Guided-topic modelling of single-cell transcriptomes enables joint cell-type-specific and disease-subtype deconvolution of bulk transcriptomes with a focus on cancer studies

Lakshmipuram Seshadri Swapna[1], Michael Huang[1], and Yue Li[1,*]

[1]School of Computer Science, McGill University, Montreal, QC, Canada

[*]Correspondence to: yueli@cs.mcgill.ca

Abstract

Cell-type composition is an important indicator of health. We developed a method called Guided Topic Model for deconvolution (GTM-decon) to automatically learn cell-type-specific gene signatures from single-cell RNA-seq data based on their cell labels to deconvolve bulk RNA-seq data. GTM-decon confers accurate cell-type deconvolution throughout comprehensive experiments on simulated and real data. GTM-decon is versatile: using cell-type-specific guides to identify cell-type of origin in pancreatic and breast cancer, phenotype-specific guides to segregate breast cancer subtypes, and a nested-guided design with breast cancer subtype at the primary level and cell type at the secondary level to identify cell-type-specific differentially expressed genes.

Introduction

Cell-type composition and its relative proportions in a tissue is an indicator of health. For example, several studies have shown that Type 2 Diabetes is characterized by reduced beta cell mass and number in pancreatic tissue [1, 2]. In acute myeloid leukemia (AML), cell-type abundance variation between patients was indicative of degree of malignancy [3]. Experimental approaches such as fluorescence-activated cell sorting (FACS) and immunohistochemistry (IHC) are used to elucidate cell-type composition of biological samples. Single-cell RNA-sequencing (scRNA-seq) technology enables high-resolution transcriptome analysis, providing molecular insights into the cell-type composition, cell-state behaviour and cell-type heterogeneity [4-7] . In the context of cancer research, scRNA-seq has led to identification of distinct cancer cell states. These are shown to occur across a range of cancer types, and are implicated in tumor progression, with high cell-type heterogeneity associated with poor prognostic outcomes [8, 9]. However, challenges such as high cost, lower throughput and difficulties in sample preparation (e.g., difficulty in cell dissociation of all cell types in solid samples), make it hard to apply these experimental approaches at the patient population scale. Subsequently, computational approaches have been developed to deconvolve cell-type proportions from bulk RNA-seq datasets, which are substantially cheaper and easier to obtain.

Bulk RNA-seq data has been the workhorse of the transcriptome community over the past decades, with their falling costs and ease of experimental setup making it an attractive tool to work with any organism [10]. Consequently, there are several databases hosting tremendous

amounts of bulk RNA-seq data – Gene Expression Omnibus (GEO) [11, 12], Genotype-Tissue Expression (GTEx) [13], The Cancer Genome Atlas (TCGA), a repository of multi-omic bulk genome wide profiles including bulk RNA-seq data for more than 11,000 primary cancer samples [14]. However, bulk RNA-seq data reflects an average mix of gene expression profiles in the tissue. Deconvolving this data into constituent cell-type components can not only yield the cell-type proportions facilitating clinical investigation but also enable high-resolution differential expression analysis of genes. Studying these genes can help uncover gene regulatory programs driving different tissue states, thereby accounting for cell type and cell-state variation. Several deconvolution methods have been developed to this end.

Most of these approaches require an arbitrary cell-type specific signature matrix for the cell types of interest. The earlier methods generated these matrices from differential expression analysis of sets of purified or enriched cell types, and achieved good performance in mixtures with highly distinct cell types such as blood [15]. CIBERSORT improved on these approaches by incorporating a feature selection step wherein genes are adaptively selected from the signature matrix based on the input bulk RNA-seq data [16] . It uses linear support vector regression (SVR) to delineate closely related cell-type subsets of leukocytes. BSEQ-sc was one of the first methods to use scRNA-seq data for training, combining it with the CIBERSORT approach to perform deconvolution [17]. CIBERSORTx also uses scRNA-seq as reference profiles, along with improved normalization schemes to suppress cross-platform variation, and an adaptive noise filter to eliminate unreliably estimated genes [18]. Another method called MuSiC adopts a weighted non-negative least-squares regression model and addresses the issue of cross-subject heterogeneity in cell-type specific gene expression as well as within-cell-type stochasticity of single-cell gene expression arising in datasets from multiple subjects [19]. Related to MuSiC, EPIC [20] uses constrained least square regression and reference profiles from scRNA-seq data to deconvolve bulk samples but accommodates user-defined reference profiles to account for the presence of uncharacterized cell types. Bisque also uses a regression-based approach, learning gene-specific transformations of the bulk data based on the single cell reference profiles and the corresponding cell proportions to account for differences in scRNA-seq protocols and bulk RNA-seq protocol [21]. A more recent method, Scaden, uses deep learning based on simulated bulk RNA-seq datasets generated from scRNA-seq data to learn ensemble models for cell-type deconvolution of RNA-seq and microarray datasets [22]. The neural network conferred accurate deconvolution results but compromised interpretability, as certain features may be unknown combinations of gene expression relating to specific cell types. Another method developed recently is BayesPrism, which uses Bayesian inference to model scRNA-seq data jointly with bulk RNA-seq data to infer cell-type composition and their proportions [23]. The joint modeling overcomes biases that may arise due to technical and biological differences.

In this study, we present guided topic model for cell-type deconvolution (GTM-decon). Because of the high interpretability of the guided topics, GTM-decon automatically learns the contribution of each gene for each cell type, deriving cell-type specific gene signatures from scRNA-seq data based on their cell-type labels. We train GTM-decon on pancreatic, breast, and immune scRNA-seq datasets and used the learned matrices to deconvolve bulk RNA-seq datasets from healthy

and diabetic patients, as well as two cancer datasets: pancreatic cancer and breast cancer. GTM-decon successfully identifies the cell type of origin for pancreatic and breast cancer datasets. Interestingly, the results for human pancreatic cancer are recapitulated even while learning cell-type specific signatures from mouse pancreas, postulating cross-species deconvolution as an option where it is difficult to obtain scRNA-seq data due to technical or ethical challenges. Furthermore, our GTM framework also enables the inference of phenotype-specific gene signatures from bulk RNA-seq data by using the cancer phenotypes (e.g., breast cancer subtypes) and cell types as the nested guides for the topic inference. We leverage this capability to distinguish Basal from Estrogen receptor (ER) positive (ER+) breast tumor samples, not only achieving high classification accuracy but also identifying the genes and pathways that segregate the cancer subtypes. Furthermore, we are able to fine-tune the inferred cell-type specific gene signatures based on the cancer phenotype labels from bulk RNA-seq data. This enables us to deconvolve the average gene expression difference among cancer subtypes in the bulk RNA-seq into cell-type specific expression changes, thereby enabling discovery of the high-resolution aberrance of gene regulatory programs implicated in breast cancer progression.

Results

Overview of GTM-decon model

We developed GTM-decon to model cell-type information from scRNA-seq data and subsequently use it to deconvolve bulk RNA-seq data (**Figure 1a**). GTM-decon uses a guided topic modeling mechanism, innovatively adapted from our recently developed MixEHR-Guided [24]. In a nutshell, GTM-decon works by factorizing the gene expression data from the large single-cell gene expression datasets (cells-by-genes matrices) into two matrices: a genes-by-topics (i.e., genes-by-cell-types) matrix ($\mathbf{\Phi}$) capturing the importance of gene expression of each gene for each cell type, and a cells-by-topics matrix ($\mathbf{\Theta}$) capturing the importance of different topics (i.e., cell types) in each cell. The modelling is guided by the prior knowledge of cell-type labels in the scRNA-seq datasets. Each cell type is modelled as a topic (or a set of topics). To ensure one-to-one mapping between topics and cell types, we cast each topic to exactly one cell type via the means of the Bayesian topic prior values assigned to the α hyperparameter, which are set to high values in comparison with the rest of the topics for topic(s) corresponding to that cell type (see **Methods**). This guides the topic inference of $\boldsymbol{\theta}_d$ for each cell *m* by anchoring each topic at the cell type observed in cell *m*, which in turn influences the genes-by-topics ($\mathbf{\Phi}$) inference. The latter as a global parameter is subsequently used to infer the cell-type proportion in a bulk RNA-seq sample, which manifests as averaged gene expression from all cell types in a tissue of interest. Deconvolution of bulk RNA-seq datasets enables their segregation into different constituent cell types, thereby elucidating their cell-type proportions. This in turn can be used as a surrogate to gauge the cognate biological states of the subjects, such as their health status or disease progression.

The basic GTM-decon uses a single topic to model each cell type. However, in order to capture possible cell-state variations in gene expression, or different sub-cell-type gene regulatory programs, GTM-decon can also represent each cell type as a combination of multiple topics per cell type, with each topic capturing sub-cell-type differences in gene-expression within the same cell type. While in most of our applications we find that the single-topic per cell type is sufficient

to capture most of the variation of gene expression, the GTM-decon with multiple topics per cell type tends provided better performance in terms of deconvolution accuracy on simulated bulk RNA-seq data using real scRNA-seq data (**Figure S1**).

We also assessed the effects of using different genes selection strategies and different data transformations on the deconvolution performance. Specifically, we tested the effect of using only the highly variable genes, which corresponds to the specialized set of genes expressed predominantly in that cell type. We identified the highly variable genes (HVG) for each cell type and used their gene expression profiles as input set for training GTM-decon. However, it is preferable to perform no preprocessing and assess if using all the genes in the dataset could lead to comparable deconvolution performance. Therefore, we used all gene expression profiles as input to build a reference model. Analyzing the performance of these two approaches, along with the log-transformations of the values, shows that using all genes performed as well as or better than using only the HVG (**Figure S2**). This also offers the added advantage of learning the individual contributions of each gene for each cell type in efforts of discovering novel marker genes.

Benchmarking bulk deconvolution using GTM-decon against SOTA methods

To quantitatively benchmark GTM-decon, we compared it against four existing deconvolution methods, namely Bisque [25], Bseq-sc [17], CIBERSORTx [18], and MuSiC [19], on three human scRNA-seq datasets (Pancreas – E-MTAB-5061, Breast Tissue – GSE113197, and Rheumatoid Arthritis (RA) Synovium – SDY998), generated using different technologies (Smart-seq2, 10x Genomics Chromium, CEL-Seq2). We performed leave-one-out cross validation, using the single-cell RNA-seq of the left-out individual to simulate the bulk RNA-seq (gene counts for that sample) as the validation data and its cell-type proportions as the ground-truth mixing proportions. As shown in **Figure 2a**, GTM-decon performs better than other models on the Pancreas and Breast Tissue datasets in terms of both Spearman R and Cross Entropy and conferred comparable performance on the RA Synovum dataset. Notably, GTM-decon achieves smaller variance for both the pancreatic and breast tissue datasets. We further ascertained the qualities of the predicted cell-type proportions of each method against the ground-truth cell-type proportions (**Figure 2b** and **2c**). Most methods perform relatively better on the Pancreas and Breast Tissue datasets compared to the RA Synovium dataset, whose cell types are similar and therefore more difficult to delineate.

GTM-decon automatically learns cell-type specific gene signatures only based on cell labels

We evaluated the performance of GTM-decon in recapitulating cell-type specific information as well as deconvolution using pancreatic tissue as a reference. The pancreas consists of several cell types including exocrine and endocrine. While the former aids digestion by secreting several enzymes, the latter regulates glucose uptake and processing by secreting hormones. With a vast literature documenting the biological roles of several cell types and their behaviour in healthy and diseased conditions, such as diabetes and cancer, the pancreatic tissue serves as a good benchmark to assess GTM-decon.

We trained GTM-decon on an scRNA-seq reference data of pancreatic tissue from Segerstolpe et. al [26], consisting of 2,209 cells, corresponding to 14 cell types from 10 individuals. The guided topic model captures distinct sets of cell-type specific gene signatures, as shown by the gene-by-topic probability distributions (i.e., the matrix $\boldsymbol{\phi}$) for the top 20 genes in each topic

(**Figure 3a**). Indeed, each topic recovers a large number of marker genes for the corresponding cell types based on two databases, namely CellMarker database [27], a manually curated resource of cell markers in human and mouse and the PanglaoDB[28], a database of marker genes generated from scRNA-seq datasets (**Figure 3a**). We further ascertained the cell-type coherence of each topic by gene set enrichment analysis (GSEA), while using the probabilities learnt for each cell type against the CellMarkerDB. For the three cell types with abundant marker genes – acinar, alpha, and beta, each of the 5 topics recovers the exact cell type as the top-most hit in the analysis, with adjusted p-value $\leq 1\times10^{-15}$ (permutation test) in most cases (**Figure 3b**).

Using Gene Set Enrichment Analysis (GSEA), we evaluated the values be assigned to the α hyperparameter (i.e., topic prior hyperparameters) for their ability to learn cell-type specific signatures. Varying values from 0.9, 5, 10, 20, 50, 100 reveals that a prior of 0.9 gives the best results (**Figure S3**). Furthermore, the enrichment of known marker genes for the main cell types suggested that 5 topics per cell type best captures the cell-type specific signatures (**Figure S4**). We also evaluated the effect of different number of cells per cell type. As expected, the topic confidence as measured by the average probabilities over the cell-type specific gene signatures increases with the increasing number of cells for that cell type (i.e., evidence) (**Figure S5**).

GTM-decon deconvolves variations in cell-type proportions between healthy and diseased tissue

Based on the inferred cell-type topic distributions $\boldsymbol{\phi}$, we used GTM-decon to deconvolve the cell-type proportions of bulk RNA-seq data from a cohort of 89 human pancreatic islet donors with and without Type 2 Diabetes (GEO: GSE50244) (**Figure 3c**). The dataset consists of 51 individuals with normal glucose tolerance (N), 15 with impaired glucose tolerance (IGT) and 12 with Type 2 Diabetes (T2D); it also has a good segregation of males (N_1=54) and females (N_2=35). As expected, the inferred proportions of most types of cells for these two sets of individuals are similar since they came from the same tissue. However, GTM-decon predicted a significant reduction in beta cells in T2D individuals (Pearson correlation with HbA1c PCC= -0.4, p-value = 0.00031; t-distribution with n-2 degrees of freedom) (**Figure 3d**). IGT and T2D individuals exhibit low Beta cell-type proportions (**Figure 3d**), which was supported by the literature [1, 2]. The increase in ductal cells is possibly caused by their regulation of glucose uptake (**Figure 3d**). These results are recapitulated even when using a different pancreatic dataset as reference (i.e., the PI-Baron reference dataset, generated via Drop-seq instead of PI-Segerstolpe, generated via Smart-seq) (**Figures 3e, 3f**), demonstrating the robustness of our method.

Deconvolution of bulk RNA-seq from pancreatic adenocarcinoma (PAAD) patients from TCGA accurately predicts cell type of origin

We next turned to deconvolving pancreatic tumor bulk samples from TCGA. For this application, we used GTM-decon with 5 topics per cell type, which showed the best performance for several datasets (**Figure S1** and **Figure S4**). Since the tumor microenvironment is known to be infiltrated with immune cells [29, 30] , we trained GTM-decon models on a reference dataset derived from not only the target tissue but also immune cells (**Methods**). Specifically, we leveraged the reference cell-type models of pancreatic tissue (PI-Segerstolpe) and immune cell-type models from RA synovium in order to deconvolve the bulk RNA-seq from patients with pancreatic cancer (PAAD, N = 174) into into specific cell types. The combined topic distributions

from both pancreatic tissue and immune tissue is harmonized to be at the same numerical scale to ensure meaningful comparisons of their relative proportions. Additionally, we also sought to identify possible novel cell types or pathways present in the bulk RNA-seq, which are not represented in the reference profiles. This is achieved by running an unguided topic model on the bulk RNA-seq data as detailed in the **Methods** section. We empirically chose the number of such *de novo* bulk RNA-seq (bulkRS) topics based on how well they could explain the variation observed in clinical phenotypes.

