# CanSig: Discovering *de novo* shared transcriptional programs in single cancer cells

Josephine Yates[1,2,3]*, Florian Barkmann[1]*, Paweł Czyż[2,3,4]*, Marc Glettig[5], Frederieke Lohmann[1], Richard von der Horst[1], Elia Saquand[1], Nicolas Volken[1], Agnieszka Kraft[1,3], and Valentina Boeva[1,2,3,6]**.

*These authors contributed equally and can use any permutation of the names for public disclosure
**corresponding author
*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*

# Abstract

Human tumors are highly heterogeneous in their cell composition; specifically, they exhibit heterogeneity in transcriptional states of malignant cells, as has been recently discovered through single-cell RNA sequencing (scRNA-seq). Distinct states of malignant cells have been linked to variability in tumorigenic properties and resistance to anti-cancer treatment. Despite the fact that scRNA-seq data contain all necessary information to uncover shared transcriptional states of malignant cells in tumors, jointly analyzing cells from multiple cancer patients comes with its set of challenges including batch correction and accounting for patient-specific genetic background driving differences between gene expression vectors. We propose CanSig, an easy-to-use approach designed to discover known and *de novo* shared signatures in cancer single cells. CanSig preprocesses, integrates and analyzes scRNA-seq data to provide new signatures of shared transcriptional states and links these states to known pathways. We show that CanSig successfully rediscovers ground truth pathways determining shared transcriptional states in two simulated and three experimental datasets; the latter spanning 135 patients and 72,000 cells. We then illustrate CanSig's investigative potential by discovering novel signatures in esophageal squamous cell carcinoma possibly linked to targeted patient treatment; we also point out a *de novo* signature in breast cancer predictive of patients' survival. In the cancer types studied, we juxtapose copy number variation with discovered shared transcriptional states and uncover a genetic component predisposing cancer cells to activation of specific transcriptional programs. In sum, CanSig, specifically developed to analyze shared transcriptional heterogeneity of malignant cells of different genetic backgrounds, can greatly facilitate the exploratory analysis of scRNA-seq cancer data and efficiently identify novel transcriptional signatures linked to known biological pathways.

# Introduction

Single-cell RNA sequencing (scRNA-seq) has been steadily used to investigate heterogeneity in tissues for years now, with a plethora of sequencing techniques available[1]. scRNA sequencing has been particularly relevant in the study of cancer biology, as malignant tissues are characterized by extensive inter- and intra-patient heterogeneity[2]. Tumor tissues consist of a mix of malignant and non-malignant cells; the non-malignant cells referred to as the tumor microenvironment (TME) include a diverse set of immune cells, fibroblasts, endothelial cells etc. Although much research has focused on the TME heterogeneity, recently, waxing interest has been allocated to the diversity of cancer cells, leading to a number of findings of so-called "signatures" describing unique cell states[3–13]. Cell states are pivotal to the study of cancer as they can influence the maintenance and progression of tumors[2]. Moreover, it has recently been shown that cells in certain transcriptional states are more resilient to treatment and their presence can then cause a relapse at a later stage[14]. Therefore, identification of rare subtypes of cells (*e.g.,* cancer stem cells) can be critical to understand cancer biology as well as to develop personalized treatments. However, due to their rarity, cells in certain shared transcriptional states could be difficult to capture on a per-sample basis. It is thus essential to be able to uncover shared transcriptional states by leveraging the heterogeneity of several tumors at once.

The study of cancer cell heterogeneity poses specific challenges. First, while non-malignant cells generally cluster by cell type, malignant cells tend to cluster by patient[2], particularly because of differences in tumor genetic backgrounds[13]. In addition, as in healthy cells, other confounding factors such as batch effects[15], the number of total counts of the cell[15] or the cell cycle stage[13] can also drive the clustering of malignant cells and mask the biological signal of interest. Therefore, a meaningful integration of patient data removing sources of unwanted variation is critical to the successful pathway enrichment analysis[16] and subsequent discovery of shared states. To tackle the challenge of integrating scRNA sequencing from several patients, numerous methods that correct for some of these effects have been designed, including widely used scVI[15], Harmony[17] or Scanorama[18]. Additionally, several recent methods can be harnessed to discover *de novo* shared transcriptional signatures[19–23] yet most of these methods do not address the aforementioned cancer-inherent challenges. Therefore, a computational approach for cancer-specific signature discovery needs to be developed.

We propose here a computational approach implemented in an easy-to-use tool, CanSig, to discover *de novo* shared transcriptional programs in cancer cells. CanSig takes raw counts from single cell RNA sequencing of several patients as input, preprocesses the data through standard quality control and copy number variation (CNV) inference, integrates the patients while removing unwanted sources of variation, and performs gene set enrichment analysis on any database of choice to uncover potential new programs, while optionally running a stability analysis to select only consistently enriched pathways. We demonstrate that our method successfully rediscovers ground truth shared transcriptional signatures in two simulated and three experimental datasets. We then showcase how CanSig can be used to identify and explore *de novo* transcriptional signatures in cancer. First, we determine signatures in

esophageal squamous cell carcinoma potentially linked to targeted treatment; second, we uncover a *de novo* signature in breast cancer predictive of survival; and third, we link copy number gains and losses in breast cancer to discovered *de novo* signatures and histopathological subtypes highlighting a link between genetic and transcriptional heterogeneity.

# Materials and Methods

## From patient raw counts to enriched gene sets

CanSig consists of three major steps: pre-processing, data integration, and postprocessing (Fig. 1). CanSig takes as input a set of matrices of raw counts per patient and cell type annotation and outputs coordinates of single cancer cells in the shared latent space and a list of enriched pathways linked to subspaces in this integrated latent space. The method is highly modular: the user can therefore easily adjust the default parameters and methods to their specific purposes.

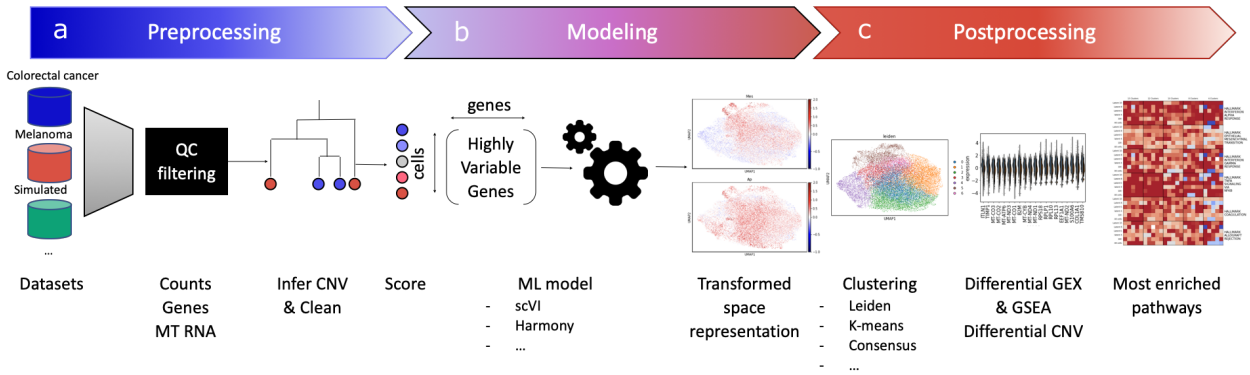


**Figure 1: CanSig pipeline consists of three modules (A) pre-processing, (B) integration and (C) postprocessing.** (A) Patient-level raw count matrices are provided as input. CanSig's pre-processing module first performs standard quality control (filter cells and genes using expression cut-offs and percentage of mitochondrial RNA). Then the CNV profile is inferred using *infercnvpy* and the cells are partitioned into malignant, non-malignant and undetermined using both provided cell type annotations and the CNV profile. Finally, the cells are scored using any previously known signatures. (B) The raw counts of all malignant cells after pre-processing is given as input. CanSig's integration module uses known models to find a latent space that preserves biologically meaningful signals while removing unwanted sources of variation. The module automatically formats the provided input so that it fits the constraints of the chosen model. The latent space can be visualized using the Uniform Manifold Approximation and Projection (UMAP). (C) The full dataset 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 an algorithm of choice. Then it performs differential gene expression analysis for each cluster versus the rest of the cells. Then the genes are ranked according to their z-score and the ranked list is used as input for the Gene Set Enrichment Analysis (GSEA) algorithm on a database of choice. For each cluster, the list of significantly associated pathways are ranked according to their Normalized Enrichment Score (NES). The cells can be optionally scored according to the set of significantly enriched pathways. Optional additional steps include performing a stability analysis to observe the consistency of the uncovered pathways through several runs of the pipeline, or performing differential Copy Number Variation (CNV) analysis to uncover potential links between signatures and CNVs.

