, including significant gains on chromosomes 3, 6, 7, 8, 12, and 20 shared on average among 13 patients (Fig. 4D, Suppl. Table 5). To evaluate the relationship between the genomic gains and gene expression, we looked at the intersection of the 757 genes from the significantly differentially gained locations and the 100 top-ranked genes of meta-signature 5. We found 31 genes that were both gained and overexpressed in the state linked to meta-signature 5 (Fig. 4E), including the *FZD1* gene encoding a receptor of the beta-catenin signaling pathway. Malignant cells assigned to meta-signature 5 were found to be heterogeneously present across patients (Fig. 4F), with the fraction of such cells ranging from 0 to 63%, and only 6 patients out of 57 having more than 25% of their cells in this transcriptional state.

To further understand the clinical implications of this signature, we sought to correlate it with clinical characteristics in an external dataset from TCGA (ESCA). We found that the meta-signature 5 scores were significantly associated with later cancer stages (Fig. 3D, Fig. 4G), consistent with the increased aggressiveness and metastatic potential of cells with this signature according to the GSEA.

Finally, to validate the link between meta-signature 5 expression and genetic variation discovered by CanSig from the scRNA-seq data of ESCC, we evaluated whether the link between the significant genomic gains and signature expression was also present in the external dataset. Three main chromosomal regions were found to have significantly ($q<0.05$) higher meta-signature 5 scores in TCGA patients with a gain/amplification (Fig. 4H): chr3:122,680,839-187,745,725, chr7:74,199,652-107,931,730, and chr20:32,358,330-46,689,444.

To sum up, we identified a novel transcriptional state in ESCC characterized by over-expression of the EMT/WNT-pathway and linked to chromosomal gains in chromosomes 3, 7, and 20. The validity of the signature was confirmed in a TCGA dataset where this state was also associated with advanced tumor stages (III/IV), suggesting the uncovered gene signature could be used as a tool for stratifying patients and potentially applying more aggressive treatments at earlier stages of the disease.



Enriched hallmark	NES
Angiogenesis	2.1
EMT	1.7
WNT beta catenin	1.5

A2M
ACSS1
AKAP9
APMAP
C12orf57
CDKN1B
EPS8
FAM210B
FGFR1
FZD1
GGT7
GNAS
GTF2I
HDAC2
IBTK
LDHB
LRP6
MANSC1
MYL9
PIGT
PLEKHA5
PNISR
PRR4
RPN2
SGCE
SMIM10L1
SSPN
STRAP
WBP11
XRN2
ZNF292

Figure 4.

In ESCC, CanSig discovers a transcriptional state associated with EMT and WNT-pathway activation and linked to gains in chromosomes 3, 7, and 20. **A,** UMAP and PCA representation of the latent space of the ESCC dataset *(d=6)*, colored according to the probability of cell assignment to a transcriptional state linked to meta-signature 5. **B**, Hallmarks of cancer significantly enriched after FWER correction using GSEA for meta-signature 5. The NES associated with each pathway is indicated. **C,** Dotplot representation of the 20 top-ranked genes of meta-signature 5. The color corresponds to the mean expression and the size of the dot corresponds to the fraction of cells expressing the gene in the group of cells assigned to meta-signature 5 (*i.e.*, with at least one mapped read). **D,** Heatmap representation of the average value of the copy number profile of cells sorted by meta-signature they were assigned to. **E,** Representation of the intersection of genes with significant copy number gain associated with meta-signature 5 and signature genes (*i.e.,* 100 top-ranked genes in the meta-signature). **F,** Proportion of malignant cells of the ESCC patients assigned to meta-signature 5. **G,** Distribution of meta-signature 5 scores in the external cohort (TCGA-ESCA) associated with the clinical stage and **H,** with gains on regions of chromosomes 3, 7, and 20. For (G) and (H), the signature is computed as the average FPKM-UQ value of the signature genes for a patient. Significance is computed with the Wilcoxon rank-sum test (*: 0.01<p<0.05, **: 0.001<p<0.01, ns: non-significant).