As shown in the heatmap of the combined deconvolution matrix (**Figure 4a**), we observe that the most abundant cell types are 3 main pancreatic cell types, namely ductal, acinar, and alpha, along with 2 immune cell types – fibroblast and monocytes. Notably, the cell type of origin is correctly predicted for all the samples: Ductal cells have the highest proportion for most PAAD samples (**Figure 4a**, brown rectangle), and acinar for a subset of the PAAD samples (**Figure 4a**, blue rectangle). This recapitulates the literature remarkably well as ductal cells are known to be the site of origin for most cases of pancreatic ductal adenocarcinoma; however, a subset arises from acinar cells [31]. More significantly, an unknown subtype is predicted to originate from alpha cells (**Figure 4a**, green rectangle). This is supported by the recent literature, which reported that these samples as the derivatives of the neuroendocrine pancreatic cancer (PNET) were in fact misclassified as pancreatic ductal adenocarcinoma (PDAC) [32] . PNET cancers are supposed to originate from alpha or beta cells [33]. In terms of immune cells, almost every sample is predicted to have a high proportion of fibroblasts, which are known to be a critical component of the tumor microenvironment (**Figure 4a**) [34].

Interestingly, very few of bulkRS topics cluster with the reference topics inferred from the scRNA-seq reference data (**Figure 4a**). This suggests that the bulkRS topics and the reference capture complementary information. Nonetheless, there is also overlap between the bulkRS topics and the scRNA-seq based topics. For example, bulkRS Topic 2 corresponds to the guided topic for Acinar cell type and bulkRS Topic 10 is similar to the guided topic for Alpha cell type.

Next, we analyzed whether the predicted cell-type proportions show any correlation with cancer stages. The proportion of ductal cell type increases with the advanced levels of T-stage (**Figure 4b**). Similarly, the proportions are also higher for the metastatic stage 'm1' in comparison with the non-metastatic 'm0' although the sample sizes are too small to draw clear conclusion. Among the bulkRS topics, Topic 5 stood out as it does not cluster with any cell type based topics and yet appears to have a significantly higher proportion in metastatic stage 'm1' in contrast to stage 'm0', suggesting that it may be enriched in pathways associated with metastasis. GSEA reveals that the top-most significant pathway is HALLMARK_EPITHELIAL_MESENCYHMAL_TRANSITION (adjusted p-value = 2.8×10^{-6}, permutation test), which consists of genes defining epithelial-mesenchymal transition, as in wound healing, fibrosis and metastasis.

We next estimated whether variation in cell-type proportions or bulkRS topics is indicative of survival time. To this end, we performed Cox Regression to regress the number of days patients lived since their cancer diagnosis on their inferred cell-type proportions. Specifically, we combined the cell-type proportions of all pancreatic, immune cell types as well as the *de novo* bulkRS topics. Overall, the Cox Regression model is statistically significant compared to the bias term (adjusted p-value = 2×10^{-7} based on likelihood ratio test). To explore the marginal effect of individual cell-type proportions on survival, we performed Kaplan-Meier analysis by separating patients into two groups based on K-means clustering (**Figure 4c**). We observe that

alpha cell type exhibits a significant hazard ratio, predicting good survival outcome (p-value = 0.049; log-rank test), which is supported by the literature as PNETs are fairly benign whereas PDACs are aggressive [35]. On the other hand, Topic 1 and Topic 5 indicate poorer survival, with hazard ratios of 4 and 40 respectively (**Figure 4c**) although their roles are unclear since they do not cluster with any of the cell-type-specific topics (**Figure 4a**).

Deconvolution of bulk RNA-seq from TCGA breast cancer differentiates Basal-like from ER+ samples

We deconvolved TCGA breast cancer (TCGA-BRCA) samples (N=1212) using a combination of cell types from normal breast tissue [36] and immune cell types from RA synovium dataset (**Figure 5**). Interestingly, the use of highly variable genes was more discriminatory in identifying cell type of origin in comparison to all genes in this analysis (**Figure S6**). The inferred cell-type-specific topic distributions from normal breast tissue recapitulates several marker genes from CellmarkerDB and PanglaoDB (**Figure 5a**). Furthermore, GTM-decon-inferred cell-type proportions clearly distinguishes TCGA breast tumor samples from GTEx normal breast tissues (**Figure 5b**). Moreover, the Basal cell type is predicted to have the highest proportion in the Basal subtype defined by the PAM50 classification in comparison to other subtypes (**Figure 5c**) [37]. Also, we observed higher predicted proportion for Luminal_2 cell type in both LumA and LumB subtypes as expected [36] (**Figure 5c**). Among the *de novo* bulkRS topics, Topic 5 is highly enriched in LumB, while Topic 7 and Topic 8 are enriched in Basal and LumA subtypes (**Figure 5c**).

The combined heatmap using breast and immune cell types along with bulkRS topics clearly segregates Basal subtype from ER+ (ER IHC positive) samples (**Figure 5d**). The guided topic score for the Basal subtype also correlates with higher proliferation score as expected [37]. Notably, the Basal cell type is enriched in this subtype (enclosed in green rectangle), whereas Luminal_2 is depleted (enclosed in green dashed rectangle). In contrast, ER+ samples appear to be enriched for Luminal_2 cells (enclosed in blue rectangle), as expected [36, 38]. As in the case of pancreatic cancer, Topic 7 clearly captures the Basal subtype (enclosed in brown rectangle), whereas there is no clear topic capturing the ER+ phenotype. Interestingly, GSEA of Topic 7 is not enriched for the Basal marker genes from CellMarker whereas the GTM-guided topic for Basal is highly enriched for the marker genes.

Deconvolution based on breast-cancer-specific single-cell transcriptomes reveals enrichment of subtype-specific cancer cells in TCGA-BRCA cohort

To capture specific subtypes of breast tumors from the TCGA, we trained GTM-decon on a more specific scRNA-seq reference data from 26 primary tumors of breast cancer patients with three major clinical subtypes of breast cancer, including 11 ER+, 5 HER2+ and 10 TNBC [39]. The data consists of about 1 million cells and covers 7 major cell types and 29 minor cell types. This produced an accurate annotation of the TCGA-BRCA tumor samples based on the major subtypes from the scRNA-seq dataset, namely Cancer-Basal, Cancer-Her2, Cancer-LumA, and Cancer-LumB (**Figure 6a**). As expected, significantly higher proportions of Endothelial and Myoepithelial cell types are found in the normal-like samples, in comparison to the cancer subtypes (**Figure 6a**). Furthermore, Basal subtype is enriched for Cancer-Basal cells (**Figure 6b**, shown in dotted brown rectangle), and the Cancer-Associated Fibroblasts (CAFs) are enriched in almost all the samples (**Figure 6b,** shown in blue rectangles). Similar to the above analyses, some of the bulkRS topics overlap with the most represented cell types. For example,

Topic 6 resembles Myofibroblast-like cancer-associated fibroblasts (myCAF-like), and Topic 8 resembles LumA subtype cancer cells.

Using GTM-decon trained on mouse pancreas scRNA-seq data to deconvolve bulk RNA-seq data from human pancreas

For cases where scRNA-seq data is not available from the organism of interest, due to either ethical, technical, or financial challenges, there is a need to leverage scRNA-seq collected from a model organism. To this end, we investigated the possibility of deconvolving bulk RNA-seq data using cell-type signatures learned from mouse gene expression data. We separately trained two GTM-decon models on the human and mouse pancreatic datasets. Specifically, the human datasets consists of the PI_Segerstolpe and Baron datasets, which consists of 2209 cells from 10 individuals and 7729 cells from 3 individuals, respectively; the mouse dataset consists of 1888 cells collected from 2 mice. For comparative analysis, we focused on only the common set of high-confidence orthologous genes between the two species mapped by ENSEMBL. We visualized the cell-type proportions comparing against the ground truth values (**Figure 7a**). As expected, the cell-type proportions deconvolved using the two human datasets recapitulate the ground-truth proportions very well (median PCC of 0.94 and 0.97). Interestingly, GTM-decon trained on the non-reference human dataset performed better than the one trained on the reference-matched dataset, which was probably due to the 10-fold higher number of cells in the former scRNA-seq dataset. Moreover, GTM-decon trained on the mouse reference dataset also performed quite well in terms of inferring the shared cell types (PCC of 0.94). Similarly, deconvolving the pancreatic cancer PAAD dataset using mouse reference dataset also conferred high quality patient clustering comparable to that of human dataset (**Figure 7b**). Notably, the pancreas adenocarcinoma other subtype is predicted to originate from alpha cells, as observed in humans (**Figure 7b**, shown in blue rectangles).

Extending GTM-decon to learn phenotype-guided topics from bulk RNA-seq data

Our guided topic mechanism is not limited to inferring cell-type-specific topics but can be extended to inferring phenotype-specific topics (**Figure 1b**), thereby discovering gene signatures of subtypes or different cancer stages. We applied this approach to study the differences between Basal and ER+ (**Figure 8a,b**) and the difference between ductal carcinoma and lobular carcinoma (**Figure 8d,e**). We modelled each phenotype using 5 topics each, based on a sparsified matrix of bulk RNA-seq data (see **Methods**), resulting in a genes-by-phenotypes $\boldsymbol{\phi}$ matrix with 5 topics per phenotype (**Figure 8a**). We then ranked genes by the topic scores under each topic. Almost all the genes identified by our approach were also deemed significant by DESeq2 differential analysis (**Figure 8a**). GSEA of the topics shows differences between the Basal and ER+ phenotypes, although there is not much difference among the 5 topics for Basal (**Figure 8c**). Moreover, the trained GTM-decon confers accurate phenotype classification with 97% accuracy for discriminating Basal and ER+ (**Figure 8b**) and 82% for discriminating ductal and lobular subtypes (**Figure 8d**), which is comparable to the traditional supervised learning methods namely logistic regression and Random Forest (**Figure 8f**). We also trained GTM-decon only on the highly variable genes in the bulk RNA-seq samples. This results in lower classification accuracy (76% for ductal-lobular samples as compared to 82%, and 95% for Basal-ER+ samples as compared to 97%), which might be due to information loss (HVG=1391

for Basal-ER+ samples, HVG=445 for ductal-lobular samples). Training using differentially expressed genes between the two phenotypes identified b DESeq2 also results in lower classification accuracy (**Figure 8b,8e**). These experiments suggest that sparsifying the data provides a practical alternative for training GTM-decon on bulk RNA-seq data.

Nested-guided topic modeling identifies cell-type-specific differentially expressed genes between breast cancer subtypes in bulk TCGA RNA-seq data

scRNA-seq can facilitate molecular understanding of genes and pathways in specific cell types with respect to phenotypic states. This level of detail is absent in bulk RNA-seq data, which profiles only the averaged gene expression from all cell types in the tissue. However, due to the cost, scRNA-seq profiles at the patient cohort size is rare. To take advantage of both types of data, we sought a way to identify cell-type specific gene expression differences corresponding to the bulk phenotype. Briefly, we took a pretrained GTM-decon on a scRNA-seq reference data and then updated its cell-type-specific topics based on the corresponding phenotypes from the bulk data (**Figure 1c**). This is similar to treating the phenotype as level 1 and cell types as level 2 in a two-stage nested factor design in statistics.

For the TCGA-BRCA data, in particular,we first initialized the genes-by-topics matrix $\boldsymbol{\phi}$ with the *pre-trained* guided topics learned from the scRNA-seq reference for normal breast tissue. We first *fine-tuned* 5 cell-type-specific topics for each phenotype (i.e., ER+ or Basal), resulting in a 10-topic model. During the fine-tuning, all cell-type-specific topics corresponding to the patient's phenotype are assigned a prior value of 0.9. Applying this approach to 798 samples from TCGA-BRCA, corresponding to Basal (N_1=140) and ER+ (N_2=658) led to a new genes-by-topics matrix $\boldsymbol{\phi}^*$. First of all, the top genes for the topics corresponding to the same cell type are similar in both phenotypes, as they are expected to capture cell-type specific signatures (**Figure 9a**). However, differential analysis of genes between two phenotypes revealed up-regulated genes in one phenotype being downregulated in the other (**Figure 9b**). These differences between Basal and ER+ are statistically significant across all genes for all the cell types based on paired Wilcoxon signed-rank tests. Next, we evaluated the ability of the fine-tuned GTM-decon to classify the held-out test set, and observed a 97% classification accuracy (**Figure 9c** top panel). A higher resolution picture of deconvolution of each of the topics reveals that ER+ samples are enriched for Luminal-2 cell type, whereas the Basal subtype is depleted for that cell type (**Figure 9c** bottom panel).

Next, we identified the differentially expressed (DE) genes per cell type by subtracting the genes-by-topics entry in $\boldsymbol{\phi}$ matrix for phenotype d (e.g., Basal) from Phenotype d' (e.g., ER+) under the same cell type (e.g., Luminal 2). To evaluate the consistency of our DE genes, we compared them against the DE genes identified by DESeq2 (**Figure 9d**). We observe that all upregulated (downregulated) DE genes nominated by DESeq2 in ER+ versus Basal comparison were also deemed upregulated (downregulated) by our approach. We visualize the expression of the 630 statistically significant DE genes (adjusted p-value < 0.05; permutation test) in the BRCA tumor samples (**Figure 9e**). We found that not only most of the DE genes identified by our approach are consistently identified by DESeq2 but we also identified several clusters of DE genes upregulated or downregulated in specific cell types. Over representation analysis (ORA) of these cell-type-specific DE genes against Hallmark pathways from MSigDb revealed several enriched pathways. As expected, Estrogen Response (Early) and (Late) pathways are highly upregulated in the ER+ phenotype, mainly in the Luminal_2 cell type (**Figure 9f**). Similarly, most DE genes in the Basal phenotype are upregulated for typical pathways involved in cancer, such

as G2M checkpoint, E2F Targets, Mitotic Spindle. This reflects the aggressive nature of Basal call type as the origin for the cancer subtype (**Figure 9f**). We obtained similar results comparing the 2 histological subtypes – ductal carcinoma and lobular carcinoma (**Figure S7**).

Discussion

In this study, we developed a highly efficient and interpretable Bayesian approach called *GTM-decon* for cell-type deconvolution, based on guided topic modeling (**Figure 1**), which learns cell-type-specific gene signatures in the form of categorical distributions (i.e., latent topics) from scRNA-seq datasets. During the topic inference of each cell, we introduce the guidance by setting the topic hyperparameter for the cell type of that cell to be relatively larger than the hyperparameters for other topics. This enables us to anchor each topic to a specific cell type and subsequently guide the inference of the global topic distributions over genes to automatically prioritize cell-type marker genes. The resulting topic distributions can then be used to infer the relative cell-type proportions from bulk RNA-seq datasets (i.e., cell-type deconvolution).