## Pre-processing: Performing Quality Control and removing misannotated cells through Copy Number Variation inference

The first module of the pipeline corresponds to data pre-processing (Fig. 1A) to obtain the gene expression matrix of all cells, classified into malignant and non-malignant cells, as well as adding any associated annotations. This step can be skipped if the user already has a pre-processed raw count matrix for their data set.

The first step consists of standard quality control measures. We filter cells presenting less or more counts than a certain threshold as well as less than a certain number of unique genes to avoid empty cells, doublets or broken cells[24,25]. We also filter out cells that present more than a certain percentage of reads mapping to mitochondrially encoded DNA[25].

The second step consists in CNV inference to remove cells annotated as malignant that do not present CNVs and cells annotated as non-malignant that present CNVs.

To ensure we analyze only malignant cells and thus avoid contamination with healthy cells (*e.g.,* immune cells), we compute the copy-number variation (CNV) associated with each sample using the infercnv implementation in Python (https://github.com/icbi-lab/infercnvpy). Specifically, we run infercnv using the annotations (T cells, B cells, etc.) as reference for healthy and unhealthy cells while removing genes with a mean read count under 0.1. We then use Ward-linkage hierarchical clustering on the CNV profile to obtain a dendrogram of the cells. By definition, a dendrogram node only contains two child nodes that represent clusters of clusters, except leaf nodes that represent cells. We then iteratively assign a CNV status to the nodes according to the composition of their subtress. In particular, if both the subtrees of a node present at least 60% of non-malignant cells, we annotate it as non-malignant. We traverse the tree until all nodes and/or a maximum depth of five are reached, allowing us to obtain an annotation as “no CNV” or “CNV” for every cell. We finally remove cells originally annotated as healthy that clustered with unhealthy cells (thus presenting CNVs), or cells originally annotated as unhealthy that clustered with healthy cells.

Additionally, CanSig can then optionally score cells according to a predetermined set of upregulated genes, corresponding to a previously uncovered signature. To do so, the cell scoring is performed using the *score_genes* function of the Scanpy package[27], originally proposed by Tirosh et al.[13]

Finally, all filtered count matrices are concatenated into three main dataset count matrices: one containing only the malignant cells, one the non-malignant cells, one the undetermined cells (corresponding to cells filtered out at the infercnv step).

## Integration: removing sources of unwanted variation

The second module of the pipeline corresponds to projection of the dataset into a meaningful latent space (Fig. 1B). The models currently implemented within the CanSig pipeline are Harmony[17] and scVI[15].

CanSig automatically processes the full count matrix to fit the requirements of the selected model. For scVI, the most highly variable genes are selected and raw counts are inputted. The batch effect and number of counts are automatically corrected for using scVI[15], however additional confounding factors can be accounted for at this stage. For Harmony, the data is normalized using counts per ten thousand (CPT) and log-transformation as follows: for a cell $c$ and the set of genes $g_n$ with the associated counts in the cell $n_k^c$ we perform the following transformation

$$CPT_c(g_k) = \frac{n_k^c}{\sum_l n_l^c} \times 10^4$$

We then log-transformed our counts by performing the following transformation $X_k^c = log(1 + CPT_c(g_k))$. We then apply Principal Component Analysis (PCA) and use the $n$ first components as input for Harmony. Harmony corrects for batch effect but does not allow for correction of other confounding factors.

The output of this module is the coordinates of all the cells of the dataset in a meaningful latent space. The latent space can be visualized using the Uniform Manifold Approximation and Projection (UMAP)[28] applied on the latent space coordinates.

## Postprocessing: rediscovering expected signatures and their link to new transcriptional signatures and copy number variations

The third and final module of CanSig consists in clustering of cells in the latent space followed by differential gene expression analysis and gene set enrichment analysis as well as differential copy number variation analysis (Fig. 1C).

The first part consists in clustering, which can be done using several algorithms currently implemented with CanSig: Leiden clustering[29], K-Means[30], GMM[31], spectral clustering[32], and HDBSCAN[33]. As per the integration module, CanSig automatically processes the data so that it corresponds to the expected input, notably by computing the k-nearest neighbor graph for Leiden clustering.

The second part is differential gene expression analysis, that computes the z-score for each gene for every cluster. The z-score is computed as follows:

$$Z_k^l = \frac{\mu_k^l - \mu_k^{\setminus l}}{\sqrt{\frac{(\sigma_k^l)^2}{n_l} + \frac{(\sigma_k^{\setminus l})^2}{n_{\setminus l}}}}$$

where $\mu_k^l$ and $\sigma_k^l$ are the mean and the standard deviation of the gene $k$ in cluster $l$, $\mu_k^{\setminus l}$ and $\sigma_k^{\setminus l}$ the mean and standard deviation of the gene $k$ in all the others cells, $n_l$ the size of cluster $l$, $n_{\setminus l}$

the number of cells not in cluster $l$. Differential gene expression can also be run with a Wilcoxon rank-sum test rather than a t-test or a logistic regression as described by Ntranos et al.[34].

The third part is gene set enrichment analysis. For each cluster, the list of genes ranked according to their z-score is used as input for the Prerank algorithm of GSEA[16], run on any valid file in the GMT format, *e.g.,* any MSigDB database (http://www.gsea-msigdb.org/gsea/msigdb/), including the hallmarks of cancer. CanSig then selects the significantly enriched gene sets after False Discovery Rate (FDR) correction[16] for each cluster that had a positive enrichment score (thus positively enriched in the cluster of interest).

An optional part consists in rescoring each discovered pathway in the cells on a sample-level using the scanpy scoring function with the list of genes of the pathway.

To link potential Copy Number Variations (CNVs) with parts of the latent space, another optional part consists in performing differential CNV analysis – indeed, much like differential gene expression, CNVs can be differentially present in a cluster versus the rest of the latent space. We can use the CNV profile per cell inferred by *infercnvpy* in the pre-processing step or any user-provided CNV profile. We give the option to compute a smoothed version of the infercnv profile by applying the following dynamic thresholding transformation for each gene (https://github.com/broadinstitute/inferCNV/wiki/De-noising-Filters):

$$CNV_{g,i} = \{1\ if\ CNV_{g,i} > \mu_g + 1.5\sigma_g\ else\ -1\ if\ \ CNV_{g,i} < \mu_g - 1.5\sigma_g\ else\ 0\}$$

where $\mu_g$, $\sigma_g$ are the CNV mean and standard deviation associated with a gene $g$. We thus obtain a profile where potential gains are indicated by 1, potential losses by -1 and no significant CNV by 0.

We then use the Wilcoxon rank sum test to compare CNV expression between clusters, either with the raw or the smoothed CNV profile. We finally apply FDR correction to the obtained *p*-values and rank the genes according to the difference in mean CNV expression between the cluster and the rest.

The second and third modules are jointly implemented so that the user can obtain the results from integration and postprocessing in a single run.

## Stability analysis: finding consistently enriched pathways in the dataset

The final step in CanSig is a stability analysis of the uncovered pathways. The underlying assumption is that if the pathways uncovered are consistently enriched independently of the choice of hyperparameters, these pathways are more likely to represent the true biology of the subset of cells, rather than a potential processing artifact. Thus, to assess the stability of the uncovered gene sets, CanSig gives the opportunity to run the pipeline with several hyperparameters to observe which gene sets systematically are significantly enriched. The normalized enrichment score (NES)[16] of the pathways can then be visualized as a heatmap (Fig. 1C).

## Datasets

We used two simulated and three real-world single cell RNA sequencing datasets for our analysis. The simulated data was created using the Splatter package[35] (details in Appendix A). The first dataset is an esophageal squamous cell carcinoma (ESCC) dataset published by Zhang et al[7]. The second is a colorectal carcinoma (CRC) dataset published by Pelka et al[5]. The third is a breast cancer dataset published by Wu et al.[36]. Known signatures were extracted from the papers of origin. Most of the analysis was conducted using the Scanpy[27] package.