Discussion

We have proposed CanSig, a modular method for the *de novo* discovery and analysis of shared transcriptional states in cancer. To the best of our knowledge, CanSig is the first automated cancer-specific approach that jointly analyzes scRNA-seq data from a patient cohort for the *de novo* discovery and annotation of shared transcriptional states in cancer. CanSig includes a pre-processing module that takes raw read counts from scRNA-seq on a patient cohort as input and infers the copy number variation (CNV) profiles of the cells further used to filter out potentially misannotated non-malignant cells. The integration module of CanSig then projects the data from malignant cells of all patients into a common latent space while controlling for sources of unwanted variation. The latent representation is used in the postprocessing step to identify potential transcriptional states through sequential clustering and differential gene expression analysis. The meta-signature discovery module of CanSig outputs consistently found states, and these states are then characterized with activated molecular pathways using GSEA and potential driver CNVs using differential CNV analysis. These characteristics set CanSig apart from previous early-integration approaches (12,13,17–19,28).

To ensure we used the best integration strategy in the cancer setting, we evaluated multiple integration techniques using simulated cancer datasets and selected the most effective approach. This turned out to be an autoencoder-based integration (14). Additionally, we compared the performance of CanSig on experimental datasets to that of *scalop (7)*, the only late integration method implemented within a reproducible computational framework, and found that CanSig performed comparably or superiorly depending on the dataset; in addition, CanSig has the advantage of automatically inferring the number of states from the correlation allowed between the states, while *scalop* requires the number of expected states to be provided in advance.

We would like to emphasize that several methods have been recently developed to discover *de novo* gene signatures from scRNA-seq data characterizing transcriptional programs, as opposed to states, in different types of tissues. These methods include a Bayesian factorization approach, scHPF (39), autoencoder-based approaches pmVAE (40) and scETM (41), the latter also implementing a topic model embedding, and a factor analysis-based method, f-scLVM (42). However, scHPF and pmVAE are not well-suited for a cancer-specific task involving data from multiple patients because they do not account for batch effects. Additionally, scETM and f-scLVM recommend to use such *a priori* knowledge as molecular pathway annotations; also, their linear models may limit the ability to correct for batch effects. Finally, the reference-query setting of ExpiMap (43), a method recently proposed to infer gene program activity, may not be suitable for identifying novel cancer signatures as cancer sets do not inherently include a perturbation or *a priori* known discriminative features between patients.

While CanSig has demonstrated promise in uncovering gene signatures of shared transcriptional states, there are certain limitations to keep in mind when using the method. First, while CanSig corrects for patient-specific effects and provides meta-signatures that are stably found across different sets of hyperparameters, one cannot completely rule out the presence of meta-signatures stemming from imperfect dataset integration or specific for only a small number of patients. To gain the first indication whether the discovered states are biologically meaningful, GSEA results are provided (21); however, it is highly recommended to further validate the states, *e.g.*, by using an external validation set to link the uncovered states to clinical characteristics, as was done in our analysis. Second, we chose to implement the early integration setting to gain statistical power to discover rare cell subtypes, but the shared latent space might only capture such rare cell states when there is a large number of patients presenting cells in these states or the rare cell state has a strong identity. Very rare or subtle cell states might not be captured. Third, currently, no single-cell-specific differential gene expression methods are implemented within CanSig, but the implemented methods have been reported to perform well on single-cell data (44).

The CanSig method could be further developed in various ways including incorporating other integration models. Another potential development could be to use CNV information as a proxy for patient identity, to better distinguish between technical artifacts and biological differences.

In conclusion, CanSig is a powerful tool that has many applications for the exploratory analysis of cancer cells. It can uncover transcriptional states of cancer cells with clinically relevant gene signatures and reveal potential links between CNVs and discovered cell states. As the field of cancer research continues to generate scRNA-seq data for human tumors, our approach provides an efficient way for researchers with limited computer science expertise to perform the analysis of their data while accounting for cancer-specific effects. The gene signatures of shared transcriptional states discovered by our approach can then be further investigated from a biological perspective by analyzing potential driver events and downstream consequences and conducting additional experimental work.