Our results suggests that most of the top genes under the cell-guided topics correspond to known marker genes (**Figure 3a**, **Figure 5a**). Because GTM-decon infers the probabilistic score of all the genes under each cell-type-guided topic, it can be used to not only quantify the contribution of the known marker genes but also score novel marker genes. In terms of cell-type deconvolution, GTM-decon confers comparable performance to the existing methods (**Figure 2**). The deconvolved cell-type proportions can be used to distinguish healthy and diseased samples as shown in the case of diabetic patients (**Figure 3c**,**e**) and normal and cancer samples from breast tissue (**Figure 5b**). This enables investigation of molecular contribution to the phenotypic differences. GTM-decon is not sensitive to different reference scRNA-seq datasets from the same tissue even if they were generated from different platforms (e.g., Smart-seq and Drop-seq) (**Figure 3d, f**). Furthermore, our approach is also generalizable to capture cell-type specific gene signatures from other species enabling cross-species deconvolution, which is especially useful in situations posing technical and ethical challenges (**Figure 7**).

Using GTM-decon, we revisited two cancer datasets from TCGA, namely pancreatic adenocarcinoma (PAAD) and breast cancer (BRCA). To dissect the heterogeneous tumor microenvironment, we combined the primary GTM-decon model trained on scRNA-seq data for the target tissue with the secondary GTM-decon model trained on immune cell types, reasoning that the latter play a predominant role in tumorigenesis. As a result, we identified the ductal and acinar origin of PAAD, the alpha origin of neuroendocrine pancreatic tumors (PNET), the basal origin of Basal subtype and the luminal origin of ER+ breast cancer (**Figures 4** and **5**). Using a cancer-specific scRNA-seq dataset for learning cell-type-specific signatures in the tumor microenvironment results in the identification of all the BRCA subtypes (**Figure 6**). By combining the *de novo* topics inferred directly from the bulk RNA-seq data with the scRNA-guided topics, we identified putative prognostic biomarkers that correlate with survival time (**Figure 4**).

We further extended GTM-decon by modelling the bulk RNA-seq data using the patient phenotype labels as the guide rather than cell types. In contrast to the traditional differential analysis approach, the phenotype-guided GTM-decon provides a different way to investigate gene signatures and molecular pathways underpinning the phenotypes of interest (**Figure 8**). To leverage the scRNA-seq reference data, we further extended this framework to a nested-

guided topic model by fine-tuning a dedicated set of cell-type-specific topic distributions for each phenotypic state (e.g., BRCA subtypes) of the patients from the bulk RNA-seq data. This enables learning not only intra-phenotype changes of cell-type distributions but inter-phenotype changes of gene expression. The latter allows us to identify cell-type-specific differential expression of genes directly from bulk RNA-seq data (**Figure 9**). This feature is not available in most approaches developed for cell-type deconvolution, except CIBERSORTx [18], which uses a non-negative matrix factorization approach based on partial observations to identify cell-type-specific differentially expressed genes across phenotypes.

As future works, we can adapt GTM-decon to leverage other single-cell omic data such as scATAC-seq for reference and deconvolve the equivalent omic in bulk samples. It can also be extended to work with different cell states, like in BayesPrism [23], by modelling the topics based on cell-states instead of cell types. Like all other deconvolution approaches, GTM-decon alone is unable to identify novel cell types from bulk RNA-seq datasets (i.e., cell types that are not present in the reference scRNA-seq data GTM-decon is trained on). while it is able to capture perturbed cell types based on the phenotypes (**Figure 9c**). This can be addressed by training GTM-decon on an atlas-level scRNA-seq data such as the Human Cell Landscape [40], which comprehensively cover most of the cell types in the primary tissues. The resulting model can then be used to deconvolve bulk samples of any given target tissue. Lastly, cell types are not independent entities but rather form a lineage. One future direction is to exploit the relations among the cell types, which may better capture the underlying phenotypic states of the subjects.

Methods

GTM-decon overview

We adapted GTM-decon from the MixEHR-Guided model [24] (which was in turn inspired by MixEHR [41] and sureLDA [42] to model gene expression from scRNA-seq data. In brief, the method assumes that cells are generated from a statistical generative process, such that each cell is a mixture of cell types, and each cell type is defined by a categorical distribution over the expression frequency of all genes.

Topic modeling infers a set of categorical distributions over a fixed vocabulary from a set of documents each treated as a bag-of-word (i.e., the word order is ignored). Each categorical distribution captures a distinct topic implicated in the documents. In our context, the documents are cells; the words are genes; the word frequencies are the transcript abundance of the genes measured by the scRNA-seq; the topics are the cell types. Standard topic model does not produce identifiable topics corresponding to the cell types. Our topic modeling uses a guided mechanism based on the cell-type labels, thereby associating each topic (or a set of topics) to exactly one unique cell type. This circumvents the topic identifiability issue [43] .

Overall, our GTM-decon workflow consists of three main steps: 1) assembling and processing of reference scRNA-seq dataset(s) (annotated with cell-type labels) to serve as training data for the tissue of interest; 2) using guided topic modeling to infer K cell-type-specific topic distributions; 3) using the inferred topic distributions as template themes to infer the relative cell-type proportions of bulk RNA-seq data. We describe each step in detail next.

Preprocessing the reference scRNA-seq data

The gene expression profiles of scRNA-seq data is used as training data. Since the performance of GTM-decon depends on how the scRNA-seq count data are processed, . we explored different gene selection and transformations of the scRNA-seq data. Specifically, two sets of genes are analyzed: a) all genes b) highly variable genes. For all genes, the following variations are considered:

- i) *all*: simple count data preprocessed to remove outliers (highly expressed genes – found in ⩾ 80% of cells, lowly expressed genes – found in ⩽ 5 cells);
- ii) *all_normr* – normalize counts (while excluding highly expressed genes for calculating the normalization factor) as counts divided by the sum of counts per cell multiplied by a scaling factor of 10,000;
- iii) *all_log1p_normr*: log-transform all gene counts and then normalize as in ii).

Highly variable genes are identified using the `highly_variable_genes` (HVG) function of `scanpy` [44]. These filtered data sets by the HVG are named based on the above steps as *hvg*, *hvg_normr*, and *hvg_log1p_normr* .

Modeling scRNA-seq reference data via topic inference guided by cell-type labels

We assume the following data generative process for the gene expression of each cell. Suppose there are K cell types. Each cell type endows a unique categorical distribution over all G genes, which is considered as the transcriptional rate of each gene. For instance, B cells have higher transcription rate for *CD19* compared to alpha cells, which have relatively high rate for *FXYD5*.

Each cell indexed by m ∈ {1, …, M} is a mixture of the K cell types. The cell-type mixture θ_m is sampled from a K-dimensional Dirichlet distribution Dir(α_m) with the hyperparameter α_m being specific to the cell. The key assumption here that separates GTM-decon from LDA [45] is the use of noisy cell-type label $y_m \epsilon \{1,..,K\}$ for the cell. The hyperparameter corresponding to the cell-type label $y_m = k$ has higher value ($\alpha_{m,k} = 0.9$ by default) in contrast to the rest of the hyperparameters $\alpha_{m,k'}$ for $k' \neq k$, each of which is sampled uniformly between 0.01 and 0.1. Note that having non-zero $\alpha_{m,k'}$ allows other cell types to be assigned to the cell (i.e., $\theta_{m,k'} \geq 0$). To generate the ith scRNA-seq read count for cell *m*, we first draw a latent topic $z_{i,m}$ from a K-dimensional categorical distribution with rate set to be θ_m: $z_{i,m} \sim Cat(\theta_m)$. Given topic $z_{i,m} = k$, we then draw a gene index $x_{g,m}$ from a G-dimensional categorical distribution: $x_{g,m} \sim Cat(\boldsymbol{\phi}_k)$, where $\boldsymbol{\phi}_k$ follows a G-dimensional Dirichlet distribution with a fixed hyperparameter β across all K dimensions: $\boldsymbol{\phi}_k \sim Dir(\beta)$. Therefore, each topic is clamped to a specific cell type such that the corresponding topic distribution $\boldsymbol{\phi}_k$ endows a cell-type-specific transcriptome landscape.

The posterior distribution for the latent variables $\boldsymbol{\theta}_m$, $z_{i,m}$ and $\boldsymbol{\phi}_k$ conditioned on the scRNA-seq reference data can be either approximated by Gibbs sampling [46] or by collapsed mean-field variational inference [47]. We took the latter approach. Under the mean-field variational distributions, all latent variables factorize into independent distributions, which allows closed-form coordinate ascent update for the variational parameters. For algorithmic convenience, we can leverage the conjugacy of the Dirichlet to Categorical distribution by integrating out $\boldsymbol{\theta}_{\mathrm{m}}$ and $\boldsymbol{\phi}_k$ and focusing on inferring the variational distribution of the topic assignments $q(z_{i,m}|\gamma_{i,m})$.

Specifically, we approximate the posterior distribution of the cell-type assignment $p(z_{i.m}|\mathbf{X},\mathbf{y})$ via the variational Categorical distribution $q(z_{i,m} = k) = \prod_k \gamma_{i.m.k}^{[z_{i.m}=k]}$, where

$$\gamma_{i,m,k} \propto \left(\alpha_{m,k} + n_{m,.,k}^{-(i,m)}\right)\left(\frac{\beta + n_{.,g,k}^{-(i,m)}}{G\beta + \sum_g [n_{g,k}]^{-(i,m)}}\right)$$

Here $n_{m,.,k}^{-(i,m)} = \sum_{g' \neq g} n_{m,g',k}$ is the total number of scRNA-seq reads allocated for topic *k* for cell *m* and $n_{.,g,k}^{-(i,m)} = \sum_{m' \neq m} n_{m',g,k}$ is the total read counts for gene *g* under topic *k* across all of the *M* cells, without counting the current i^{th} read for cell *m*. The usage of cell-specific $\boldsymbol{\alpha}_m$ in this step ensures cell-type-specific topic guidance during training. The sufficient statistics are updated as follows:

$$n_{m,.,k}^{(-i,m)} = \sum_{m' \neq m}^{M_m} \gamma_{i,m',k}$$

$$n_{.,g,k}^{-(i,m)} = \sum_{m' \neq m, i' \neq i} [x_{i',m'} = g]\ \alpha_{m,k}\, \gamma_{i,m,k}$$

Here in updating the global gene distribution in the second equation, we further make use of the cell-specific $\boldsymbol{\alpha}_m$ to obtain more interpretable results.

Given the sufficient statistics under mean-field, the expected value for $\boldsymbol{\theta}_{\mathrm{m}}$ and $\boldsymbol{\phi}_k$ are:

$$\mathbb{E}_q[\theta_{m,k}] = \frac{\alpha_{m,k} + n_{m,.,k}}{\sum_k \alpha_{m,k} + n_{m,.,k}}$$

$$\mathbb{E}_q[\phi_{\mathrm{g,k}}] = \frac{\beta + n_{.,g,k}}{G\beta + \sum_g n_{.,g,k}}$$

More algorithmic details were described in elsewhere [24, 48].

Usage of multiple topics per cell type

Since it is known that each cell type can exist in different cell states and express different pathways involving different sets of genes, we extended the above model by modelling each cell type via multiple topics. To this end, we evaluated 2, 3, 4, 5, and 10 topics per cell type, with the hyperparameter $\alpha_{m,k}$ for each cell of cell type k set to 0.45, 0.3, 0.22, 0.18, and 0.09, respectively and the remaining topics assigned random probabilities between 0.001 and 0.01. These priors were heuristically chosen based on hyperparameter of value 0.9 for one topic divided by the number of topics per cell type.

Inferring mixed cell-type proportions in bulk transcriptome

To perform deconvolution on bulk RNA-seq data, we used the $G \times K$ genes-by-topics estimates $\widehat{\Phi}$ inferred from the scRNA-seq reference data to infer cell-type mixture (θ_{n}) for the nth bulk RNA-seq profile. This involves alternating between the topic assignment inference $\gamma_{i,n,k}$ and computing the sufficient statistics $n_{n,.,k}^{-(i,n)}$ but fixing the global sufficient statistics $n_{.,g,k}^{-(i,n)}$ as the values fitting the reference scRNA-seq data.

In order to identify *de novo* topics that are not present in the reference scRNA-seq data, we applied standard LDA using the implementation from MixEHR [48] to the sparsified bulk RNA-seq data (see method below on "Learning phenotype-specific gene signatures from bulk RNA-seq data for rationale"). To detect any common or new cell types, we compared the topic distributions from the bulk data with the cell-type-guided topics inferred by GTM-decon from reference scRNA-seq data.

Identifying statistically significant marker genes per cell type

After training, marker genes for each cell type were identified using the permutation test. Specifically, for gene g under topic k, we computed the difference of its topic score $\phi_{g,k}$from the average topic score over the rest of the *K*-1 topics $\phi_{g,k\prime\neq k}$ For example, for 3 cell types, the test statistic for cell type 1 is calculated as $\phi_{g,1} - (\phi_{g,2} + \phi_{g,3})/2$. More generally, for *K* cell types and *T* topics per cell topic, the test statistic for cell type k and gene g is

$$\frac{1}{T}\left(\sum_{t=1}\phi_{k,g}^{\{t\}} - \frac{1}{K-1}\sum_{k'\neq k}\sum_{t=1}\phi_{k',g}^{\{t\}}\right)$$

The significance of the observed statistic is compared against the same statistic calculated from 100,000 permutations. The empirical p-value is computed as fraction of permutations, where the test statistic is greater than the observed value.

Metrics for evaluation of deconvolution accuracy

We evaluated the deconvolution accuracy by comparing the inferred cell-type proportions against the ground truth values for the datasets. Specifically, we simulated bulk RNA-seq data using the scRNA-seq dataset. We randomly sampled a set of cells from the scRNA-seq data and take the sum of their gene expression counts to represent the bulk data. The ground-truth cell-type proportions are the fraction of cells for each cell type. Each artificial bulk transcriptome was constructed from the same subject (from whom the scRNA-seq data were collected). To avoid information bias in the evaluation, we performed leave-one-out-cross-validation (LOOCV) by inferring topics from scRNA-seq data for N - 1 subjects as the training examples and deconvolve the held-subject as the validation example. To evaluate the concordance between the ground truth $\mathbf{y}_m$ and our inferred cell-type mixture $\boldsymbol{\theta}_m$, we computed three common metric scores: 1) Pearson correlation coefficient (PCC), 2) Spearman correlation (SCC), and 3) cross entropy (CE).

Gene set enrichment analysis (GSEA)

GSEA was performed using the `fgsea` package from R, on two different gene sets: a) Gene sets corresponding to specific cell types from CellMarkerDB [27]; b) Gene set corresponding to the HALLMARK pathways from MSigDb using `msigdbr` [49].

Survival analysis

For TCGA datasets, after cell-type proportions are inferred using GTM-decon, survival analysis is performed using the `survival` and `survminer` R packages in order to assess if there is any correlation between survival probability of the cancer subject and cell-type proportions for the cell types. Kaplan-Meier curves were generated for those cell types, where the association is deemed statistically significant at p-value < 0.05 based on log-rank test.

Learning *phenotype*-specific gene signatures from bulk RNA-seq data

To identify phenotype-specific genes, we applied GTM-decon to directly model the bulk RNA-seq data (**Figure 1b**). Bulk RNA-seq data is a dense matrix with most genes having non-zero entries while topic models excel at modeling sparse matrices (e.g., scRNA-seq data). To make the bulk RNA-seq data amenable to our approach, we sparsified the dense matrix by setting all values below the 75th percentile to 0. This cutoff was derived from scRNA-seq datasets, where

on average 25% of genes in a cell have non-zero count values. While sparsifying works well in our applications, as a caveat, we acknowledge that it will lead to information loss. We also experimented with two other approaches to sparsify the matrix: training GTM-decon only using HVG genes in bulk RNA-seq data, training GTM-decon using differentially expressed genes identified between the phenotypes using DESeq2 [50].