## Applying CanSig to simulated datasets

To investigate whether CanSig is capable of uncovering true biological signals present within the data while correcting for technical bias, we applied CanSig to two simulated datasets created with Splatter[35]. We sought to simulate a dataset similar to the real-world data encountered (see Appendix A). Notably, we simulated a dataset (simulated dataset A) with eight different uneven batches and three programs expressed in each batch. The programs corresponded to a gradient rather than to a discrete identity. We also used the same dataset to generate a new one (simulated dataset B) by removing two programs for three different patients, effectively simulating inter-patient heterogeneity in the dataset. We applied CanSig, without the pre-processing module, to both these datasets. We created an artificial database that consisted of three subsets of the most highly differentially expressed genes of the three programs present in the data and three random control programs to use for input in the GSEA algorithm. We performed the stability analysis and assessed via a heatmap which of the programs were consistently discovered.

## Applying CanSig to real-world datasets

To assess the performance of CanSig on real-world scRNA sequencing data and to verify that CanSig can successfully rediscover known signatures, we ran the pipeline and performed a stability analysis on a set of three cancer types with known signatures. We filtered cells presenting less than 1,500 total counts (potentially empty or broken cells) and more than 50,000 counts (to exclude doublets). We also filtered cells whose transcripts mapped to less than 700 genes. The mitochondrial content threshold was dependent on the dataset: for breast and ESCC the cells with a mitochondrial content over 20% had been prefiltered. For CRC, the threshold was set at 30%.

We then applied scVI to all three datasets, using between 4 and 10 dimensions in the latent space. Additionally to the batch and count correction, we used the percentage of mitochondrial counts as a confounding factor. We ran the integration module both by conditioning on the cell cycle (referred to as “all cells”) or by removing the cells in G2 or S phase according to the function *score_genes_cell_cycle* by scanpy, and we performed Leiden clustering using between 6 and 15 clusters, followed by differential gene expression analysis with a *t*-test or with a Wilcoxon rank-sum test and computed GSEA on the set of cancer hallmarks which we retrieved from MSigDB (http://www.gsea-msigdb.org/gsea/msigdb/genesets.jsp?collection=H).

## Use case: Finding new signatures in Esophageal Squamous Cell Carcinoma using WikiPathways

To show how one can use CanSig to uncover novel associations with known signatures, we applied CanSig to the ESCC dataset[7] using WikiPathways[37] as an exploratory database in the GSEA step. We investigated the significantly enriched pathways with Leiden clustering using $k = 10$ clusters to partition the latent space obtained with scVI with latent space dimensionality $d = 6$.

## Use case: Investigating Copy Number Variations associated with the basal and luminal subtypes in breast cancer

To show how one can use CanSig to understand the underlying biology of a signature, we investigated the link between the CNV profile and the known histopathological signatures in the breast cancer dataset[36]. We applied differential CNV analysis with Leiden clustering using $k = 10$ clusters to partition the latent space obtained with scVI with latent space dimensionality $d = 6$.

## Use case: Discovering a *de novo* signature predictive of survival in breast cancer

To show how one can use CanSig to find *de novo* signatures, we applied CanSig to the breast cancer dataset[36] and investigated the signatures associated with the different clusters using Leiden with $k = 10$ clusters to partition the latent space obtained with scVI with latent space dimensionality $d = 6$. A signature was defined as the genes with the 100 highest z-scores associated with a cluster. To link the signatures with survival, we used as validation cohort The Cancer Genome Atlas (TCGA) breast cancer dataset (BRCA), using Fragments per Million Kilobases Upper Quartile (FPKM-UQ) gene expression and clinical information retrieved from the GDC (https://portal.gdc.cancer.gov/), as well as Consensus Purity Estimates[38] for purity values for each tumor, a metric that aggregates results from several purity estimation methods. The samples of TCGA were scored according to the signature by using the mean of the standardized FPKM-UQ values of the genes in the signature. We then computed a Cox regression using the min-max normalization of the signature score, the standardized age, the CPE purity value and the stage as covariates.

# Results

## CanSig allows for cancer-specific *de novo* discovery of transcriptional signatures

CanSig is a method that allows for *de novo* discovery of shared transcriptional signatures in single-cell RNA sequencing (scRNA-seq) data. This method takes patient-level raw counts as input and outputs significantly enriched pathways associated with a subset of cells. CanSig consists of three main modules: pre-processing, integration, and postprocessing (Fig. 1).

The first module, pre-processing, takes as input patient-level raw counts with which we first perform standard quality control (QC)[24]. We then infer the copy number variation (CNV) profile of the cells and use it jointly with provided cell type annotation to partition the data into malignant, non-malignant and cells for which we are uncertain of the origin (termed undetermined). The malignant cells are used for the rest of the analysis; these cells can be scored according to user-provided known signatures. The second module, integration, finds a meaningful latent space representation of the data that retains biological signals while removing sources of unwanted variation such as batch or cell cycle for instance. The third module, postprocessing, partitions the latent space using a user-defined clustering algorithm, performs differential gene expression analysis and subsequent gene set enrichment analysis to uncover pathways significantly enriched in parts of the latent space. Some optional steps can be used to investigate the latent space, such as differential CNV analysis or performing a stability analysis to evaluate the consistency of the uncovered pathways. CanSig is implemented as a documented Python package at https://github.com/BoevaLab/CanSig.

## CanSig successfully rediscovers known pathways in simulated data

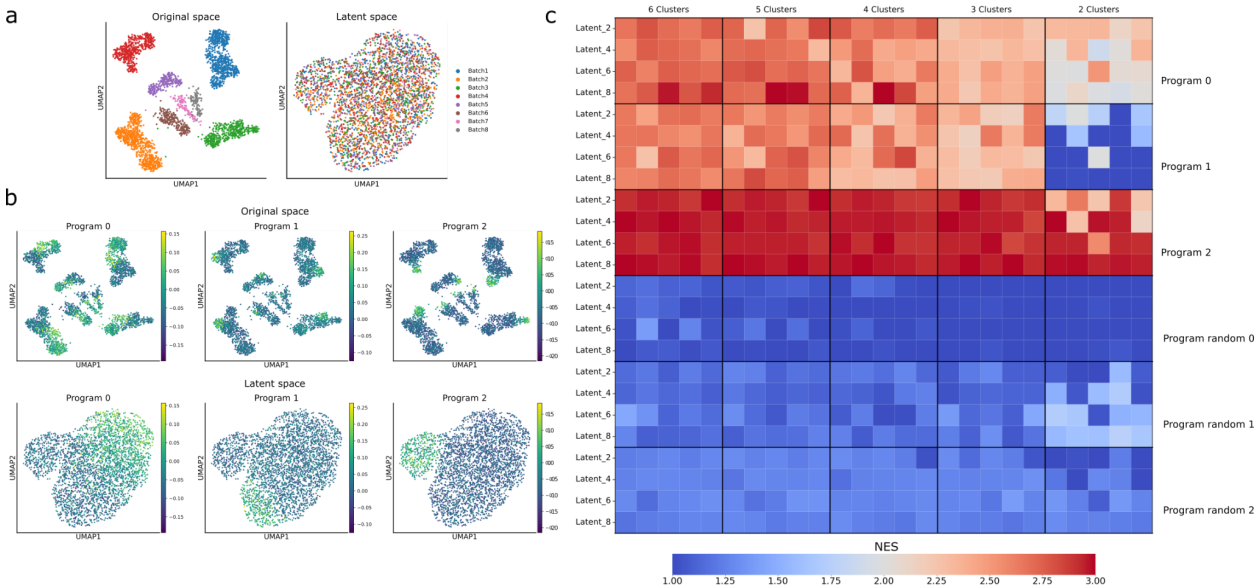


**Figure 2: CanSig rediscovers known transcriptional programs in simulated data.** UMAP representation of the simulated dataset A (see Methods) in the raw count space (original space) and the latent space computed by scVI (*d=6*), (**A**) colored according to the batch and (**B**) colored according to the

program expression (programs 0, 1, 2). (**C**) Heatmap representing 100 runs of CanSig on the simulated dataset A with different hyperparameter configurations, using scVI in the integration module and Leiden clustering in the postprocessing. Each square corresponds to a single run. Each meta-row represents the expression of a program. Programs 0, 1, 2 are expressed in the dataset; programs random 0, 1, 2 are not expressed in the dataset and used as random controls. Each meta-column represents a different configuration for Leiden clustering (2 to 6 clusters). Each sub-row represents a different configuration for scVI (dimensionality of the latent space from 2 to 8). Each sub-column represents a different random seed to compute the scVI latent space. The squares are colored according to the maximum Normalized Enrichment Score (NES) associated with this pathway computed with GSEA across all clusters.