Data availability statement

The high-grade glioma data (HGG) was obtained from the Gene Expression Omnibus (GEO) repository GSE103224 at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103224. The glioblastoma data (GBM) was obtained on request on Broad Institute Single-Cell Portal: https://singlecell.broadinstitute.org/single_cell/study/SCP393/single-cell-rna-seq-of-adult-and-pediatric-glioblastoma

The cutaneous squamous cell carcinoma (SCC) single-cell and spatial transcriptomics data was obtained from the GEO repository GSE144240 at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE144240

The colorectal cancer data (CRC) was obtained from the GEO repository GSE178341 at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178341

The colorectal cancer data (CRC/iCMS) was obtained on request through Synapse under the accession code syn26844071 at https://www.synapse.org/#!Synapse:syn26844071/wiki/615389 The breast cancer data (BRCA) was obtained from the GEO repository GSE176078 at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE176078

The esophageal squamous cell carcinoma data (ESCC) was obtained from the GEO repository GSE160269 at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160269

For TCGA analysis, clinical and survival information was retrieved from Liu et al. (45). Subtype information was retrieved from https://www.synapse.org/#!Synapse:syn8402849 (a list of the papers describing the subtypes can be found at https://bioconductor.org/packages/release/bioc/vignettes/TCGAbiolinks/inst/doc/subtypes.html). We retrieved the RSEM (RNA-Seq by Expectation Maximization) gene expression pan-cancer file from the GDC (https://portal.gdc.cancer.gov/).

ESCA gene expression data (Fragments Per Kilobase of exon per Million mapped fragments normalized with Upper Quartiles, FPKM-UQ), CNV information, ESTIMATE (46) values of purity for each tumor sample, and clinical information were retrieved from the GDC (https://portal.gdc.cancer.gov/).

Code availability

CanSig is freely available as a documented Python package at https://github.com/BoevaLab/CanSig.

Acknowledgments

We would like to thank Gunnar Rätsch, Mitch Levesque, Stefan Stark, Antoine Combremont, Gian Hiltbrunner, and Laure Ciernik for engaging in helpful discussions. We are grateful to Julie Laffy for quick access to the glioblastoma dataset and for her generosity in sharing the code and tutorials for *scalop*.

Author contributions

V.B. conceived and directed the study. J.Y., F.B., and P.C. designed the model and implemented the computational framework. J.Y. and F.B. designed simulated data sets and analyzed simulated and experimental data. M.G., N.V. and R.H. participated in the initial model design and development. F.L. conducted preliminary studies on model benchmarking. E.S. contributed to the implementation of the computational framework. A.K. and J.Y conducted the analysis on TCGA. A.K. and N.B. contributed to model development and result interpretation. J.Y. wrote the manuscript with input from all authors.