To evaluate phenotype prediction accuracy, we randomly split the bulk RNA-seq samples into training and test datasets in 80:20 ratio. The training dataset is guided in a manner similar to GTM-decon for the scRNA-seq datasets. Briefly, for each subject, the hyperparameter(s) for the topic(s) corresponding to the observed phenotype was (were) set to the prior value of 0.9, and the hyperparameters for the rest of the topics were set a small value between 0.01 and 0.1. Since a model of 5 topics per cell type worked well for several scRNA-seq datasets, we opted to model each phenotype by 5 topics as well. For the test subject in the 20% test data, we summed up the inferred phenotype mixture $\theta_{d,k}$ values for all topics corresponding to the same phenotype. The predicted phenotype is the one with the highest summed up topic score.

As a comparison, we also evaluated two other common machine learning methods namely logistic regression and random forest. We used their scikit-learn implementations with the default settings [51].

Identifying cell-type specific differentially expressed genes between phenotypes from bulk RNA-seq data using a nested-guided topic approach

In identifying phenotype-specific differentially expressed genes, we developed a way to leverage the cell-type specific topics inferred from the scRNA-seq reference data. In a nutshell, we extended the GTM-decon framework to fine-tune the cell-type-specific topics based on gene expression from bulk RNA-seq data for different phenotypes (**Figure 1c**). Specifically, we initialized the genes-by-topics matrix Φ_d for *each* phenotype *d*. For example, for D=3 phenotypes in a tissue with K=5 cell types, we will have 15 topics, comprising of 3 sets of cell-type-specific Φ_{d} matrices of 5 columns each. These matrices are guided by the phenotype labels during the variational inference.

Specifically, for subject *n*, suppose his/her phenotype label is *d* $d \in \{1, .., D\}$. The topic hyperparameters $\alpha_{n,d,k}$'s for all the *K* topics corresponding to the phenotype label *d* were set to 0.9, and the $\alpha_{n,d',k}$ for the other phenotypes were set to a small value between 0.01 and 0.1. These values were propagated throughout the variational inference, guiding the phenotypic inference appropriately.

From the learned matrix, the test statistic of a differentially expressed (DE) candidate gene *g* between phenotypes *d* and $d' \neq d$ per cell type *k* were identified by subtracting the $\phi_{g,d,k}$ value for phenotype d from the $\phi_{g,d',k}$ value for phenotype d'. We identify statistically significant genes with a p-value < 0.05 by permutation test calculated from the 100,000 randomly shuffled Φ_d matrices. The p-value is the fraction of times the null statistic from the permutations is greater than the observed test statistic. As a reference, we compared our approach against the DE genes identified from the same samples using DESeq2 [50]. We identified HALLMARK pathways from MSigDb which were enriched for the DE genes by Over Representation Analysis (ORA) using the clusterProfiler package from R [52].

Deconvolving bulk RNA-seq datasets using multiple scRNA-references

Before combining the Φ_r matrices from two different scRNA-seq datasets, we harmonized them first. Specifically, we fixed one of the matrices as the primary tissue reference $\mathbf{\Phi}_p$ and processed the other one as the secondary tissue reference $\mathbf{\Phi}_s$. We aligned the genes of the two matrices. Genes present in $\mathbf{\Phi}_s$ but absent in $\mathbf{\Phi}_p$ were set to 0. We then scaled all entries in $\mathbf{\Phi}_s$ based on the ratio of their averages such that $\mathbf{\Phi}_s$ is at the same numerical scale as $\boldsymbol{\phi}_\mathbf{p}$. Finally, we concatenated the two matrices for deconvolution.

Figure Legends

Figure 1. **GTM-decon overview**. **a**) Inferring cell-type-specific topics from scRNA-seq reference data. For each cell *m*, GTM-decon clamps its hyperparameter α_{m,y_m} of the topic prior distribution based on cell-type label y_m of the cell. As depicted in the probabilistic graphical model (PGM) for the data generative process of each ith read in cell m, the cell-type topic mixture θ_m is sampled from Dirichlet distribution with hyperparameter α_m; cell-type topic $z_{i,m}$ is then sampled from the Categorical distribution with rate fixed to θ_m; given topic assignment $z_{i,m}$ = k, the gene index of the ith read is then sampled from a Categorical distribution with rate set to be $\boldsymbol{\phi}_k$, which is a G-dimensional Dirichlet variable of flat hyperparameter β. To infer the latent variables (i.e., the unshaded nodes in the PGM), we use a collapsed mean-field variational inference algorithm as detailed in the **Methods** section. The inferred genes-by-cell-types $\widehat{\mathbf{\Phi}}$ matrix from the scRNA-seq reference then serves as a set of fixed topic distributions in the inference of cell-type proportion θ_d of a bulk RNA-seq sample *d*. **b**) Phenotype-guided modeling of bulk RNA-seq data. GTM-decon can also use phenotype labels as a guide for topic inference to model bulk transcriptomes in a disease study. For each subject *n*, we set the topic hyperparameter α_{n,y_n} based on the phenotype label y_n of the subject. The inference algorithm is the same as in modeling the scRNA-seq reference data. Given a test subject, the inferred topic mixture $\boldsymbol{\theta}_n$ represents the phenotypic probabilities of the subject. **c**) Nested-guided topic approach for detecting cell-type-specific differentially expressed genes between phenotypes. In this nested design, we treat the phenotype as level 1 and the cell types as level 2. The pre-trained genes-by-cell-types distribution $\widehat{\mathbf{\Phi}}$ learned from panel **a**) are used to initialize the topic distributions for each phenotype in a bulk transcriptome disease study. For example, for 2 phenotypes and 3 cell types, there will be 6 topics. GTM-decon then fine-tunes the combined cell-type-specific topic distribution by running the same algorithm described in panel **b**). The resulting topic distributions reflect the phenotypic influences on cell-type-specific gene expression distributions.

Figure 2. Evaluation of deconvolution accuracy. a) Evaluation of cell-type deconvolution accuracy. We simulated the bulk data by summing up counts per gene for all cells per individual. For unbiased evaluation, we adopted a leave-one-out cross-validation (LOOCV) design. Each deconvolution method was trained on N-1 individuals and deconvolve the bulk transcriptome simulated using the scRNA-seq data from the held-out individual for validation. The inferred cell proportions by each method were evaluated by Spearman correlation and Cross Entropy. The box and the whiskers in each boxplot indicate the 25%-75% quartile and min-max of the evaluation scores over the LOOCV individuals, respectively. **b-d**) Ground truth and Inferred cell-type proportions by each method on Pancreatic, Breast and RA Synovium

Tissues simulated bulk data, respectively. In each heatmap, the rows are individuals and the columns are cell types. The color intensity is proportional to the inferred cell-type proportions.

Figure 3. Cell-type-specific topic inference and deconvolution of pancreatic tissue. a) Gene signatures of cell-type-specific topics in pancreas. We trained GTM-decon on the PI-Segerstolpe scRNA-seq dataset of pancreas tissue. We used 5 topics per cell type, allowing sub cell-type inference. For the inferred genes-by-cell-types matrix Φ, we took the top 20 genes under each topic and visualized their topic distributions in heatmap. Known marker genes from CellMarkerDB and PanglaoDB are indicated on the left, along with the number of statistically significantly different genes in each cell type based on their topic scores (p-value < 0.05; permutation test). **b**) Gene set enrichment analysis (GSEA) of inferred topics based on known marker genes. Cell-type-specific topics for Acinar, Alpha, and Beta were evaluated based on whether the top genes are enriched for the known marker genes under that cell type. In each of the leading-edge plots, genes were ordered in decreasing order from left to right. The green curves indicate the running scores of enrichments. The barcode bars indicate cell-type marker genes. Adjusted p-values based on the GSEA enrichment scores are indicated in each panel. The three large panels display the most significantly enriched topic of among the five topics for each cell type and the 12 small panels display the remaining topics. **c**) and **e**) Deconvolution of bulk RNA-seq samples of 89 human pancreatic islet donors. The GTM-decon models separately trained on the PI_Segerstolpe (i.e., panel c) and PI_Baron scRNA-seq (i.e., panel e) reference datasets were used to deconvolve the 89 bulk transcriptomes. As indicated by the legend, the 89 subjects consist of 51 normal, 15 impaired glucose tolerance, and 12 T2D individuals. In the heatmap, the rows represent subject and the columns represent cell types; the color intensity are proportional to the inferred cell-type proportions. **d**) and **f**) Deconvolved cell-type proportions as a function of Hemoglobin A1c (HbA1c) level. GTM-decon were trained on PI_Segerstolpe (i.e., panel d) and PI_Baron scRNA-seq (i.e., panel f) reference datasets. Each of the 10 panels displays a scatter plot of inferred cell-type proportion (y-axis) and HbA1c level (x-axis). The color legend indicates the 3 phenotypes. The heatmap on the right shows the deconvolved proportion of 3 most indicative cell types with subjects (rows) ordered on the basis of inferred beta cell-type proportions.

Figure 4. Deconvolution of bulk RNA-seq samples for pancreatic cancer from TCGA-PAAD. **a**) Inferred cell-type proportions of TCGA-PAAD tumor samples. GTM-decon was trained separately on pancreatic and immune scRNA-seq datasets. The trained GTM-decon models were then used to deconvolve the 174 TCGA-PAAD bulk RNA-seq profiles. In addition, we also ran unguided topic model (i.e., LDA) on the TCGA-PAAD bulk RNA-seq profiles directly to detect *de novo* topics that are not present in scRNA-seq reference. The above heatmap visualizes the combined deconvolution results based on the 10 pancreatic cell types, 4 immune cell types, and 10 *de novo* topics (i.e., columns). Each of the 174 rows represents a subject. Five types of demographic or clinical phenotypes were shown in the legend to aid result interpretation. These include days to death, cancer stage, sex, misclassified status, and cancer subtype. The regions in highlighted boxes are described in the main text. **b**) Association between cancer stages and inferred cell-type proportions. The box in boxplots in each panel display the 25%-75% quartiles of inferred cell-type proportions Ductal or *de novo* Topic 5 (y-axis) for each cancer stage (x-axis). Statistically significant differences in proportions across different clinical stages were indicated with the corresponding p-value based on Kruskal-Wallis test. Significant individual pairwise comparison between cancer stages based on Wilcox Rank Sum test were also indicated by the p-values in each panel. **c**) Survival analysis using inferred

cell-type proportions. The 174 subjects were divided into two groups based on K*-means cluster with K* set to 2 (not to be confused with the K cell types or topics). Kaplan-Meier curves for cell types and the *de novo* topics that resulted in significant differences in terms of their hazard ratios were displayed. The curve and shaded area represent the mean and standard deviation of the cell-type proportions in the two groups, respectively. The number of subjects for each cluster were indicated below the Kaplan-Meier plot.

Figure 5. Cell-type-specific inference and deconvolution of breast transcriptomes using HVG. a) Gene signatures of cell-type-specific topics in breast tissue. We trained GTM-decon on a scRNA-seq reference of normal breast tissue [36] using 5 topics per cell type, on the set of highly variable genes. Same as Figure 3, the heatmap visualizes the top 20 genes per topic for the 5 main cell types in normal breast tissue (i.e., 25 topics in total for the 5 cell types). The heatmap intensity is proportional to the gene topic probabilities. The known marker genes based on PanglaoDB and CellMarkerDB were indicated on the left color bar. **b**) Heatmap of deconvolved cell-type proportions for the main cell types for GTEx (normal samples) and TCGA-BRCA (cancer samples). The rows represent samples and columns represent the reference cell types. The heat intensity is proportional to the inferred cell-type proportions. Agglomerative clustering was applied to both the rows and columns to cluster samples and cell types respectively using Euclidean distance metric and complete linkage clustering. **c**) Comparison of inferred cell type or *de novo* topic proportions across breast cancer types. Each boxplot displays the distribution of cell-type proportions or topic proportions (y-axis) for the breast tumor samples separated by their cancer classes based on Prediction Analysis of Microarray 50 (PAM50) (x-axis). In each panel, p-value based on Kruskal-Wallis test was indicated on the top left corner and p-values based on Wilcox Rank Sum test for the significant pairwise comparison between PAM50 classes were also indicated. **d**) Comprehensive visualization of deconvolution results for the 1212 breast tumor samples from TCGA. The deconvolution results were consolidated from the inferred topics of three topic models, namely GTM-decon trained on breast scRNA-seq reference (i.e., the same model from panel a), GTM-decon trained on immune scRNA-seq reference, and unguided 10-topic LDA directly trained on the TCGA-BRCA. The three sets of topics were indicated by blue, green, and grey colors, respectively on the top color bar. The row represents the 1212 tumor samples with 6 types of clinical phenotypes annotation on the left color bar to aid result interpretation. The regions in the highlighted boxes were discussed in the main text. Deconvolution results using all genes were shown in **Figure S6**.

Figure 6. Deconvolving breast cancer bulk transcriptome based on single-cell transcriptomes from breast tumors. **a**) Inferred cell-type proportions matched the cancer subtypes. We trained GTM-decon on the scRNA-seq breast cancer dataset using the 8 cell types as a guide. Each boxplot shows the distribution of the cell-type proportion for the samples separated by the cancer subtypes as defined by the PAM50 system. P-values based on the Kruskal-Wallis were shown. The subtype that exhibits the highest cell-type proportion in each panel was circled. **b**) Detailed visualization of the deconvolved proportions. Similar to Figure 5, the consolidated heatmap illustrates the cell type and *de novo* topic proportions from GTM-decon and unguided LDA separately trained on the scRNA-seq breast cancer data and TCGA-BRCA data, respectively. These two sets of topics were colored in grey and blue, respectively. ell-typeRows represents the 1212 tumor samples and columns represents the cell types and *de novo* topics. The clinical phenotypes were annotated in the left color bar to help interpret the topics. Abbreviationsell-type: CAFs – cancer associated fibroblasts; myCAF – myofibroblast;

iCAF – inflammatory CAF; DCs – dendritic cells; NK – natural killer; NKT – natural killer T cells; PVL – perivascular-like.

Figure 7. Cross-reference and cross-species deconvolution. **a**) Deconvolution of PI-Segerstolpe human pancreas samples using different scRNA-seq reference data. The four Hinton plots from left to right display the ground truth cell-type proportions, the deconvolved proportions using the same reference dataset as target (i.e., PI-Segerstolpe), the deconvolved proportions using PI_Baron scRNA-seq data as the reference, the deconvolved proportions using mouse pancreas scRNA-seq data as the reference. Row are subjects and columns are cell types. **b**) Deconvolution of TCGA-PAAD dataset using human and mouse pancreatic datasets as reference. The samples (rows) were clustered by agglomerative clustering based on the GTM-decon model trained on the mouse reference (i.e., left heatmap). The same row order was applied to the heatmap based on human reference on the right. PSC in human dataset represents pancreatic stellate cells, which are represented as two different subtypes in mouse dataset. Four clinical phenotypes were shown as color bar on the left of each heatmap. Highlighted region for the alpha cell type was described in the main text.