To investigate if CanSig could rediscover known pathways while removing unwanted sources of variation in a controlled setting, we applied CanSig to two simulated datasets (Appendix A) with three expressed programs and eight samples. The latent spaces computed with scVI successfully removed the strong batch effect present in the data (Fig 2A) while retaining an meaningful organization according to the three expressed programs (Fig 2B). The three expressed programs were consistently discovered using CanSig over a range of hyperparameters while the random controls were never uncovered (Fig 2C). Noteworthily, the only setting in which the three programs were not recovered is when the number of clusters used in CanSig was smaller than the number of programs ($k=2$). The results obtained on simulated dataset B that contained more inter-patient heterogeneity were similar (Supplementary Fig. 1).

## CanSig rediscovers previously investigated signatures in three experimental datasets

We applied CanSig to three experimental scRNA-seq datasets: esophageal squamous cell carcinoma[7] (ESCC), colorectal cancer[5] (CRC), and breast cancer[36] (Fig 3A). After pre-processing (see Methods), between 19 and 62 individual patients were used per study, comprising around 15,000 to 41,000 malignant cells and around 14,000 to 30,000 genes.

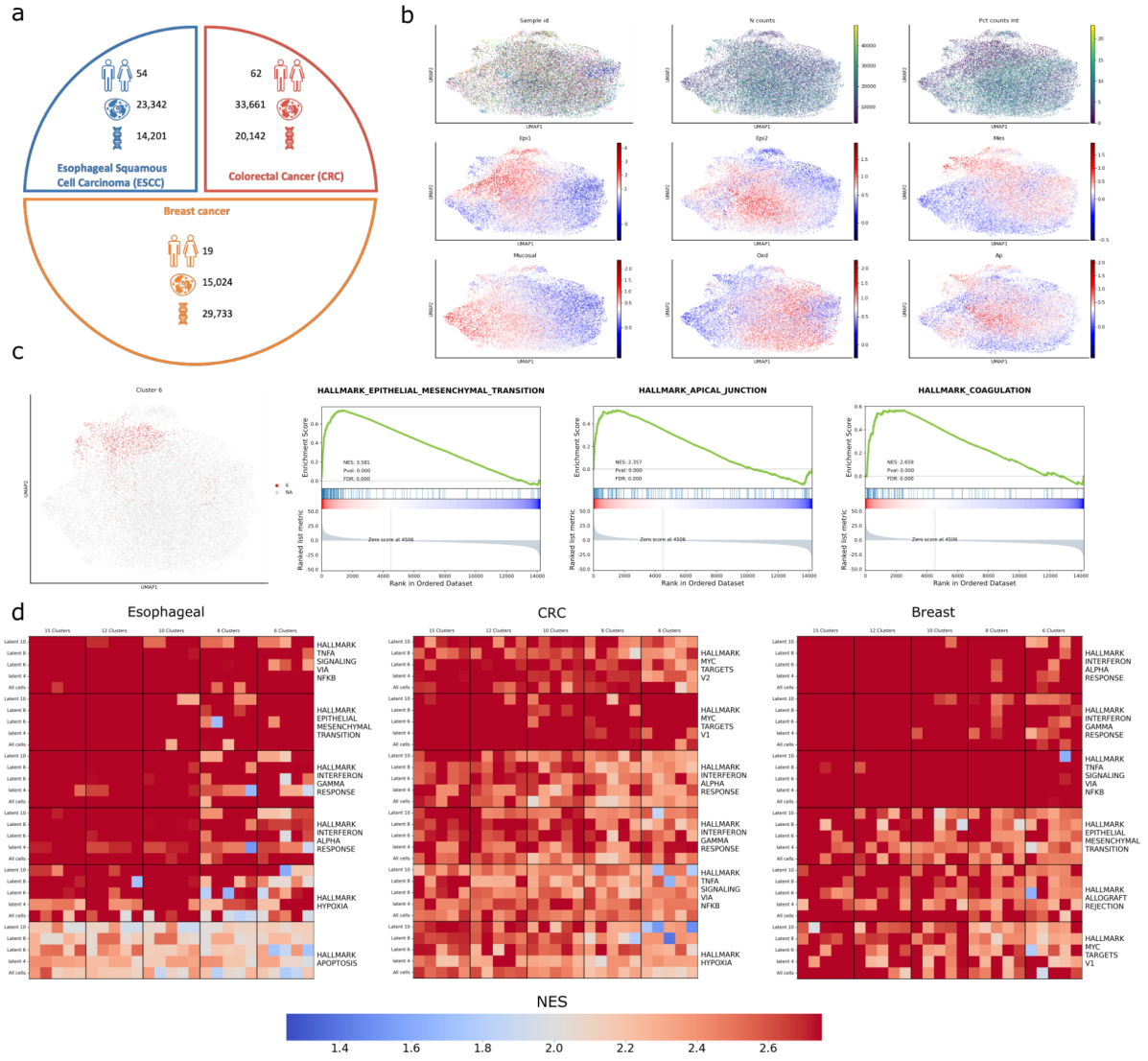


**Figure 3: CanSig rediscovers known signatures in three experimental scRNA-seq datasets.** (**A**) The number of samples, of malignant cells and of genes retained after pre-processing for the three analyzed datasets. (**B**) UMAP representation of the latent space of Esophageal Squamous Cell Carcinoma (ESCC) obtained using scVI *(d=6)*, colored according to the sample ID, the number of counts in the cell, the percentage of counts mapping to the mitochondrial genome, and six signatures described in the original paper[7]: Epi1, Epi2, Mesenchymal (Mes), Mucosal, Oxidative Stress (Oxd) and Antigen Processing (AP). For signatures, the color scheme is centered around the associated score median. (**C**) UMAP representation of the same latent space, cluster 6 colored in orange; GSEA plots showing the three most enriched pathways of the Hallmarks of Cancer discovered in cluster 6. (**D**) Heatmap representing 125 runs of CanSig on all three scRNA-seq datasets with different hyperparameter configurations, using scVI in the integration module and Leiden clustering in the postprocessing. Each square corresponds to a single run. Each meta-row represents the expression of a hallmark of cancer. Each meta-column represents a different configuration for Leiden clustering (6 to 15 clusters). Each sub-row represents a different configuration for scVI (dimensionality of the latent space from 4 to 10). “All cells” corresponds to a run without filtering out cells in the G2M or S phase. Each sub-column represents

a different random seed to compute the scVI latent space. The squares are colored according to the maximum Normalized Enrichment Score (NES) associated with this pathway computed with GSEA in any cluster.

The latent spaces computed with scVI successively captured previously known signatures as described in the original papers describing the studied datasets (Fig 3B, Supplementary Fig. 2-3). For instance, we see for ESCC that the latent space captured a gradient for the Mucosal signature and the Antigen Processing (AP) signature. The Mucosal signature is associated with the innate immunity while the AP signature is associated with the adaptive immunity. Thus, the gradient of Mucosal towards AP as well as the overlap of both signatures was consistent with the transition from an innate immune response to an adaptive immune response.

Of note, some of the described pathways were not clearly detected in our latent spaces. For example, the signatures associated with cycling cells in breast (GM4) and ESCC (Cycling) did not show any strong gradient (Supplementary Fig. 2). A reason for this is that we either condition on the cell cycle in our pipeline or remove cycling cells, thus effectively removing the latent organization of cycling pathways. Another example is that we did not find a clear organization for the Her2 signature in breast cancer, potentially linked to the clonal nature of the subtype that could drive inter- rather than intra-patient heterogeneity.

We found that the pathways enriched in specific clusters were consistent with the scores enriched in the cluster. For instance, we see that the three most significantly enriched hallmarks of cancer found in cluster 6 of ESCC, enriched in the Mesenchymal score, were epithelial-to-mesenchymal transition (EMT), coagulation and apical junction, all strongly correlated with the mesenchymal score in the paper of origin[7] (Fig. 3C).