Bibliography

1. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol Cell. 2017;65:631–43.e4.
2. Suvà ML, Tirosh I. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. Mol Cell. 2019;75:7–12.
3. Pelka K, Hofree M, Chen JH, Sarkizova S, Pirl JD, Jorgji V, et al. Spatially organized multicellular immune hubs in human colorectal cancer. Cell. 2021;184:4734–52.e20.
4. Zhang X, Peng L, Luo Y, Zhang S, Pu Y, Chen Y, et al. Dissecting esophageal squamous-cell carcinoma ecosystem by single-cell transcriptomic analysis. Nat Commun. 2021;12:5291.
5. Tirosh I, Izar B, Prakadan SM, Wadsworth MH 2nd, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016;352:189–96.
6. Yuan J, Levitin HM, Frattini V, Bush EC, Boyett DM, Samanamud J, et al. Single-cell transcriptome analysis of lineage diversity in high-grade glioma. Genome Med. 2018;10:57.
7. Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, et al. An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. Cell. 2019;178:835–49.e21.
8. Joanito I, Wirapati P, Zhao N, Nawaz Z, Yeo G, Lee F, et al. Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. Nat Genet. 2022;54:963–75.
9. Ji AL, Rubin AJ, Thrane K, Jiang S, Reynolds DL, Meyers RM, et al. Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma. Cell. 2020;182:497–514.e22.
10. Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A, et al. A single-cell and spatially resolved atlas of human breast cancers. Nat Genet. 2021;53:1334–47.
11. Rambow F, Rogiers A, Marin-Bejar O, Aibar S, Femel J, Dewaele M, et al. Toward Minimal Residual Disease-Directed Therapy in Melanoma. Cell. 2018;174:843–55.e19.
12. Liu Y, He S, Wang X-L, Peng W, Chen Q-Y, Chi D-M, et al. Tumour heterogeneity and intercellular networks of nasopharyngeal carcinoma at single cell resolution. Nat Commun. 2021;12:741.
13. Gouin KH 3rd, Ing N, Plummer JT, Rosser CJ, Ben Cheikh B, Oh C, et al. An N-Cadherin 2 expressing epithelial cell subpopulation predicts response to surgery, chemotherapy and immunotherapy in bladder cancer. Nat Commun. 2021;12:4906.
14. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15:1053–8.
15. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16:1289–96.
16. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform. 2020;2:lqaa078.
17. Dong X, Wang F, Liu C, Ling J, Jia X, Shen F, et al. Single-cell analysis reveals the intra-tumor heterogeneity and identifies MLXIPL as a biomarker in the cellular trajectory of hepatocellular carcinoma. Cell Death Discov. 2021;7:14.

18. Ho DW-H, Tsui Y-M, Chan L-K, Sze KM-F, Zhang X, Cheu JW-S, et al. Single-cell RNA sequencing shows the immunosuppressive landscape and tumor heterogeneity of HBV-associated hepatocellular carcinoma. Nat Commun. 2021;12:3684.
19. Song H, Weinstein HNW, Allegakoen P, Wadsworth MH 2nd, Xie J, Yang H, et al. Single-cell analysis of human primary prostate cancer reveals the heterogeneity of tumor-associated epithelial cell states. Nat Commun. 2022;13:141.
20. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9:5233.
21. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
22. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19:15.
23. Zappia L, Phipson B, Oshlack A. Splatter: simulation of single-cell RNA sequencing data. Genome Biol. 2017;18:174.
24. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19:41–50.
25. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020;21:12.
26. Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat Biotechnol. 2019;37:685–91.
27. Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park J-E. BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics. 2020;36:964–5.
28. Rashid S, Shah S, Bar-Joseph Z, Pandya R. Dhaka: Variational Autoencoder for Unmasking Tumor Heterogeneity from Single Cell Genomic Data. Bioinformatics [Internet]. 2019; Available from: http://dx.doi.org/10.1093/bioinformatics/btz095
29. Büttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing single-cell RNA-seq batch correction. Nat Methods. 2019;16:43–9.
30. Jiménez-Sánchez J, Bosque JJ, Jiménez Londoño GA, Molina-García D, Martínez Á, Pérez-Beteta J, et al. Evolutionary dynamics at the tumor edge reveal metabolic imaging biomarkers. Proc Natl Acad Sci U S A [Internet]. 2021;118. Available from: http://dx.doi.org/10.1073/pnas.2018110118
31. Zhan T, Rindtorff N, Boutros M. Wnt signaling in cancer. Oncogene. 2017;36:1461–73.
32. Feng Q, Li S, Ma H-M, Yang W-T, Zheng P-S. LGR6 activates the Wnt/β-catenin signaling pathway and forms a β-catenin/TCF7L2/LGR6 feedback loop in LGR6high cervical cancer stem cells. Oncogene. 2021;40:6103–14.
33. Xu S, Xu H, Wang W, Li S, Li H, Li T, et al. The role of collagen in cancer: from bench to bedside. J Transl Med. 2019;17:309.
34. Sun M-C, Fang K, Li Z-X, Chu Y, Xu A-P, Zhao Z-Y, et al. ETV5 overexpression promotes progression of esophageal squamous cell carcinoma by upregulating SKA1 and TRPV2. Int J Med Sci. 2022;19:1072–81.