Figure 8. Cancer-subtype-guided topic modeling of bulk RNA-seq data of breast cancer. **a**) Predicted top genes for the phenotype-guided topics for Basal and ER+ subtypes. GTM-decon was trained on the TCGA-BRCA bulk RNA-seq data with the Basal and ER+ cancer subtypes as the guide. Five topics were used per subtype and therefore 10 topics in total. The heatmap illustrates the topic probabilities of the top 20 genes from each topic. As a comparison, the genes were also labelled as up or down-regulated if they were deemed differentially expressed by the DESeq2 analysis. **b**) Classification of Basal and ER+ subtypes based on phenotype-guided topic scores. The 5 topics for the same subtype were summed to obtain the overall score for Basal and ER+ subtype. Subjects in the rows were sorted by their Basal topic scores. **c**) GSEA analysis of the Basal and ER+ subtype topics. Significantly enriched MSigDb HALLMARK pathways were identified for each topic and displayed as barplots. The heights of the bar indicate the -log10 adjusted p-values and the colors indicate enriched pathways. **d**) Predicted top genes for the phenotype-guided topics for histological subtypes. Same as in panel **a**) but for Ductal and Lobular subtypes. **e)** Classification of histological subtypes. Same as in panel **b**) but for Ductal and Lobular subtypes. **f**) Evaluation of the subtype classification accuracy on the test breast tumor samples. We trained the phenotype-guided GTM-decon separately on 80% TCGA-BRCA tumor samples using Basal/ER+ and Histological types as the guides and evaluated its phenotype prediction accuracy on the 20% held samples. As a comparison, we also trained and evaluated Logistic Regression and Random Forest on the same training and test split, respectively. The classification accuracy on the test set by each method were displayed in the barplots.

Figure 9. Identification of cell-type-specific differentially expressed genes from bulk RNA-seq data for Basal vs. ER+ subtypes. **a**) Top cell-type-specific gene signatures for Basal and ER+. GTM-decon was pretrained on a scRNA-seq reference dataset from normal breast tissue to infer the expression distribution of 5 cell types, namely Basal, Basal myoepithelial, Luminal 1-1, Luminal 1-2, and Luminal 2. The resulting genes-by-cell-types estimates were then used as the initial topic distributions for another GTM-decon, which is guided by the Basal and ER+ cancer subtypes in modeling the TCGA-BRCA data. This led to a 10-topic distribution, each of which was specifically tailored for a combination of cell type and cancer subtype. The heatmap displays the probabilities of the top 20 genes for each topic. The left half displays the cell-type-

specific topic distribution for Basal and the right half for ER+. **b**) Predicted differentially expressed (DE) genes for each cell type between Basal and ER+. The top DE genes for Basal in contrast to ER+ were identified by subtracting the gene topic scores for ER+ from the gene topic score for Basal under the same cell type. The resulting DE scores were shown in the top half of the heatmap. The bottom half displays the DE scores of the top genes for ER+ in contrast to Basal. The pairwise Wilcox signed-rank tests were performed to compare the gene topic scores across all genes between the two subtypes for the same cell type. All tests yielded p-values lower than 2.2e-16. **c**) Classification of Basal and ER+ based on the phenotype probabilities. As a validation for our nested phenotype-cell-type guided approach, we evaluated the classification accuracy on the 160 held-out breast tumor samples. For each subtype, we summed the cell-type-specific topic probabilities from bottom heatmap for each sample to obtain the phenotype scores, which are shown in the top heatmap. **d**) Comparison of DE genes detected by our approach and by DESeq2. DESeq2 was applied to the bulk RNA-seq gene expression data to compare gene expression between ER+ and Basal samples. In total, 6815 DE genes were deemed significant by adjusted p-value < 0.05 (Wald test) with 2952 up-regulated and 3863 down-regulated genes in ER+ relative to Basal. The grey bar and the heatmap on the left display the -log adjusted p-value for all of the upregulated genes (top half) and the down-regulated genes (bottom half). Genes were ordered in decreasing order of the absolute test statistic for each half. The corresponding $\log_2$ Fold-Change of ER+ over Basal was also shown as heatmap. The heatmap on the right displays the change of gene topic score from Basal subtype to ER+ subtype. **e**) Cell-type-specific DE genes identified by nested-guided topic approach. The top and bottom part of the heatmap displays the topic scores for the up-regulated and down-regulated genes in Basal relative to ER+, respectively (p-value < 0.05; permutation test). Genes that were also detected by DESeq2 were labeled in the color bar. **f)** ORA was applied to the differential topic scores of up-regulated and down-regulated genes in ER+ relative to Basal. MSigDb HALLMARK pathway gene sets was used in ORA. The -log p-values for the significant pathways were shown in the bar plot.

Acknowledgements

Y.L. is supported by Canada Research Chair (Tier 2) in Machine Learning for Genomics and Healthcare, New Frontier Research Fund -- Exploration (NFRFE-2019-00980), and Canada First Research Excellence Fund Healthy Brains for Healthy Life (HBHL) initiative New Investigator start-up award (G249591). We thank Yixuan Li for helping troubleshoot the GTM-decon software at its later stage.

Author contributions

Y.L. conceived of the study. S.S. and Y.L. analyzed and interpreted the data, wrote the manuscript, and wrote the code for GTM-decon. M.H. ran the experiments on benchmarking the deconvolutional accuracy. Y.L. supervised the project. All authors approved the final manuscript.

Competing interest declaration

They declare no competing interests.

Data and software availability

The datasets analyzed during the current study were obtained from publicly available repositories or data portals. The Human pancreatic islet datasets from Segerstolpe et. al. used in this study is available in the ArrayExpress database under the accession code E-MTAB-5061. The Human and Mouse pancreatic islet datasets from Baron et. al. used in this study are available in the GEO database under the accession code GSE84133. The normal breast tissue dataset is available in the GEO database under the accession code GSE113197. The cancerous breast tissue dataset is available in the GEO database under the accession code GSE176078. The RA synovium dataset is available in the ImmPort database under the accession code SDY998. The bulk RNA-seq dataset for normal and type 2 diabetes patients is available in the GEO database under the accession code GSE50244. The bulk RNA-seq datasets for pancreatic cancer (PAAD) and breast cancer (BRCA) datasets from TCGA can be downloaded from the GDC data portal (https://portal.gdc.cancer.gov/).

GTM-decon source codes have been deposited at the GitHub repository (https://github.com/li-lab-mcgill/gtm-decon).

Supplementary Figure Legends

Figure S1. Deconvolution accuracy for different topics per cell type. Each GTM-decon model was trained by modelling a cell type using K number of topics, with K varying from 1 to 5, using all genes. We evaluated the deconvolution accuracy of these models on simulated bulk data from three datasets: Pancreatic (Segerstolpe), Breast (Normal), and RA Synovium. We simulated the bulk data by summing up counts per gene for all cells per individual. For unbiased evaluation, we adopted a leave-one-out cross-validation (LOOCV) design. Each model was trained on N-1 individuals and used to deconvolve the simulated bulk transcriptome from the held-out individual for validation. The inferred cell proportions by each model were evaluated by Pearson correlation coefficient (PCC) (top row) and Spearman correlation coefficient (PCC) (bottom row). The box and the whiskers in each boxplot indicate the 25%-75% quartile and min-max of the evaluation scores over the LOOCV individuals, respectively.

Figure S2. Deconvolution accuracy for different gene sets. Each GTM-decon model was trained by modelling a cell type using different gene selection strategies and their transformations i) all genes – corresponding to raw gene counts ii) all_normr – corresponding to normalized of raw genes by total counts in a cell iii) all_log1p_normr – corresponding to log-transformed form of all_normr iv) hvg – corresponding to raw counts of highly variable genes v) hvg_log1p_normr – corresponding to log-transformed form of normalized hvg. Each of these models was further trained based on K number of topics, with K varying from 1 to 5. We evaluated the deconvolution accuracy of these models on simulated bulk data from two datasets: Pancreatic (Segerstolpe) and Breast (Normal). We simulated the bulk RNA-seq (bulkRS) data by randomly sampling cells from the held-out set of scRNA-seq data (20%) and summing up the values for each gene from all cells belonging to an individual. The cell proportions inferred for each individual was compared with the ground truth values and evaluated by Pearson correlation. The box and the whiskers in each boxplot indicate the 25%-75% quartile and min-max of the evaluation scores over the individuals, respectively.

Figure S3. Evaluating hyperparameter α based on the cell-type-specific topic enrichment for the known marker genes. The pancreatic dataset from Segerstolpe was trained using different cell-type-guided topic priors with hyperparameter α set: 0.9, 5, 10, 20, 50, 100**.** The effect of these hyperparameter values in recovering the specific cell type on which the topic was guided was evaluated using gene set enrichment analysis (GSEA). Briefly, cell-type-specific topics for Acinar, Alpha, and Beta cell types were evaluated based on whether the top genes are enriched for the known marker genes from CellMarkerDB under that cell type. In each of the leading-edge plots, genes were ordered in decreasing order from left to right. The green curves indicate the running scores of enrichments. The barcode bars indicate cell-type marker genes. Leading plots are generated for the three cell types for different priors assigned to hyperparameter α: 0.9, 5, 10, 20, 50, 100. The closer to the y-axis the leading plots are, the better the prior. Adjusted p-values based on the GSEA enrichment scores are indicated for the best hyperparameter.

Figure S4. Evaluating the number of topics per cell-type based on GSEA of known marker genes. Cell-type-specific topics for Acinar, Alpha, and Beta cell types from Pancreatic dataset (PI_Segerstolpe) were evaluated based on whether the top genes are enriched for the known marker genes (from CellMarkerDB) under that cell type using gene set enrichment analysis. Each model was trained with varying numbers of topics per cell type (from 1 to 5). The plot shows the average (-log10 adjusted p-value) for the cell type of interest for the different topics. An overall value is calculated by summing up the values for the three cell types.

Figure S5. Evaluating topic confidence in terms of number of cells per cell type. We evaluated topic confidence by correlating the topic confidence scores with the number of cells for each cell type available for training. The topic confidence scores were calculated by the first taking the maximum probability from the G-by-1 column vector Φ_k for each cell type $k \in \{1, \ldots, K\}$ and then averaging over all K cell types. The resulting topic confidence scores were plotted as a function of number of cells per cell type. Each panel correspond to the number of topics per cell type (varying from 1 to 5) for two pancreatic datasets, Segerstolpe and Baron, generated using different approaches, Smart-seq2 and Drop-seq, respectively.

Figure S6. Deconvolution of breast cancer transcriptomes from TCGA using *all genes*. Visualization of deconvolution results for the 1212 breast tumor samples from TCGA. The deconvolution results were consolidated from the inferred topics of three topic models, namely GTM-decon trained on breast scRNA-seq reference using all genes with 5 topics per cell type, GTM-decon trained on immune scRNA-seq reference using all genes with 5 topics per cell type, and unguided 10-topic LDA directly trained on the TCGA-BRCA. The three sets of topics were indicated by blue, green, and grey colors, respectively on the top color bar. The row represents the 1212 tumor samples with 6 types of clinical phenotypes annotation on the left color bar to aid result interpretation.

Figure S7. Identification of cell-type-specific differentially expressed genes from bulk RNA-seq data comparing between ductal and lobular breast carcinoma subtypes. Bulk RNA-seq samples of ductal and lobular subtypes of breast carcinoma from TCGA were analyzed using the same approach described in **Figure 9** to identify cell-type-specific differentially expressed (DE) genes corresponding to these two subtypes. **a)** Top cell type specific genes in each subtype. Visualization of top cell-type-specific gene signatures for Ductal and Lobular carcinoma in terms of gene-by-cell-type proportions for each subtype. **b)** Top DE genes in each subtype. Visualization of top predicted differentially expressed (DE) genes for

each cell type between Ductal and Lobular carcinoma in terms of gene-by-cell-type proportions for each subtype. Paired Wilcoxon-ranked tests for same cell types in the two subtypes reveals significant differences in gene-cell-type probabilities for all cell types (p-value < 2.2×10^{-16} for all comparisons). **c)** Classification of subtypes based on phenotype probabilities. Visualization of phenotype probabilities inferred for ductal and lobular subtypes for the held-out set. The plot is ordered based on decreasing values of inferred phenotype probabilities for ductal subtype (rows), with the phenotype labels indicated on the left color bar. **d)** Comparison with DE genes from DESeq2. Comparison of gene-topic scores inferred by our approach for DE genes detected by DESeq2 in a Lobular versus Ductal comparison. On the left, the $\log_2$ fold change in expression of upregulated genes (top half) and downregulated genes (bottom half) identified by DESeq2 are arranged in decreasing order of the adjusted p-values. On the right, a heatmap of differences in gene-by-topic proportions for the same cell type, inferred by our method for a lobular versus ductal comparison, is shown for corresponding genes. **e)** DE genes identified by our method. Heatmap of 284 statistically significant DE genes (p-value < 0.05) identified by our method (rows), depicting difference in gene-topic-scores for the same cell types. The top cluster shows the upregulated genes in ductal subtype while the bottom cluster shows the downregulated genes. DE genes corresponding to those identified by DESeq2 are indicated in the left color bar. **f)** ORA analysis of differences in gene-topic proportions. Statistically-significantly enriched HALLMARK pathways identified in an ORA analysis based on differences in gene-topic-scores per cell type for ductal versus lobular comparison are indicated, for both upregulated and downregulated genes. Their significance level is indicated by the -log (adjusted p-value) on the y-axis, for the different cell types on the x-axis.

References

1. Cho J-H, Kim J-W, Shin J-A, Shin J, Yoon K-H: **β-cell mass in people with type 2 diabetes.** *Journal of diabetes investigation* 2011, **2:**6-17.
2. Sasaki H, Saisho Y, Inaishi J, Watanabe Y, Tsuchiya T, Makio M, Sato M, Nishikawa M, Kitago M, Yamada T, Itoh H: **Reduced beta cell number rather than size is a major contributor to beta cell loss in type 2 diabetes.** *Diabetologia* 2021, **64:**1816-1821.
3. van Galen P, Hovestadt V, Wadsworth Ii MH, Hughes TK, Griffin GK, Battaglia S, Verga JA, Stephansky J, Pastika TJ, Lombardi Story J, et al: **Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity.** *Cell* 2019, **176:**1265-1281.e1224.
4. Chen G, Ning B, Shi T: **Single-Cell RNA-Seq Technologies and Related Computational Data Analysis.** *Frontiers in genetics* 2019, **10:**317-317.
5. Haque A, Engel J, Teichmann SA, Lönnberg T: **A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications.** *Genome medicine* 2017, **9:**75-75.
6. Hwang B, Lee JH, Bang D: **Single-cell RNA sequencing technologies and bioinformatics pipelines.** *Experimental & molecular medicine* 2018, **50:**1-14.
7. Van den Berge K, Hembach KM, Soneson C, Tiberi S, Clement L, Love MI, Patro R, Robinson MD: **RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis.** *Annual Review of Biomedical Data Science* 2019, **2:**139-173.
8. Barkley D, Rao A, Pour M, França GS, Yanai I: **Cancer cell states and emergent properties of the dynamic tumor system.** *Genome research* 2021, **31:**1719-1727.
9. Davis-Marcisak EF, Deshpande A, Stein-O'Brien GL, Ho WJ, Laheru D, Jaffee EM, Fertig EJ, Kagohara LT: **From bench to bedside: Single-cell analysis for cancer immunotherapy.** *Cancer cell* 2021, **39:**1062-1080.

10. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A: **A survey of best practices for RNA-seq data analysis.** *Genome biology* 2016, **17:**13-13.
11. Clough E, Barrett T: **The Gene Expression Omnibus Database.** *Methods in molecular biology (Clifton, NJ)* 2016, **1418:**93-110.
12. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic acids research* 2002, **30:**207-210.
13. Consortium GT: **The Genotype-Tissue Expression (GTEx) project.** *Nature genetics* 2013, **45:**580-585.
14. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al: **Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer.** *Cell* 2018, **173:**291-304.e296.
15. Shen-Orr SS, Gaujoux R: **Computational deconvolution: extracting cell type-specific information from heterogeneous samples.** *Current opinion in immunology* 2013, **25:**571-578.
16. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA: **Robust enumeration of cell subsets from tissue expression profiles.** *Nature methods* 2015, **12:**453-457.
17. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, Ryu JH, Wagner BK, Shen-Orr SS, Klein AM, et al: **A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure.** *Cell systems* 2016, **3:**346-360.e344.
18. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, Khodadoust MS, Esfahani MS, Luca BA, Steiner D, et al: **Determining cell type abundance and expression from bulk tissues with digital cytometry.** *Nature biotechnology* 2019, **37:**773-782.
19. Wang X, Park J, Susztak K, Zhang NR, Li M: **Bulk tissue cell type deconvolution with multi-subject single-cell expression reference.** *Nature communications* 2019, **10:**380-380.
20. Racle J, Gfeller D: **EPIC: A Tool to Estimate the Proportions of Different Cell Types from Bulk Gene Expression Data.** *Methods in molecular biology (Clifton, NJ)* 2020, **2120:**233-248.
21. Jew B, Alvarez M, Rahmani E, Miao Z, Ko A, Garske KM, Sul JH, Pietiläinen KH, Pajukanta P, Halperin E: **Accurate estimation of cell composition in bulk expression through robust integration of single-cell information.** *Nature communications* 2020, **11:**1971-1971.
22. Menden K, Marouf M, Oller S, Dalmia A, Magruder DS, Kloiber K, Heutink P, Bonn S: **Deep learning-based cell composition analysis from tissue expression profiles.** *Science advances* 2020, **6:**eaba2619-eaba2619.
23. Chu T, Wang Z, Pe'er D, Danko CG: **Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology.** *Nature Cancer* 2022, **3:**505-517.
24. Ahuja Y, Zou Y, Verma A, Buckeridge D, Li Y: **MixEHR-Guided: A guided multi-modal topic modeling approach for large-scale automatic phenotyping using the electronic health record.** *Journal of biomedical informatics* 2022, **134:**104190-104190.
25. Jew B, Alvarez M, Rahmani E, Miao Z, Ko A, Garske KM, Sul JH, Pietiläinen KH, Pajukanta P, Halperin E: **Accurate estimation of cell composition in bulk expression through robust integration of single-cell information.** *Nature Communications* 2020, **11**.
26. Segerstolpe Å, Palasantza A, Eliasson P, Andersson E-M, Andréasson A-C, Sun X, Picelli S, Sabirsh A, Clausen M, Bjursell MK, et al: **Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes.** *Cell metabolism* 2016, **24:**593-607.
27. Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, Luo T, Xu L, Liao G, Yan M, et al: **CellMarker: a manually curated resource of cell markers in human and mouse.** *Nucleic Acids Research* 2019, **47:**D721-D728.

28. Franzén O, Gan L-M, Björkegren JLM: **PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data.** *Database* 2019, **2019**.
29. Anderson NM, Simon MC: **The tumor microenvironment.** *Current biology : CB* 2020, **30:**R921-R925.
30. Lei X, Lei Y, Li J-K, Du W-X, Li R-G, Yang J, Li J, Li F, Tan H-B: **Immune cells within the tumor microenvironment: Biological functions and roles in cancer immunotherapy.** *Cancer letters* 2020, **470:**126-133.
31. Xu Y, Liu J, Nipper M, Wang P: **Ductal vs. acinar? Recent insights into identifying cell lineage of pancreatic ductal adenocarcinoma.** *Annals of pancreatic cancer* 2019, **2**.
32. Peran I, Madhavan S, Byers SW, McCoy MD: **Curation of the Pancreatic Ductal Adenocarcinoma Subset of the Cancer Genome Atlas Is Essential for Accurate Conclusions about Survival-Related Molecular Mechanisms.** *Clinical cancer research : an official journal of the American Association for Cancer Research* 2018, **24:**3813-3819.
33. Di Domenico A, Pipinikas CP, Maire RS, Bräutigam K, Simillion C, Dettmer MS, Vassella E, Thirlwell C, Perren A, Marinoni I: **Epigenetic landscape of pancreatic neuroendocrine tumours reveals distinct cells of origin and means of tumour progression.** *Communications biology* 2020, **3:**740-740.
34. Alkasalias T, Moyano-Galceran L, Arsenian-Henriksson M, Lehti K: **Fibroblasts in the Tumor Microenvironment: Shield or Spear?** *International journal of molecular sciences* 2018, **19**.
35. Starzyńska T, Karczmarski J, Paziewska A, Kulecka M, Kuśnierz K, Żeber-Lubecka N, Ambrożkiewicz F, Mikula M, Kos-Kudła B, Ostrowski J: **Differences between Well-Differentiated Neuroendocrine Tumors and Ductal Adenocarcinomas of the Pancreas Assessed by Multi-Omics Profiling.** *International journal of molecular sciences* 2020, **21**.
36. Nguyen QH, Pervolarakis N, Blake K, Ma D, Davis RT, James N, Phung AT, Willey E, Kumar R, Jabart E, et al: **Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity.** *Nature communications* 2018, **9:**2028-2028.
37. Bertucci F, Finetti P, Birnbaum D: **Basal breast cancer: a complex and deadly molecular subtype.** *Current molecular medicine* 2012, **12:**96-110.
38. Dai X, Cheng H, Bai Z, Li J: **Breast Cancer Cell Line Classification and Its Relevance with Breast Tumor Subtyping.** *Journal of Cancer* 2017, **8:**3131-3141.
39. Wu SZ, Al-Eryani G, Roden DL, Junankar S, Harvey K, Andersson A, Thennavan A, Wang C, Torpy JR, Bartonicek N, et al: **A single-cell and spatially resolved atlas of human breast cancers.** *Nature genetics* 2021, **53:**1334-1347.
40. Han X, Zhou Z, Fei L, Sun H, Wang R, Chen Y, Chen H, Wang J, Tang H, Ge W, et al: **Construction of a human cell landscape at single-cell level.** *Nature* 2020, **581:**303-309.
41. Li Y, Nair P, Lu XH, Wen Z, Wang Y, Dehaghi AAK, Miao Y, Liu W, Ordog T, Biernacka JM, et al: **Inferring multimodal latent topics from electronic health records.** *Nature Communications* 2020, **11**.
42. Ahuja Y, Zhou D, He Z, Sun J, Castro VM, Gainer V, Murphy SN, Hong C, Cai T: **sureLDA: A multidisease automated phenotyping method for the electronic health record.** *Journal of the American Medical Informatics Association* 2020, **27:**1235-1243.
43. Nguyen D, Liakata M, DeDeo S, Eisenstein J, Mimno D, Tromble R, Winters J: **How We Do Things With Words: Analyzing Text as Social and Cultural Data.** *Frontiers in artificial intelligence* 2020, **3:**62-62.
44. Wolf FA, Angerer P, Theis FJ: **SCANPY: large-scale single-cell gene expression data analysis.** *Genome biology* 2018, **19:**15-15.
45. Blei DM, Ng AY, Jordan MI: **Latent dirichlet allocation.** *The Journal of Machine Learning Research* 2003, **3:**993-1022.

46. Griffiths TL, Steyvers M: **Finding scientific topics.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101 Suppl 1:**5228-5235.
47. Teh YW, Newman D, Welling M: **A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation.** *Advances in neural …* 2006.
48. Li Y, Nair P, Lu XH, Wen Z, Wang Y, Dehaghi AAK, Miao Y, Liu W, Ordog T, Biernacka JM, et al: **Inferring multimodal latent topics from electronic health records.** *Nature communications* 2020, **11:**2536-2536.
49. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P: **The Molecular Signatures Database (MSigDB) hallmark gene set collection.** *Cell systems* 2015, **1:**417-425.
50. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome biology* 2014, **15:**550-550.
51. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Müller A, Nothman J, Louppe G, et al: **Scikit-learn: Machine Learning in Python.** 2012.

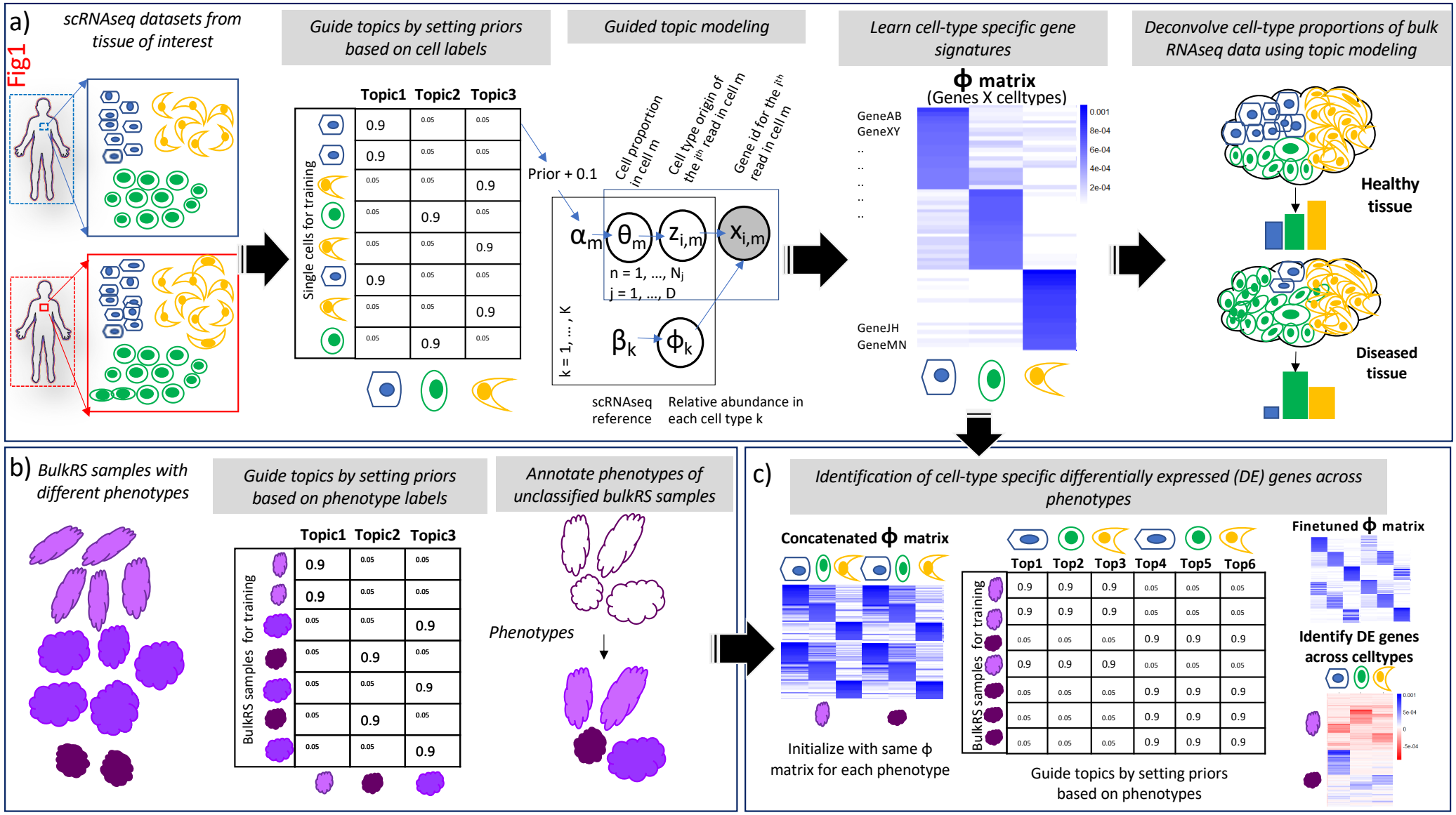
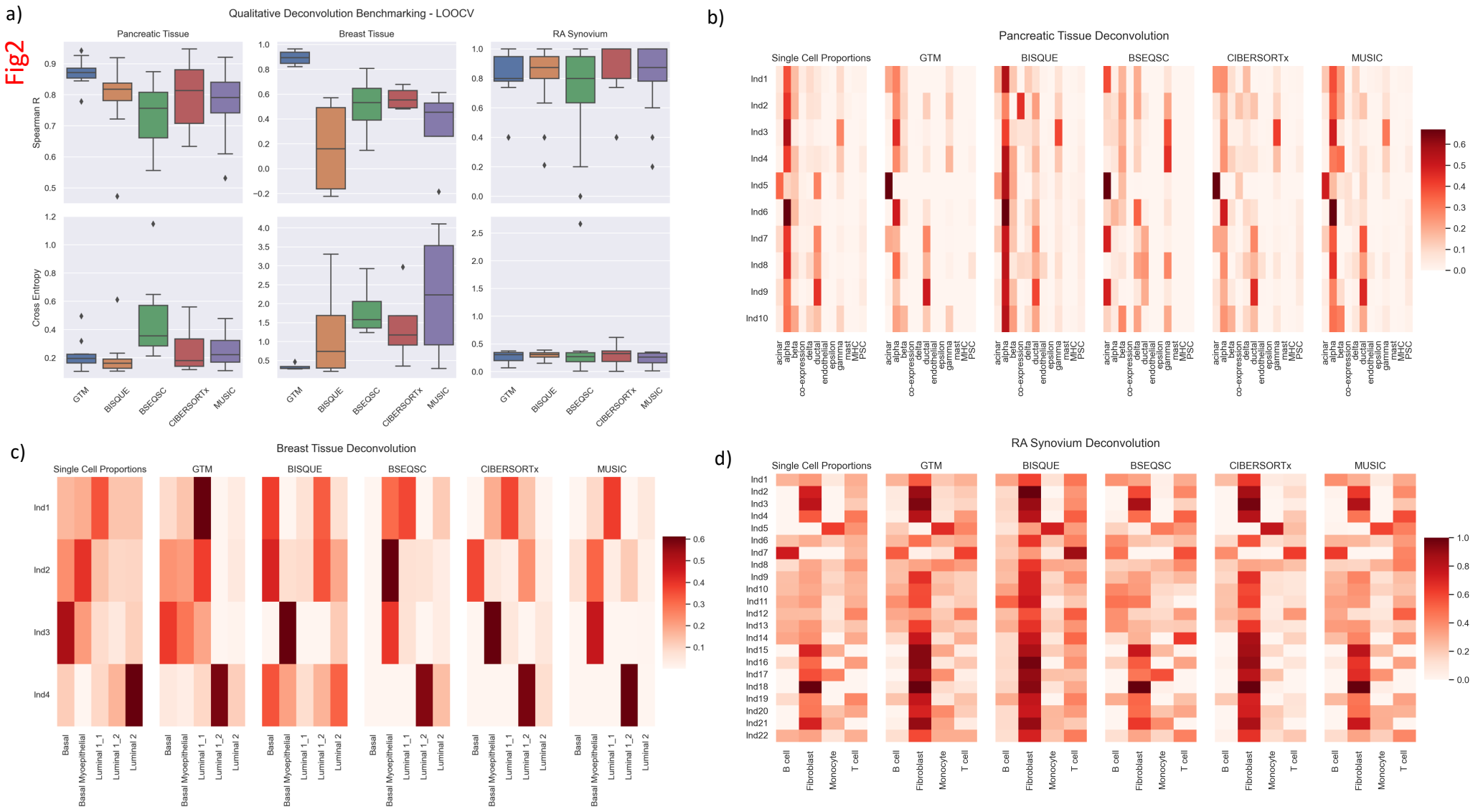