The stability analysis run using a t-test showed that the hallmarks of cancer significantly enriched in a cluster were consistently uncovered over a range of hyperparameters (Fig. 3D, Supplementary Fig. 4-5). In particular, the most strongly associated pathways with at least one cluster were present over a range of hyperparameter configurations, thus pointing towards a true biological signal present in the data. For instance, we see Tumor Necrosis Factor (TNF-)α signaling was systematically uncovered in ESCC with NES>2 regardless of the hyperparameter configuration. Some hallmarks were consistently uncovered across cancer types (Fig 3D), such as TNF-α signaling (ESCC, CRC, breast), interferon (IFN-)γ response (ESCC, CRC, breast), apoptosis (ESCC, CRC, breast), interferon (IFN-)α response (ESCC, CRC, breast), MYC targets v1 (CRC, breast), hypoxia (ESCC, CRC) and epithelial-to-mesenchymal transition (ESCC, breast). These have generally already been described in the context of these cancer types, either directly in the original paper[7,36] or in external literature[39,40]. Of note, we also computed the stability analysis with the Wilcoxon rank-sum test and obtained similar results, although the order of some pathways was changed (Supplementary Fig. 6).

To further understand the link between the found pathways and known signatures, we scored the cells according to the discovered pathways and computed the Pearson correlation between the uncovered pathways and known signatures. The discovered hallmarks were mostly

consistent with the previously reported links between the uncovered signatures and hallmarks of cancer (Supplementary File 1). For instance, in breast cancer, TNF-α signaling was the most strongly correlated with GM1 (r=0.78), followed by GM7 (r=0.44), and IFN-γ response most strongly correlated with GM3 (r=0.70), followed by GM6 (r=0.44) as described in the original paper[36]. Another example is the strong correlation between EMT and the Mesenchymal signature (r=0.85) and between TNF-α signaling and the Stress signature (r=0.90) in ESCC as previously described[7].

## CanSig uncovers the expression of orexin and corticotropin releasing hormone receptors that can inform potential treatment resistance and therapies

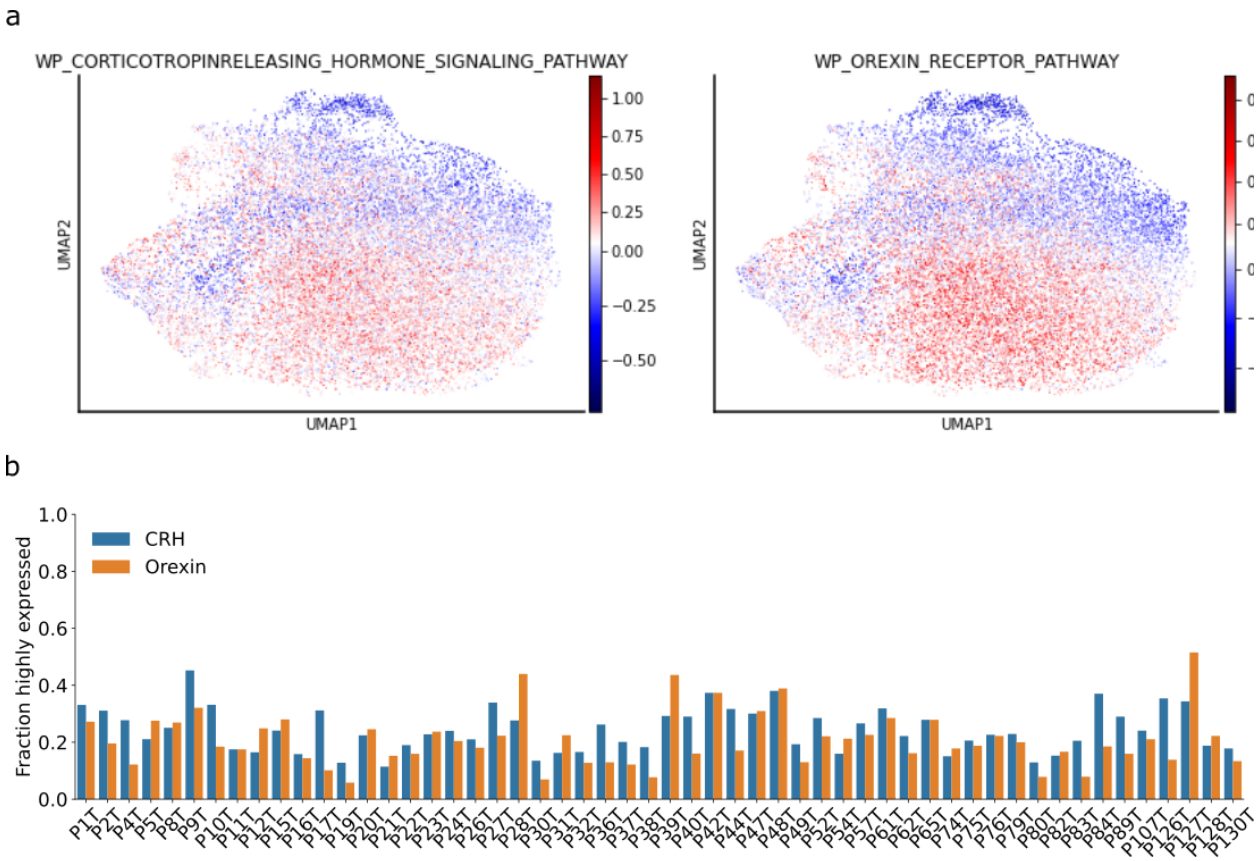


**Figure 4: CanSig uncovers new programs in esophageal squamous carcinoma associated with potential targeted therapies.** (**A**) UMAP representation of the latent space of ESCC using scVI (*d=6*) colored according to the corticotropin releasing hormone (CRH) signaling pathway and the orexin receptor pathway. (**B**) Proportion of cells of ESCC patients with a score in the highest 75 percentile in the CRH signaling or orexin receptor pathway, showing extensive inter-patient heterogeneity in the activation of the pathway

To showcase how CanSig can be used to investigate uncovered signatures to inform treatment resistance or therapy, we applied CanSig using the WikiPathways[37] dataset to ESCC.

We uncovered a number of significantly enriched pathways in at least one cluster that showed a strong Normalized Enrichment Score (NES) (Fig 4A, Supplementary Fig. 7). In particular, the Corticotropin Releasing Hormone (CRH) signaling pathway and the orexin receptor pathway were upregulated in a subspace of the ESCC latent space. These pathways have been previously shown to be overexpressed in several cancers and can be used for targeted therapy[41–44].

The effect of CRF is cancer-type specific: CRF binding to the CRF receptor can potentiate an anti-tumor response in endometrial cancer[45] and promote cancer progression in colon cancer[46] or breast cancer for instance[41]. Orexin receptors are expressed in gastric cancer (including esophageal cancer) and treatment with orexin has been shown to induce apoptosis in colorectal, neuroblastoma and pancreatic cancer cells[43]. Orexin A was even able to induce apoptosis in drug resistant colorectal cancer cell lines[47].

Interestingly, the cells presenting high expression of these pathways - defined as presenting a score above the 75 percentile - were present in different proportions across patients, showing strong inter-patient heterogeneity (Fig 4B). Indeed, we found that cells with a strong CRH (resp. Orexin receptor) signal consisted on average of 24% (IQR [18-30]) (resp. 21% (IQR [15-25])) of the cells of a patient, with some patients showing as little as 11% (resp. 6%) or as high as 45% (resp. 52%) of the cell type composition. These results could help stratify patients for targeted therapy using CRH agonists or orexin A into potential responders and non-responders.

## CanSig uncovers regional copy number variations associated with breast cancer programs favoring proliferation and aggressiveness