35. Flahaut M, Meier R, Coulon A, Nardou KA, Niggli FK, Martinet D, et al. The Wnt receptor FZD1 mediates chemoresistance in neuroblastoma through activation of the Wnt/beta-catenin pathway. Oncogene. 2009;28:2245–56.

36. Takano A, Ishikawa N, Nishino R, Masuda K, Yasui W, Inai K, et al. Identification of nectin-4 oncoprotein as a diagnostic and therapeutic target for lung cancer. Cancer Res. 2009;69:6694–703.

37. Huang PY, Kandyba E, Jabouille A, Sjolund J, Kumar A, Halliwill K, et al. Lgr6 is a stem cell marker in mouse skin squamous cell carcinoma. Nat Genet. 2017;49:1624–32.

38. Zhang J, Cao H, Xie J, Fan C, Xie Y, He X, et al. The oncogene Etv5 promotes MET in somatic reprogramming and orchestrates epiblast/primitive endoderm specification during mESCs differentiation. Cell Death Dis. 2018;9:224.

39. Levitin HM, Yuan J, Cheng YL, Ruiz F Jr, Bush EC, Bruce JN, et al. De novo gene signature identification from single-cell RNA-seq with hierarchical Poisson factorization. Mol Syst Biol. 2019;15:e8557.

40. Gut G, Stark SG, Rätsch G, Davidson NR. pmVAE: Learning Interpretable Single-Cell Representations with Pathway Modules [Internet]. bioRxiv. 2021 [cited 2022 Mar 4]. page 2021.01.28.428664. Available from: https://www.biorxiv.org/content/10.1101/2021.01.28.428664v1

41. Zhao Y, Cai H, Zhang Z, Tang J, Li Y. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. Nat Commun. 2021;12:5261.

42. Buettner F, Pratanwanich N, McCarthy DJ, Marioni JC, Stegle O. f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. Genome Biol. 2017;18:212.

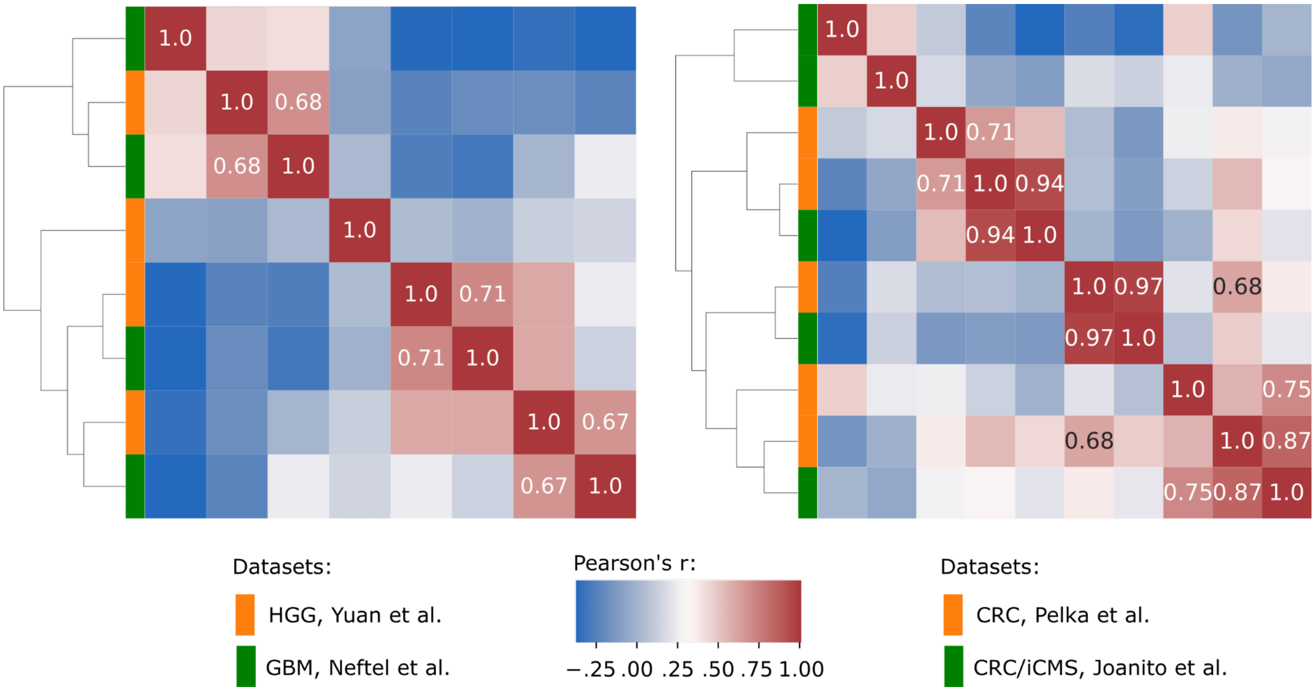
43. Lotfollahi M, Rybakov S, Hrovatin K, Hediyeh-zadeh S, Talavera-López C, Misharin AV, et al. Biologically informed deep learning to infer gene program activity in single cells [Internet]. bioRxiv. 2022 [cited 2022 Mar 4]. page 2022.02.05.479217. Available from: https://www.biorxiv.org/content/10.1101/2022.02.05.479217v2

44. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. 2018;15:255–61.

45. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell. 2018;173:400–16.e11.

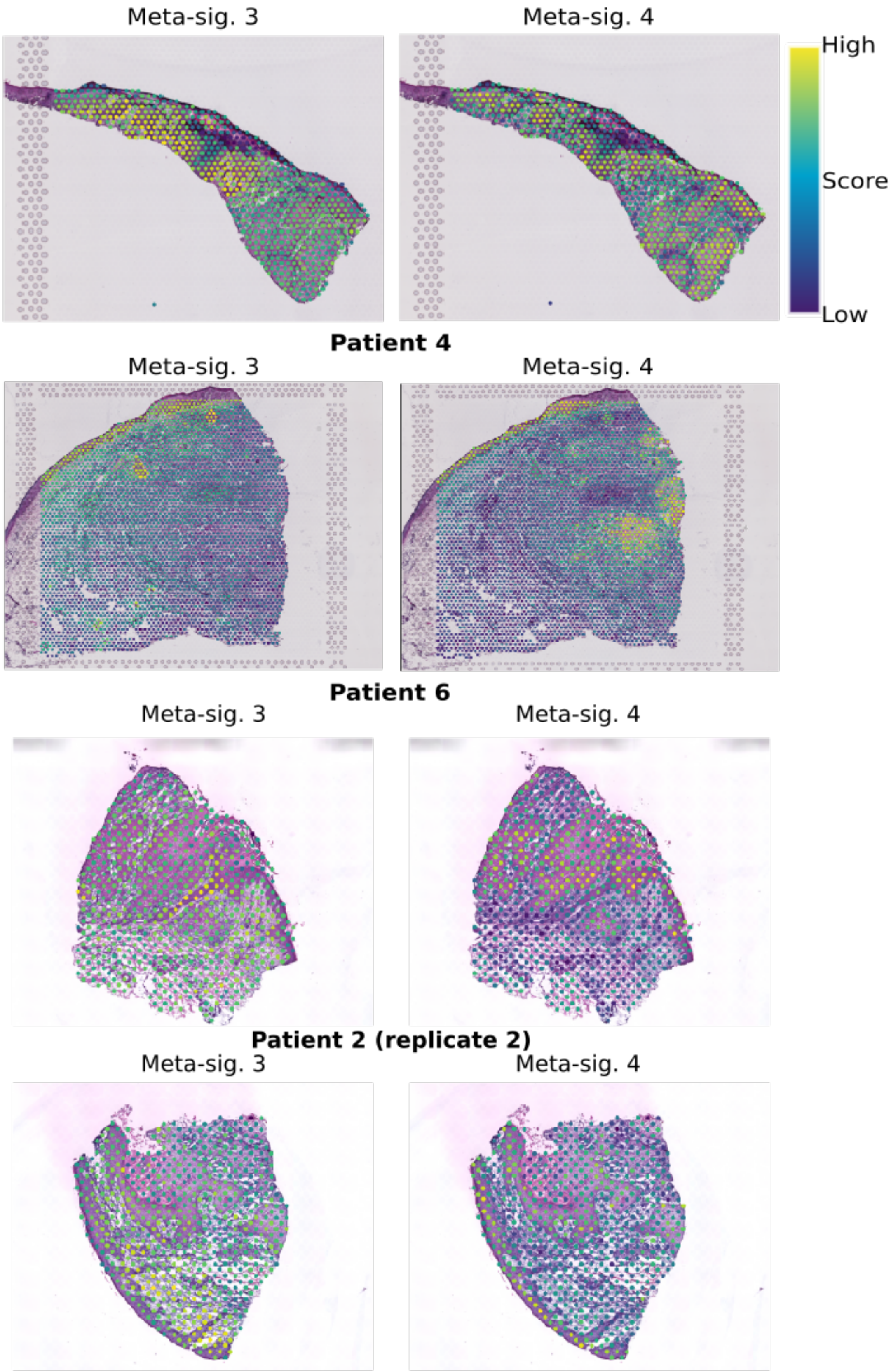
46. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.