Fig1
a)
scRNAseq datasets from tissue of interest
Guide topics by setting priors based on cell labels
Guided topic modeling
Learn cell-type specific gene signatures
Deconvolve cell-type proportions of bulk RNAseq data using topic modeling
Topic1 Topic2 Topic3
Single cells for training
Prior + 0.1
Cell proportion in cell m
Cell type origin of the ith read in cell m
Gene id for the ith read in cell m
n = 1, ..., N_j
j = 1, ..., D
k = 1, ... , K
scRNAseq reference
Relative abundance in each cell type k
ϕ matrix
(Genes X celltypes)
GeneAB
GeneXY
GeneJH
GeneMN
Healthy tissue
Diseased tissue
b)
BulkRS samples with different phenotypes
Guide topics by setting priors based on phenotype labels
Annotate phenotypes of unclassified bulkRS samples
Topic1 Topic2 Topic3
BulkRS samples for training
Phenotypes
c)
Identification of cell-type specific differentially expressed (DE) genes across phenotypes
Concatenated ϕ matrix
Initialize with same ϕ matrix for each phenotype
Top1 Top2 Top3 Top4 Top5 Top6
BulkRS samples for training
Guide topics by setting priors based on phenotypes
Finetuned ϕ matrix
Identify DE genes across celltypes

Fig2


a)
Qualitative Deconvolution Benchmarking - LOOCV
Pancreatic Tissue
Breast Tissue
RA Synovium
Spearman R
Cross Entropy
GTM
BISQUE
BSEQSC
CIBERSORTx
MUSIC
b)
Pancreatic Tissue Deconvolution
Single Cell Proportions
GTM
BISQUE
BSEQSC
CIBERSORTx
MUSIC
Ind1
Ind2
Ind3
Ind4
Ind5
Ind6
Ind7
Ind8
Ind9
Ind10
acinar
alpha
beta
co-expression
delta
ductal
endothelial
epsilon
gamma
mast
MHC
PSC
c)
Breast Tissue Deconvolution
Single Cell Proportions
GTM
BISQUE
BSEQSC
CIBERSORTx
MUSIC
Ind1
Ind2
Ind3
Ind4
Basal
Basal Myoepithelial
Luminal 1_1
Luminal 1_2
Luminal 2
d)
RA Synovium Deconvolution
Single Cell Proportions
GTM
BISQUE
BSEQSC
CIBERSORTx
MUSIC
Ind1
Ind2
Ind3
Ind4
Ind5
Ind6
Ind7
Ind8
Ind9
Ind10
Ind11
Ind12
Ind13
Ind14
Ind15
Ind16
Ind17
Ind18
Ind19
Ind20
Ind21
Ind22
B cell
Fibroblast
Monocyte
T cell

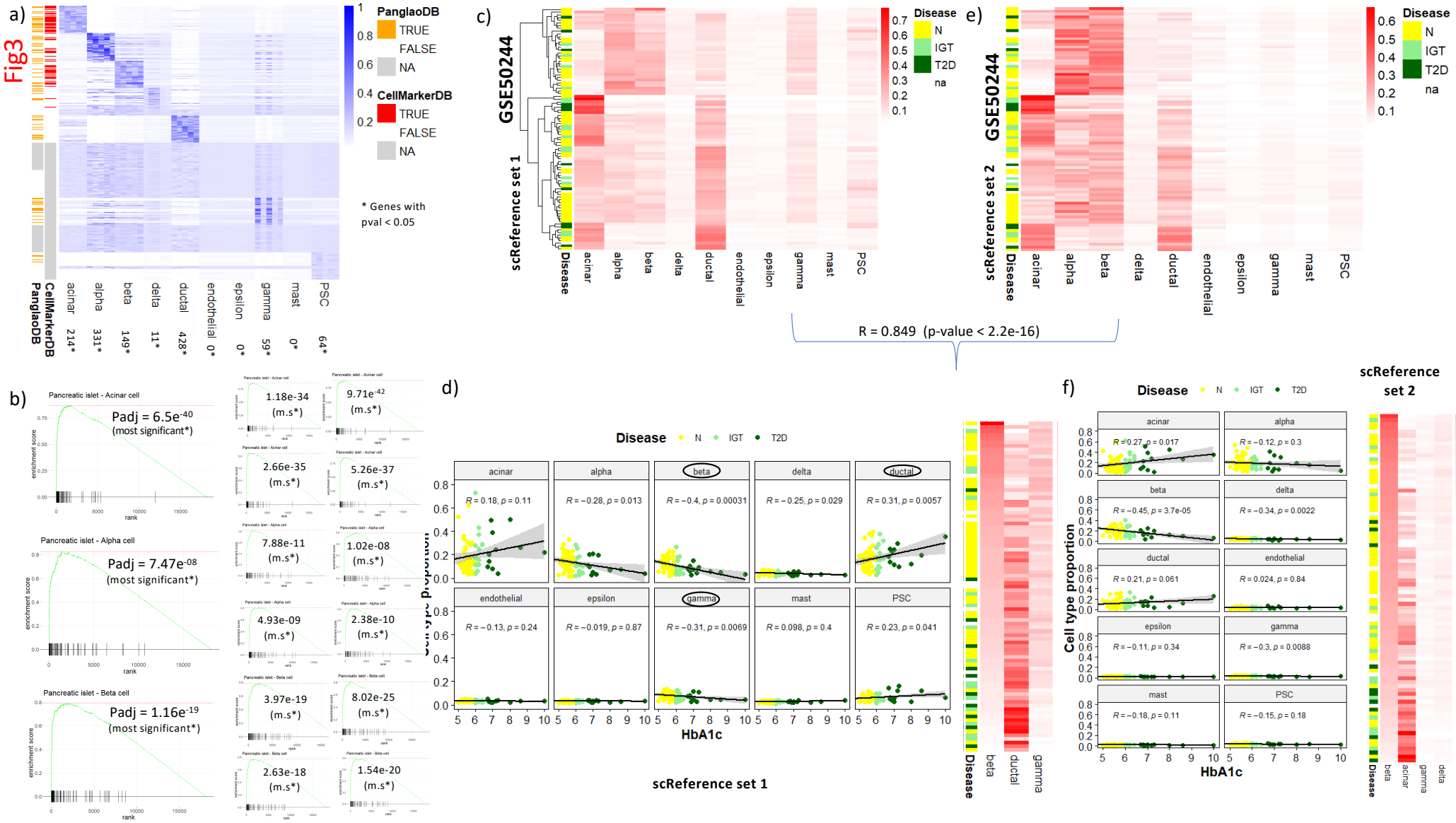
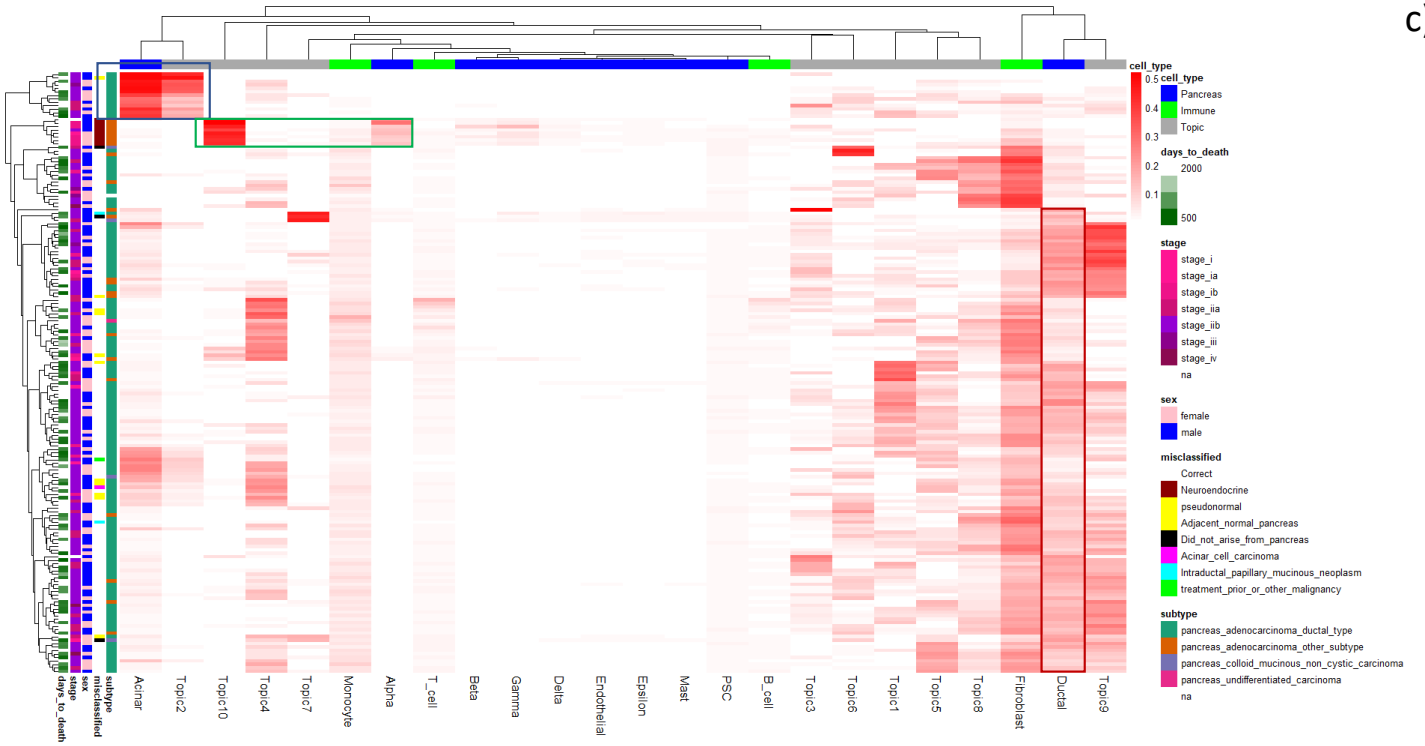


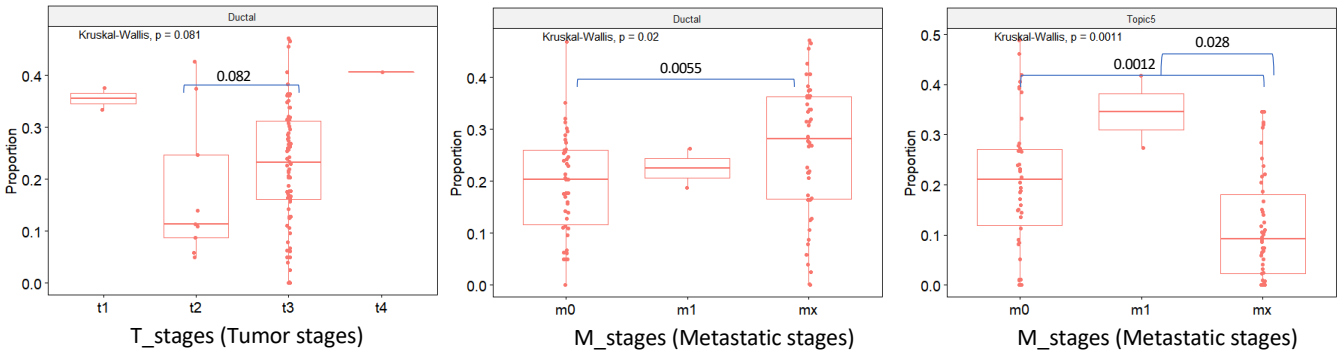
Fig3
a)
PanglaoDB
TRUE
FALSE
NA
CellMarkerDB
TRUE
FALSE
NA
* Genes with pval < 0.05
acinar alpha beta delta ductal endothelial epsilon gamma mast PSC
PanglaoDB CellMarkerDB
214* 331* 149* 11* 428* 0* 0* 59* 0* 64*
b)
Pancreatic islet - Acinar cell
Padj = 6.5e-40 (most significant*)
Pancreatic islet - Alpha cell
Padj = 7.47e-08 (most significant*)
Pancreatic islet - Beta cell
Padj = 1.16e-19 (most significant*)
1.18e-34 (m.s*)
9.71e-42 (m.s*)
2.66e-35 (m.s*)
5.26e-37 (m.s*)
7.88e-11 (m.s*)
1.02e-08 (m.s*)
4.93e-09 (m.s*)
2.38e-10 (m.s*)
3.97e-19 (m.s*)
8.02e-25 (m.s*)
2.63e-18 (m.s*)
1.54e-20 (m.s*)
c)
GSE50244
scReference set 1
Disease
N
IGT
T2D
na
R = 0.849 (p-value < 2.2e-16)
e)
GSE50244
scReference set 2
d)
Disease N IGT T2D
acinar R = 0.18, p = 0.11
alpha R = −0.28, p = 0.013
beta R = −0.4, p = 0.00031
delta R = −0.25, p = 0.029
ductal R = 0.31, p = 0.0057
endothelial R = −0.13, p = 0.24
epsilon R = −0.019, p = 0.87
gamma R = −0.31, p = 0.0069
mast R = 0.098, p = 0.4
PSC R = 0.23, p = 0.041
Cell type proportion
HbA1c
scReference set 1
f)
Disease N IGT T2D
acinar R = 0.27, p = 0.017
alpha R = −0.12, p = 0.3
beta R = −0.45, p = 3.7e-05
delta R = −0.34, p = 0.0022
ductal R = 0.21, p = 0.061
endothelial R = 0.024, p = 0.84
epsilon R = −0.11, p = 0.34
gamma R = −0.3, p = 0.0088
mast R = −0.18, p = 0.11
PSC R = −0.15, p = 0.18
Cell type proportion
HbA1c
scReference set 2

Fig4

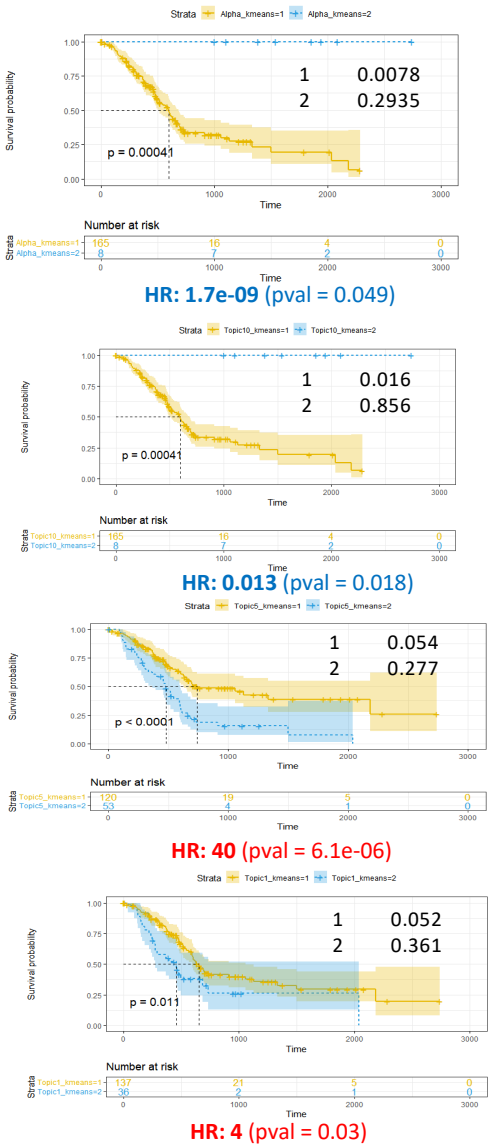
a)


cell_type
Pancreas
Immune
Topic
days_to_death
2000
500
stage
stage_i
stage_ia
stage_ib
stage_iia
stage_iib
stage_iii
stage_iv
na
sex
female
male
misclassified
Correct
Neuroendocrine
pseudonormal
Adjacent_normal_pancreas
Did_not_arise_from_pancreas
Acinar_cell_carcinoma
Intraductal_papillary_mucinous_neoplasm
treatment_prior_or_other_malignancy
subtype
pancreas_adenocarcinoma_ductal_type
pancreas_adenocarcinoma_other_subtype
pancreas_colloid_mucinous_non_cystic_carcinoma
pancreas_undifferentiated_carcinoma
na
Acinar
Topic2
Topic10
Topic4
Topic7
Monocyte
Alpha
T_cell
Beta
Gamma
Delta
Endothelial
Epsilon
Mast
PSC
B_cell
Topic3
Topic6
Topic1
Topic5
Topic8
Fibroblast
Ductal
Topic9
days_to_death
stage
sex
misclassified
subtype

b)