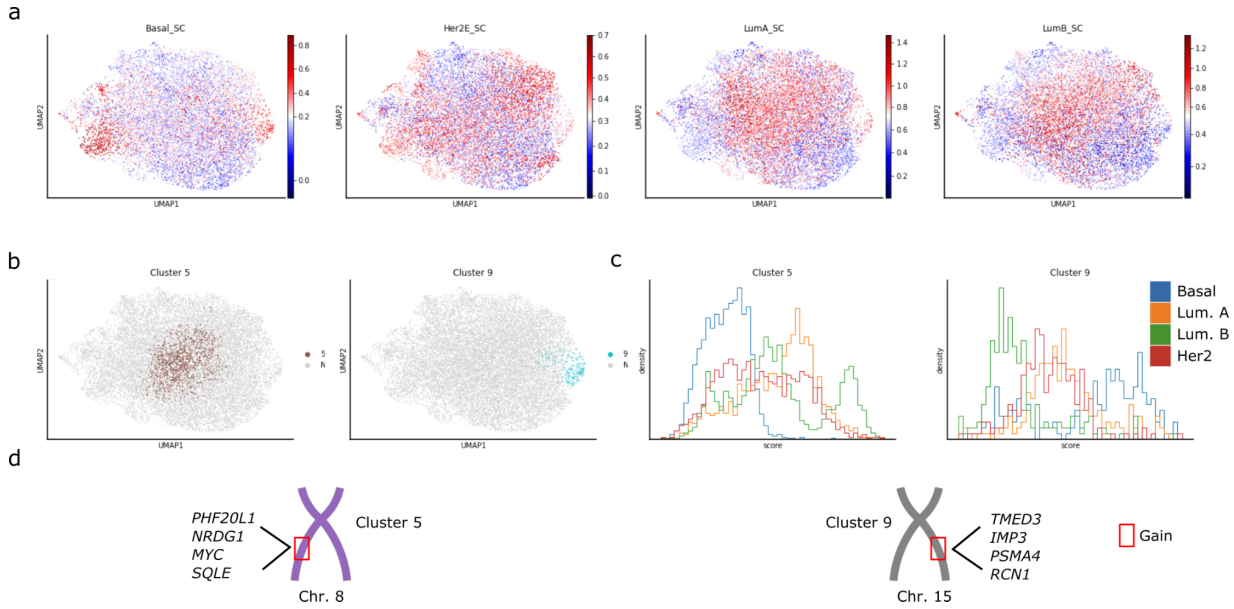


**Figure 5: CanSig uncovers significant Copy Number Variations associated with breast cancer programs.** (**A**) UMAP representation of the latent space of breast cancer obtained using scVI *(d=6)*, colored according to the four histopathological signatures described in the original paper[36]: Basal, Her2 positive (Her2E), Luminal A (LumA) and Luminal B (LumB). (**B**) UMAP representation of the same latent

space, colored according to the cluster of interests (clusters 5 and 9). (**C**) Associated density plots representing the distribution of the min-max normalized signature scores in the clusters 5 and 9. (**D**) Representation of the significant uncovered CNV gains and losses associated with cluster 5 (gain in chromosome 8q) and cluster 9 (gain in chromosome 15q). Known tumor suppressor genes or oncogenes present in the most differential gains are indicated.

CanSig also gives the possibility to uncover potential copy number gains or losses associated with previously known or uncovered signatures by performing differential copy number variation analysis. To showcase how CanSig can be used to uncover potential new links between copy number variation and known signatures, we applied differential CNV analysis (see Methods) to the breast cancer dataset.

We observed significant gains in cluster 9 associated with the Basal phenotype (Fig. 5C-F, Supplementary Table 1). Notably, we observed gains in the 15q region. Interestingly, a number of oncogenes and tumor suppressors were present in these CNVs. Some of the strongest gains in 15q included *TMED3, IMP3, PSMA4* and *RCN1*. In the context of breast cancer, *TMED3* and *RCN1* overexpression promote proliferation[48,49], *IMP3* is associated with EMT[50] and *PSMA4* correlated with poorer survival[51]. The basal phenotype in breast cancer is an aggressive type associated with poor prognosis, that typically expresses genes associated with proliferation[52] and EMT[53]. The CNVs uncovered thus are consistent with the basal phenotype enrichment in the cluster.

We observed significant gains in cluster 5 associated with the Luminal phenotype (Fig. 5C-F, Supplementary Table 1). In particular, we observed gains in the 8q region. Some of the strongest gains in 8q included *PHF20L1, NRDG1, MYC* and *SQLE*. In the context of breast cancer, *PHF20L1* overexpression promotes tumorigenesis and is associated with higher grades[54], *NRDG1* overexpression is associated with metastasis and mortality[55], *MYC* is a classical amplified oncogene[56] and *SQLE* overexpression promoted proliferation and invasiveness[57]. The luminal B subtype is an aggressive breast cancer subtype, thus consistent with these amplifications[58].

## CanSig finds a *de novo* signature linked to oxidative phosphorylation predictive of survival in breast cancer

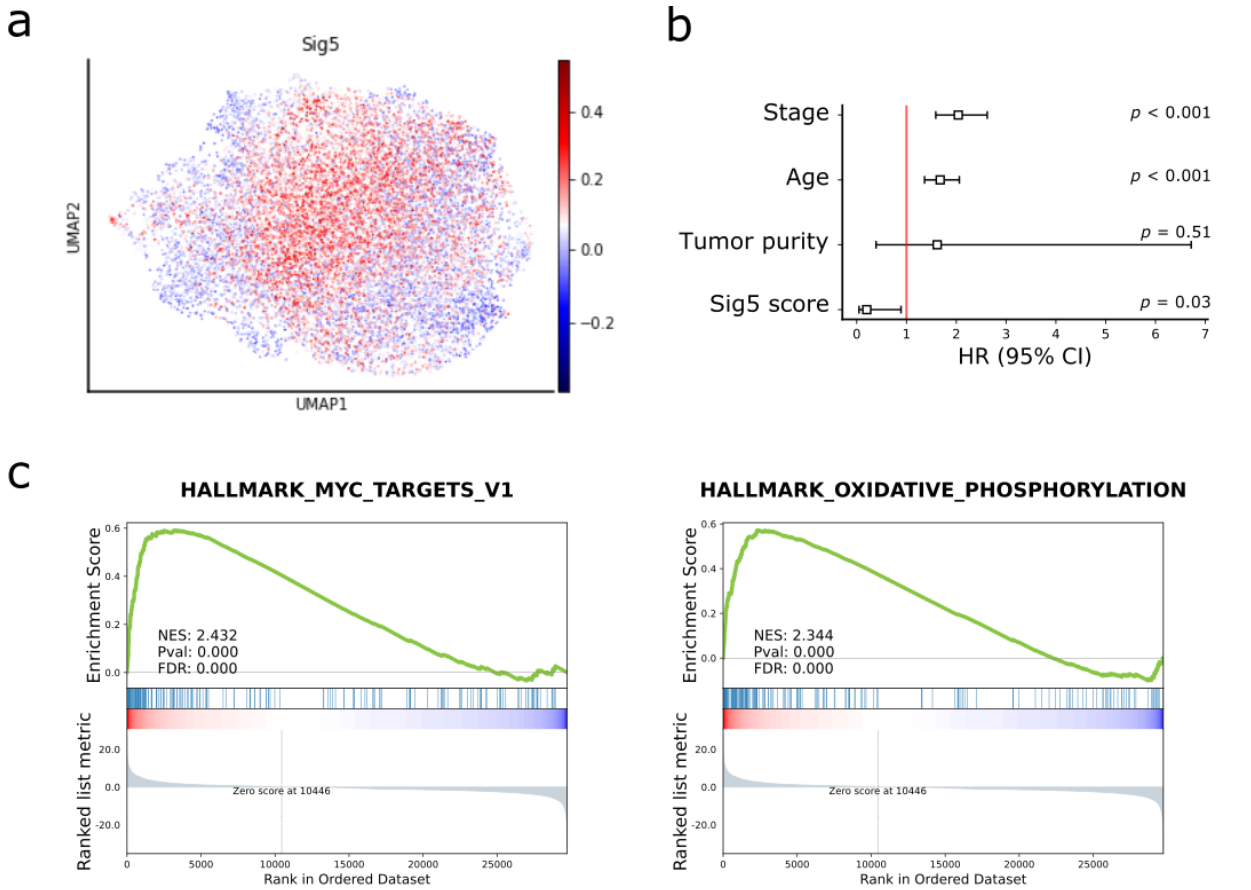


**Figure 6: CanSig uncovers a *de novo* signature in breast predictive of survival.** (**A**) UMAP representation of the latent space of breast cancer obtained using scVI *(d=6)*, colored according to the uncovered signature associated with cluster 5. Signature is defined as the 100 genes with the highest z-score. The color scheme is centered around the median. (**B**) Cox regression model representation of hazards for significant associations on the TCGA breast cancer dataset. The hazard ratios (HR, with 95% CI) associated with each variable for all significant cancer types are represented. The Sig5 score is computed as the mean of the standardized FPKM-UQ value of the genes in the uncovered signature. (**C**) GSEA plots of the two most significantly enriched hallmarks of cancer associated with cluster 5.

Finally, to showcase how CanSig could be harnessed to uncover *de novo* signatures leading to clinical and biological insights, each cluster was associated with a signature characterized as the top 100 most differentially expressed genes (*i.e.* with the highest z-scores). We found that the signature associated with the cluster 5 (Fig. 6A), Sig5, was predictive of survival in the TCGA breast cancer validation cohort after correction for age, stage and tumor purity (HR: 0.2 [0.05 - 0.89], $p$ = 0.03, Fig. 6B, Supplementary Table 2). Of note, this signature was the most significantly associated with survival among all the signatures previously uncovered in the paper of origin (GM1-7, Basal, Luminal A and B, Her2 positive).