Supplementary Figures and Tables



Suppl. Figure 1.

Comparison of *scalop* meta-signatures across two colorectal cancer datasets (CRC and CRC/iCMS) and two glioma datasets (HGG and GBM). *scalop* meta-signatures are scored on CRC/iCMS or GBM respectively and the Pearson correlation is computed between the cell scores. Agglomerative clustering with average linkage was performed to group meta-signatures. We only annotate correlations above 0.65. CRC: colorectal cancer from Pelka et al. (3); CRC/iCMS: colorectal cancer from Joanito et al. (8); HGG: high-grade glioma from Yuan et al. (6); GBM: glioblastoma from Neftel et al. (7).



Patient 4

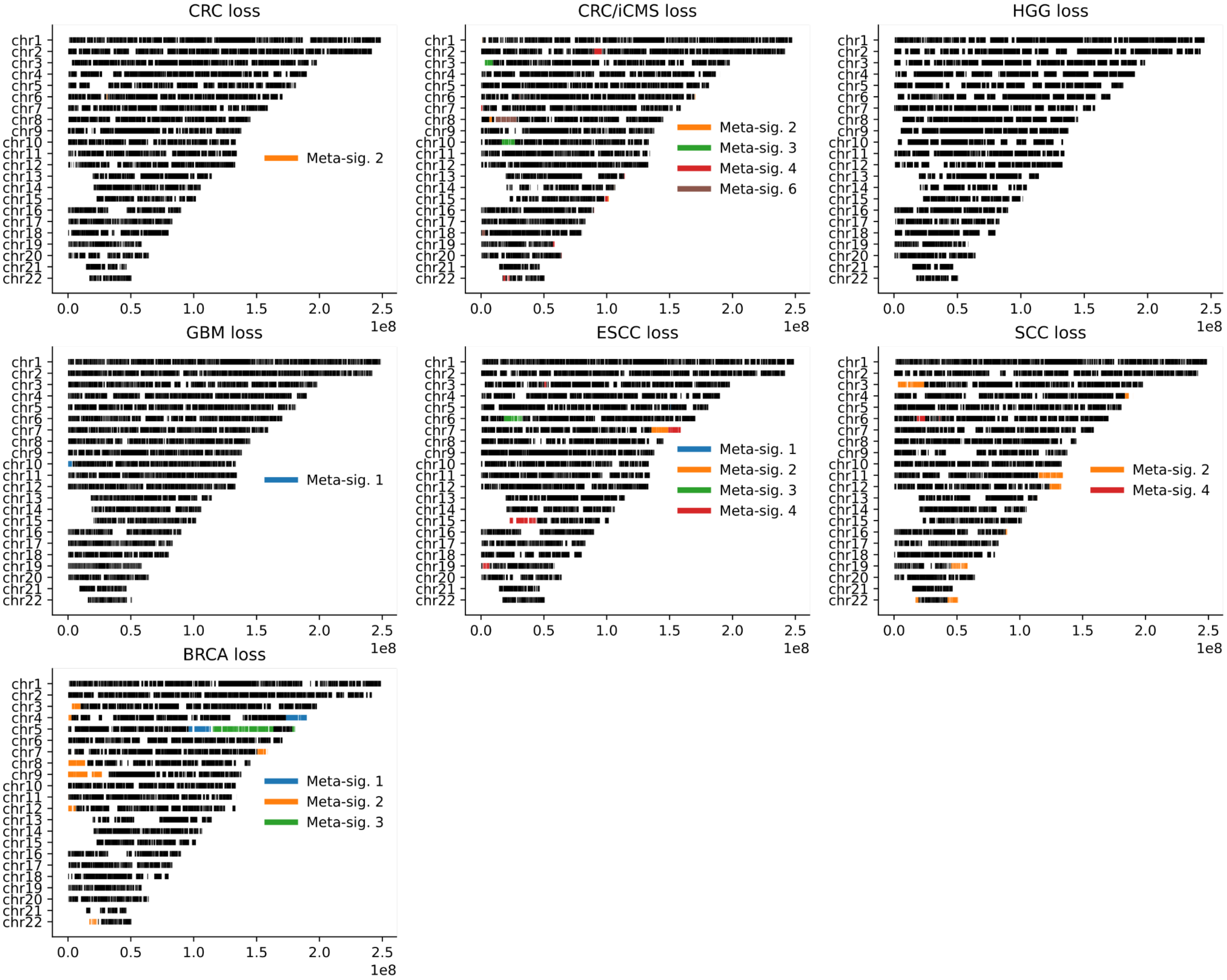
Patient 6

Patient 2 (replicate 2)

Patient 10 (replicate 2)

Suppl. Figure 2.

Hematoxylin and eosin (H&E) staining of tissue sections and scoring of spatial transcriptomics (ST) spots in SCC patients 2, 4, 6 and 10. Spots are scored using the 50 top-ranked genes of the meta-signatures 3 and 4.



Suppl. Figure 3.

Ideogram representation of significant copy number losses associated with meta-signatures in all seven experimental datasets. Each row corresponds to a chromosome; the x-axis corresponds to chromosomal position. Chromosomal regions inferred by *infercnvpy* are represented in black. Regions associated with a meta-signature with FDR p<0.05 and with at least 25% more cells with a loss in the meta-signature are highlighted in the corresponding color.

CRC: colorectal cancer, CRC/iCMS: colorectal cancer iCMS, HGG: high-grade glioma, GBM: glioblastoma, ESCC: esophageal squamous cell carcinoma, SCC: cutaneous squamous cell carcinoma; BRCA: breast cancer; Meta-sig.: meta-signature.

Supplementary Tables are provided in a separate .xlsx files.

Suppl. Table 1. Benchmark results for the 6 integration methods on the 12 simulated datasets. Mean, IQR and median are reported for all metrics.

Suppl. Table 2. Associations between CanSig meta-signatures and survival in TCGA. Univariate analysis corresponds to a Cox model with the score as predictor. Multivariate analysis corresponds to a Cox model with the score, the age, the stage and the tumor purity as predictors.

Suppl. Table 3. Associations between CanSig meta-signatures and known molecular subtypes in TCGA. The -log(q-value) significance of the FDR corrected p-value of the Kruskal-Wallis test across groups is reported.

Suppl. Table 4. Associations between CanSig meta-signatures and age, stage and gender in TCGA. For age, the Pearson correlation coefficient is reported, as well as the FDR corrected p-value. For age and gender, the Kruskal Wallis FDR-corrected p-value is reported. The mean signature score in each group is reported.

Suppl. Table 5. Gains significantly associated with meta-signature 5 in ESCC.