The cluster 5 was enriched in MYC targets v1, oxidative phosphorylation, estrogen response early and DNA repair with NES>2 (Figure 6C). Although patients positive for the estrogen receptor typically show better prognosis than negative cases[59] and the switch from oxidative phosphorylation to glycolysis in breast cancer is typically associated with a more malignant subtype[59], no signature has previously been uncovered linking oxidative phosphorylation activation and prognosis. Some studies did however find signatures linked with bad prognosis that are negatively enriched in oxidative phosphorylation[60,61].

# Discussion

We propose CanSig, a modular method that enables *de novo* discovery of transcriptional signatures in cancer and their subsequent analysis. CanSig contains a pre-processing module that takes patient-level raw counts as input and performs standard quality control and additional cleaning using CNVs. CanSig then integrates the data from several patients while removing sources of unwanted variation through its integration module and uses the transformed space latent representation to obtain pathways enriched in subspaces through sequential clustering, differential gene expression analysis and gene set enrichment analysis in the postprocessing module. CanSig outputs consistently enriched pathways through a stability analysis and additionally gives the opportunity to study potential driving CNVs through optional differential CNV analysis.

To the extent of our knowledge, CanSig is the first cancer-specific method for *de novo* transcriptional signature discovery. Working with cancer data poses a unique set of challenges, notably from the integration perspective. Indeed, cancer cell expression can be driven by CNVs, which makes the batch effect between patients all the more potent and hinders the discovery of shared transcriptional states. Additionally, strong confounders such as cell cycle can also mask the signal from more tumor-specific sources of heterogeneity[2]. In effect, signature discovery in cancer is generally inconsistently performed across studies: some authors use NMF on each patient individually and investigate the shared uncovered signatures[5]; others use trajectory inference[6] or patient-level clustering followed by subsequent grouping of the signatures[7,36]. It is often the case that the issue of patient integration is not addressed at all[9,11,12]. Some studies perform integration followed by clustering similar to CanSig[10,62], however, these studies do not directly address all sources of unwanted variation and focus on removing batch effects. CanSig allows to address all challenges inherent to cancer data while obtaining cancer-specific information, *e.g.* the link between signatures and CNV profiles. Another strength of CanSig is that it provides signatures as a continuum rather than a discrete classification into subtypes, which is more adapted to the cancer setting[2,63].

In recent years, methods have been developed to discover de novo signatures in scRNA-seq, although none are specific to a cancer setting. scHPF[19] is a Bayesian factorization method developed for de novo discovery of gene signatures in scRNA-seq. RCC[22] is a recursive consensus clustering algorithm that can be applied to scRNA-seq to uncover novel subtypes. scETM[20] computes an autoencoder-based topic model embedding that can be used to deconvolve cell expression in several signatures. pmVAE[23] learns an interpretable latent space representation where each embedding represents a separate pathway that can be used to analyze the organization of cells. Finally, ExpiMap[21] is a deep learning model that enables reference mapping through a similar organization as pmVAE, while also giving the opportunity to learn new pathways through unconstrained nodes.
However, scHPF, RCC, and pmVAE are not adapted to a cancer specific task because of their lack of integrative power. ExpiMap's reference-query setting is not ideal in a general setting for cancer *de novo* signature discovery, as there is no perturbation setting or *a priori* discriminative

feature between cancer sets. Finally, the linear setting of scETM makes for less expressive power, notably to correct for batch effects.

CanSig shows some limitations. First, it is not necessarily adapted to datasets with very strong inter-patient heterogeneity. Indeed, if the inter-patient heterogeneity is too strong, the latent space might be driven by inter- rather than intra-patient heterogeneity, thus effectively masking the shared signal. Second, to discover rare subtypes, one needs either a large number of patients presenting said subtype or a strong identity of the subtype that will enable the latent space to capture it. Third, no single-cell specific differential gene expression methods are currently implemented. This was motivated by the ease of use of the implemented methods paired with reported good performance on single cell data[64]. Fourth, the *de novo* uncovered signatures are somewhat dependent on the cluster; however, as seen in the stability analysis, uncovered known pathways are consistent, which argues in favor of the stability of the biological meaning of the differentially expressed genes. Finally, CanSig uncovers signatures in a sequential manner as opposed to an integrated manner, which might be argued to yield false positives in the differential gene expression step[24]. However, as we do not use the significance associated with differentially expressed genes but rather the ranking of z-scores, we argue that this should not strongly impact the enriched pathways obtained.

The choice of a sequential discovery scheme was motivated by the desire to have a very modular pipeline that could subsequently integrate novel methods and give the user flexibility to adapt each step to the needs of their dataset. For instance, the pre-processing module gives the opportunity to further classify the data into malignant and nonmalignant cells by jointly considering cell type annotation and the CNV profile. This step might not be adapted to a cancer type presenting very little CNVs. In this case, the user can choose to skip the cleaning step altogether.

CanSig can be used in a number of manners. As illustrated by the discovery of a receptor-based signature potentially beneficial in stratification for patient treatment in ESCC, researchers can use CanSig to discover signatures that can then be further investigated for applicability in the clinics. As showcased by the discovery of significant gains and losses linked to the basal and luminal subtypes in breast cancer, researchers can use CanSig to investigate potential links between CNVs and discovered signatures. Finally, as shown by the discovery of a *de novo* signature predictive of survival in breast cancer, researchers can use CanSig as an exploratory tool to find new signatures to then further validate experimentally. CanSig can also be harnessed for other investigations, *e.g.,* linking the Tumor Microenvironment (TME) composition with a specific signature.

Some developments of the method include the addition of other integration models or the conditioning on CNV information as a proxy for patient identity in the integration module to allow for better patient-level integration without over-correction (in cases where the CNV profiles do not drive the program expression).

To conclude, CanSig is a user-friendly tool that enables known and *de novo* discovery of transcriptional programs in cancer. CanSig was capable of rediscovering known pathways in two simulated datasets and three real-world datasets. We showed how CanSig was able to identify new signatures in esophageal squamous cell carcinoma that could potentially inform patient stratification for targeted treatment. Furthermore, we showcased how differential CNV analysis could be used to complement signature discovery to uncover potential links between the CNV profile and signatures. Finally, we showed how CanSig was able to uncover a novel *de novo* signature in breast cancer predictive of survival in a validation cohort. In the context of rising generation and analysis of scRNA-seq in cancer, our approach can greatly facilitate the exploratory analysis of such data while taking into consideration cancer-specific effects, allowing for quick discovery of potential signatures for users with little to no expertise in computer science. These signatures can then be investigated from a biological perspective, for instance through differential CNV analysis or by performing further experimental work.

## Acknowledgments

We would like to thank Gunnar Rätsch, Mitch Levesque, Stefan Stark, Antoine Combremont, Gian Hiltbrunner, and Laure Ciernik for engaging in helpful discussions.

P.C. is funded through the ETH AI Center fellowship programme.

# Bibliography


1. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* **65**, 631–643.e4 (2017).
2. Suvà, M. L. & Tirosh, I. Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. *Mol. Cell* **75**, 7–12 (2019).
3. Suphavilai, C. *et al.* Predicting heterogeneity in clone-specific therapeutic vulnerabilities using single-cell transcriptomic signatures. *Genome Med.* **13**, 189 (2021).
4. Chen, F. *et al.* Single-Cell Transcriptomic Heterogeneity in Invasive Ductal and Lobular Breast Cancer Cells. *Cancer Res.* **81**, 268–281 (2021).
5. Pelka, K. *et al.* Spatially organized multicellular immune hubs in human colorectal cancer. *Cell* **184**, 4734–4752.e20 (2021).
6. Kim, N. *et al.* Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.* **11**, 2285 (2020).
7. Zhang, X. *et al.* Dissecting esophageal squamous-cell carcinoma ecosystem by single-cell transcriptomic analysis. *Nat. Commun.* **12**, 5291 (2021).
8. Brady, L. *et al.* Inter- and intra-tumor heterogeneity of metastatic prostate cancer determined by digital spatial gene expression profiling. *Nat. Commun.* **12**, 1426 (2021).
9. Dong, X. *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* **7**, 14 (2021).
10. Liu, Y. *et al.* Tumour heterogeneity and intercellular networks of nasopharyngeal carcinoma at single cell resolution. *Nat. Commun.* **12**, 741 (2021).
11. Ho, D. W.-H. *et al.* Single-cell RNA sequencing shows the immunosuppressive landscape and tumor heterogeneity of HBV-associated hepatocellular carcinoma. *Nat. Commun.* **12**, 3684 (2021).

12. Song, H. *et al.* Single-cell analysis of human primary prostate cancer reveals the heterogeneity of tumor-associated epithelial cell states. *Nat. Commun.* **13**, 141 (2022).
13. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
14. Rambow, F. *et al.* Toward Minimal Residual Disease-Directed Therapy in Melanoma. *Cell* **174**, 843–855.e19 (2018).
15. 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).
16. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
17. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
18. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
19. Levitin, H. M. *et al.* De novo gene signature identification from single-cell RNA-seq with hierarchical Poisson factorization. *Mol. Syst. Biol.* **15**, e8557 (2019).
20. Zhao, Y., Cai, H., Zhang, Z., Tang, J. & Li, Y. Learning interpretable cellular and gene signature embeddings from single-cell transcriptomic data. *Nat. Commun.* **12**, 5261 (2021).
21. Lotfollahi, M. *et al.* Biologically informed deep learning to infer gene program activity in single cells. *bioRxiv* 2022.02.05.479217 (2022) doi:10.1101/2022.02.05.479217.
22. Sonpatki, P. & Shah, N. Recursive Consensus Clustering for novel subtype discovery from transcriptome data. *Sci. Rep.* **10**, 11005 (2020).
23. Gut, G., Stark, S. G., Rätsch, G. & Davidson, N. R. pmVAE: Learning Interpretable Single-Cell Representations with Pathway Modules. *bioRxiv* 2021.01.28.428664 (2021) doi:10.1101/2021.01.28.428664.

24. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
25. Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol.* **17**, 29 (2016).
26. Prummer, M. *et al.* scROSHI - robust supervised hierarchical identification of single cells. *bioRxiv* 2022.04.05.487176 (2022) doi:10.1101/2022.04.05.487176.
27. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
28. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).
29. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
30. MacQueen, J. Some methods for classification and analysis of multivariate observations. in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* vol. 5.1 281–298 (University of California Press, 1967).
31. Reynolds, D. Gaussian Mixture Models. in *Encyclopedia of Biometrics* (eds. Li, S. Z. & Jain, A.) 659–663 (Springer US, 2009).
32. Ng, A., Jordan, M. & Weiss, Y. On spectral clustering: Analysis and an algorithm. *Adv. Neural Inf. Process. Syst.* **14**, (2001).
33. Campello, R. J. G. B., Moulavi, D. & Sander, J. Density-Based Clustering Based on Hierarchical Density Estimates. in *Advances in Knowledge Discovery and Data Mining* 160–172 (Springer Berlin Heidelberg, 2013).
34. Ntranos, V., Yi, L., Melsted, P. & Pachter, L. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. *Nat. Methods* **16**, 163–166 (2019).
35. Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18**, 174 (2017).

36. Wu, S. Z. *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 (2021).
37. Slenter, D. N. *et al.* WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* **46**, D661–D667 (2018).
38. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015).
39. Al Obeed, O. A. *et al.* Increased expression of tumor necrosis factor-α is associated with advanced colorectal cancer stages. *World J. Gastroenterol.* **20**, 18390–18396 (2014).
40. Kosmidis, C. *et al.* Interferon-γ and Colorectal Cancer: an up-to date. *J. Cancer* **9**, 232–238 (2018).
41. Kaprara, A., Pazaitou-Panayiotou, K., Kortsaris, A. & Chatzaki, E. The corticotropin releasing factor system in cancer: expression and pathophysiological implications. *Cell. Mol. Life Sci.* **67**, 1293–1306 (2010).
42. Couvineau, A. *et al.* The Anti-tumoral Properties of Orexin/Hypocretin Hypothalamic Neuropeptides: An Unexpected Therapeutic Role. *Front. Endocrinol.* **9**, 573 (2018).
43. Couvineau, A., Voisin, T., Nicole, P., Gratio, V. & Blais, A. Orexins: A promising target to digestive cancers, inflammation, obesity and metabolism dysfunctions. *World J. Gastroenterol.* **27**, 7582–7596 (2021).
44. Kim, B.-J. & Jones, H. P. Implications of corticotropin releasing factor in targeted anticancer therapy. *J. Pharm. Pract.* **23**, 86–90 (2010).
45. Graziani, G. *et al.* CRH inhibits cell growth of human endometrial adenocarcinoma cells via CRH-receptor 1-mediated activation of cAMP-PKA pathway. *Endocrinology* **143**, 807–813 (2002).
46. Lee, Y. *et al.* Corticotropin-Releasing Hormone Receptor Alters the Tumor Development and Growth in Apcmin/+ Mice and in a Chemically-Induced Model of Colon Cancer. *Int. J. Mol. Sci.* **22**, (2021).

47. Voisin, T. *et al.* Aberrant Expression of OX1 Receptors for Orexins in Colon Cancers and Liver Metastases: an Openable Gate to Apoptosis. *Cancer Research* **71**, 3341–3351 (2011).

48. Pei, J. *et al.* TMED3 promotes cell proliferation and motility in breast cancer and is negatively modulated by miR-188-3p. *Cancer Cell Int.* **19**, 75 (2019).

49. Nakakido, M. *et al.* Phosphatidylinositol glycan anchor biosynthesis, class X containing complex promotes cancer cell proliferation through suppression of EHD2 and ZIC1, putative tumor suppressors. *Int. J. Oncol.* **49**, 868–876 (2016).

50. Su, P. *et al.* IMP3 expression is associated with epithelial-mesenchymal transition in breast cancer. *Int. J. Clin. Exp. Pathol.* **7**, 3008–3017 (2014).

51. Chiao, C.-C. *et al.* Prognostic and Genomic Analysis of Proteasome 20S Subunit Alpha (PSMA) Family Members in Breast Cancer. *Diagnostics (Basel)* **11**, (2021).

52. Rakha, E. A., Reis-Filho, J. S. & Ellis, I. O. Basal-like breast cancer: a critical review. *J. Clin. Oncol.* **26**, 2568–2581 (2008).

53. Sarrió, D. *et al.* Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype. *Cancer Res.* **68**, 989–997 (2008).

54. Hou, Y. *et al.* PHF20L1 as a H3K27me2 reader coordinates with transcriptional repressors to promote breast tumorigenesis. *Sci Adv* **6**, eaaz0356 (2020).

55. Sevinsky, C. J. *et al.* NDRG1 regulates neutral lipid metabolism in breast cancer cells. *Breast Cancer Res.* **20**, 55 (2018).

56. Tokgun, P. E., Tokgun, O., Kurt, S., Tomatir, A. G. & Akca, H. MYC-driven regulation of long non-coding RNA profiles in breast cancer cells. *Gene* **714**, 143955 (2019).

57. Tang, W., Xu, F., Zhao, M. & Zhang, S. Ferroptosis regulators, especially SQLE, play an important role in prognosis, progression and immune environment of breast cancer. *BMC Cancer* **21**, 1160 (2021).

58. Li, Z.-H., Hu, P.-H., Tu, J.-H. & Yu, N.-S. Luminal B breast cancer: patterns of recurrence

and clinical outcome. *Oncotarget* **7**, 65024–65033 (2016).

59. Avagliano, A. *et al.* Mitochondrial Flexibility of Breast Cancers: A Growth Advantage and a Therapeutic Opportunity. *Cells* **8**, (2019).
60. Li, X. *et al.* A Signature of Autophagy-Related Long Non-coding RNA to Predict the Prognosis of Breast Cancer. *Front. Genet.* **12**, 569318 (2021).
61. Huang, Y., Wang, H. & Yang, Y. Annexin A7 is correlated with better clinical outcomes of patients with breast cancer. *J. Cell. Biochem.* **119**, 7577–7584 (2018).
62. Gouin, K. H., 3rd *et al.* An N-Cadherin 2 expressing epithelial cell subpopulation predicts response to surgery, chemotherapy and immunotherapy in bladder cancer. *Nat. Commun.* **12**, 4906 (2021).
63. Fan, J., Slowikowski, K. & Zhang, F. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Exp. Mol. Med.* **52**, 1452–1465 (2020).
64. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).