Ductal
Kruskal-Wallis, p = 0.081
0.082
Proportion
t1
t2
t3
t4
T_stages (Tumor stages)
Ductal
Kruskal-Wallis, p = 0.02
0.0055
m0
m1
mx
M_stages (Metastatic stages)
Topic5
Kruskal-Wallis, p = 0.0011
0.0012
0.028
m0
m1
mx
M_stages (Metastatic stages)

c)


Strata
Alpha_kmeans=1
Alpha_kmeans=2
1 0.0078
2 0.2935
p = 0.00041
Survival probability
Time
Number at risk
HR: 1.7e-09 (pval = 0.049)
Topic10_kmeans=1
Topic10_kmeans=2
1 0.016
2 0.856
p = 0.00041
HR: 0.013 (pval = 0.018)
Topic5_kmeans=1
Topic5_kmeans=2
1 0.054
2 0.277
p < 0.0001
HR: 40 (pval = 6.1e-06)
Topic1_kmeans=1
Topic1_kmeans=2
1 0.052
2 0.361
p = 0.011
HR: 4 (pval = 0.03)

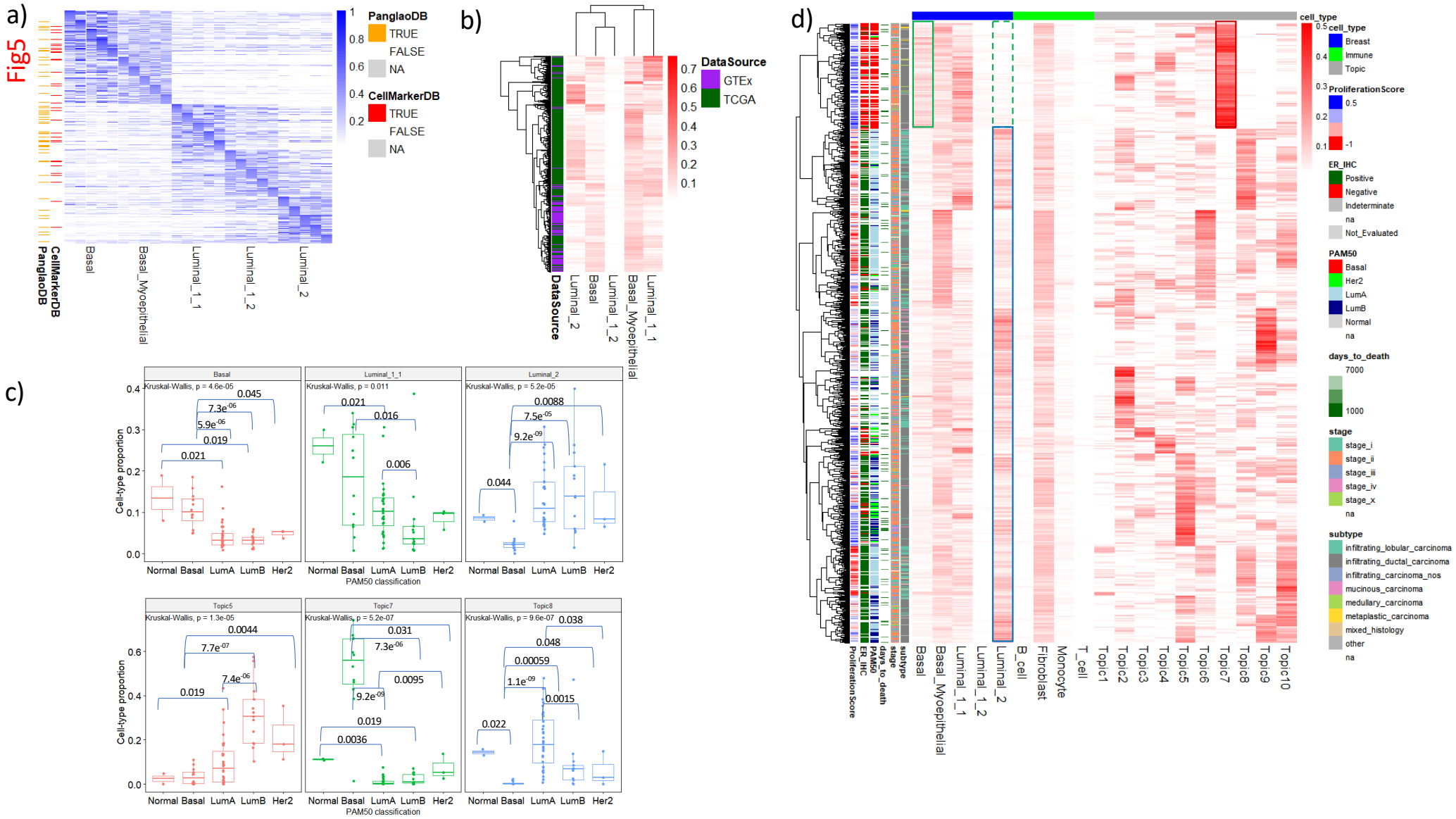
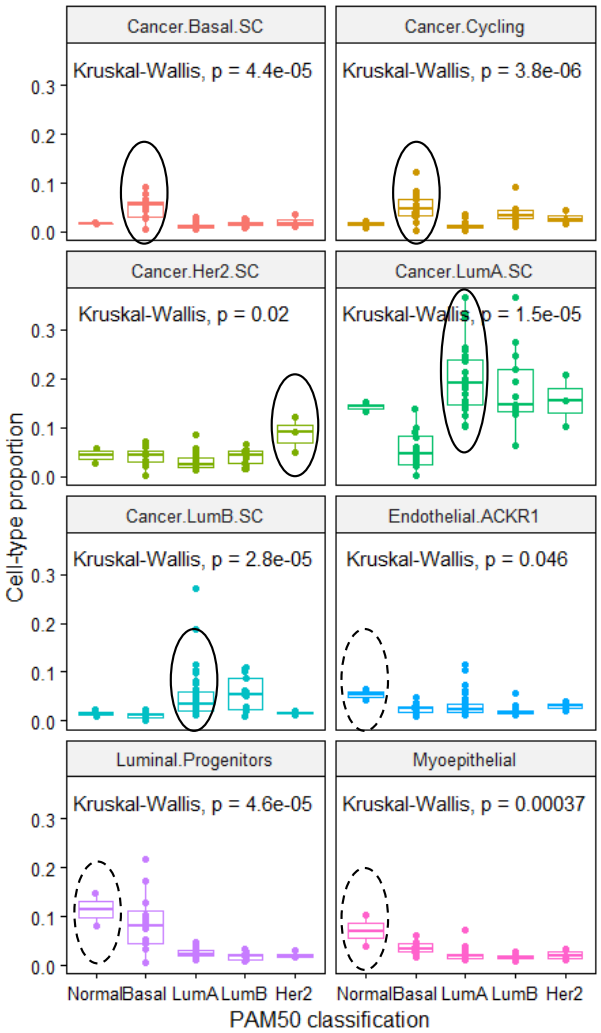
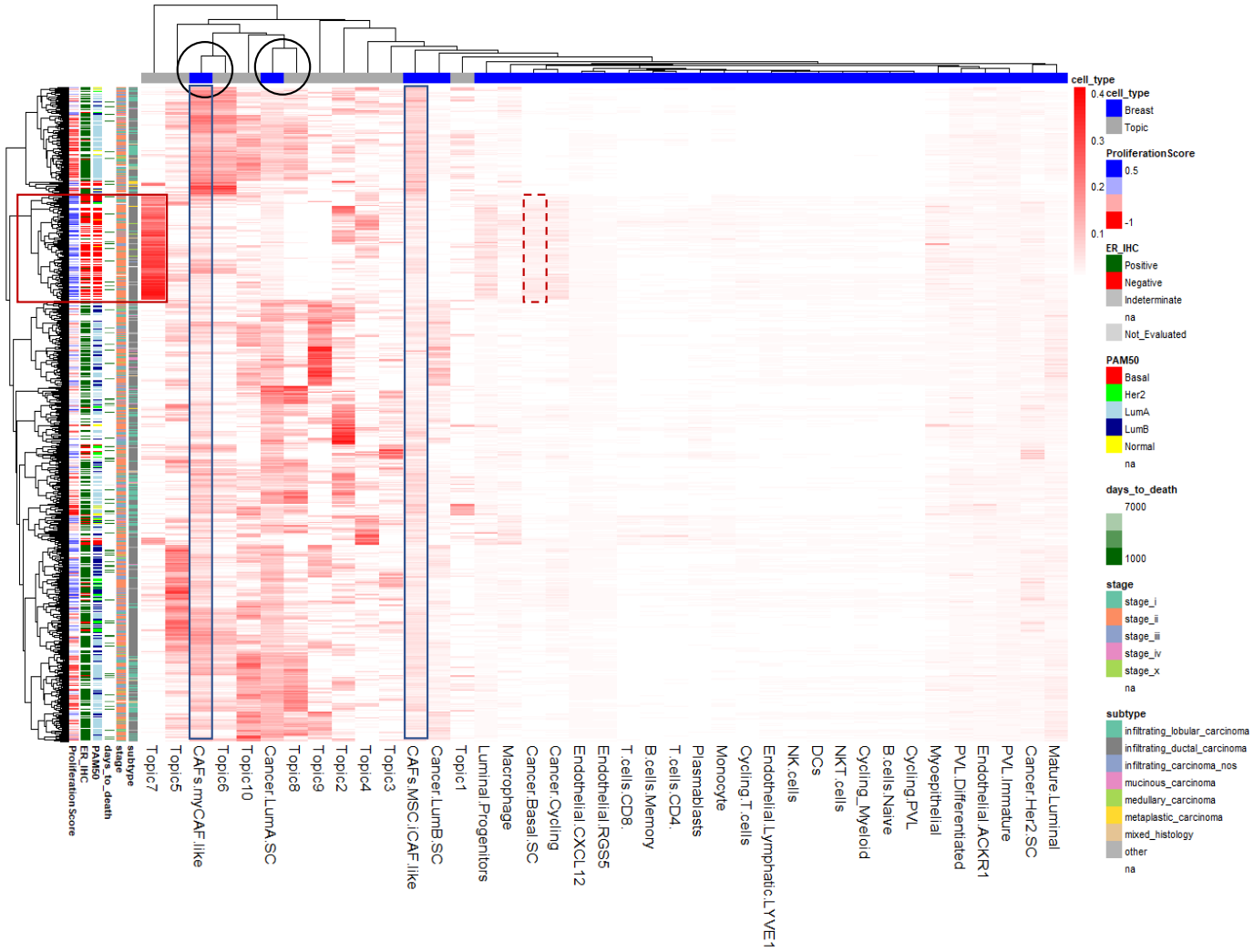


Fig5
a)
PanglaoDB
TRUE
FALSE
NA
CellMarkerDB
TRUE
FALSE
NA
CellMarkerDB
PanglaoDB
Basal
Basal_Myoepithelial
Luminal_1_1
Luminal_1_2
Luminal_2
b)
DataSource
GTEx
TCGA
DataSource
Luminal_2
Basal
Luminal_1_2
Basal_Myoepithelial
Luminal_1_1
c)
Basal
Luminal_1_1
Luminal_2
Topic5
Topic7
Topic8
Kruskal-Wallis, p = 4.6e-05
Kruskal-Wallis, p = 0.011
Kruskal-Wallis, p = 5.2e-05
Kruskal-Wallis, p = 1.3e-05
Kruskal-Wallis, p = 5.2e-07
Kruskal-Wallis, p = 9.6e-07
Cell-type proportion
Normal Basal LumA LumB Her2
PAM50 classification
d)
cell_type
Breast
Immune
Topic
ProliferationScore
0.5
-1
ER_IHC
Positive
Negative
Indeterminate
na
Not_Evaluated
PAM50
Basal
Her2
LumA
LumB
Normal
na
days_to_death
7000
1000
stage
stage_i
stage_ii
stage_iii
stage_iv
stage_x
na
subtype
infiltrating_lobular_carcinoma
infiltrating_ductal_carcinoma
infiltrating_carcinoma_nos
mucinous_carcinoma
medullary_carcinoma
metaplastic_carcinoma
mixed_histology
other
na
ProliferationScore
ER_IHC
PAM50
days_to_death
stage
subtype
Basal
Basal_Myoepithelial
Luminal_1_1
Luminal_1_2
Luminal_2
B_cell
Fibroblast
Monocyte
T_cell
Topic1
Topic2
Topic3
Topic4
Topic5
Topic6
Topic7
Topic8
Topic9
Topic10

a) Fig6


Cancer.Basal.SC
Kruskal-Wallis, p = 4.4e-05
Cancer.Cycling
Kruskal-Wallis, p = 3.8e-06
Cancer.Her2.SC
Kruskal-Wallis, p = 0.02
Cancer.LumA.SC
Kruskal-Wallis, p = 1.5e-05
Cancer.LumB.SC
Kruskal-Wallis, p = 2.8e-05
Endothelial.ACKR1
Kruskal-Wallis, p = 0.046
Luminal.Progenitors
Kruskal-Wallis, p = 4.6e-05
Myoepithelial
Kruskal-Wallis, p = 0.00037
Cell-type proportion
0.0
0.1
0.2
0.3
Normal Basal LumA LumB Her2
PAM50 classification

b)


ProliferationScore
ER_IHC
PAM50
days_to_death
stage
subtype
Topic7
Topic5
CAFs.myCAF.like
Topic6
Topic10
Cancer.LumA.SC
Topic8
Topic9
Topic2
Topic4
Topic3
CAFs.MSC.iCAF.like
Cancer.LumB.SC
Topic1
Luminal.Progenitors
Macrophage
Cancer.Basal.SC
Cancer.Cycling
Endothelial.CXCL12
Endothelial.RGS5
T.cells.CD8.
B.cells.Memory
T.cells.CD4.
Plasmablasts
Monocyte
Cycling.T.cells
Endothelial.Lymphatic.LYVE1
NK.cells
DCs
NKT.cells
Cycling_Myeloid
B.cells.Naive
Cycling.PVL
Myoepithelial
PVL.Differentiated
Endothelial.ACKR1
PVL.Immature
Cancer.Her2.SC
Mature.Luminal
cell_type
0.4
0.3
0.2
0.1
cell_type
Breast
Topic
ProliferationScore
0.5
-1
ER_IHC
Positive
Negative
Indeterminate
na
Not_Evaluated
PAM50
Basal
Her2
LumA
LumB
Normal
na
days_to_death
7000
1000
stage
stage_i
stage_ii
stage_iii
stage_iv
stage_x
na
subtype
infiltrating_lobular_carcinoma
infiltrating_ductal_carcinoma
infiltrating_carcinoma_nos
mucinous_carcinoma
medullary_carcinoma
metaplastic_carcinoma
mixed_histology
other
na

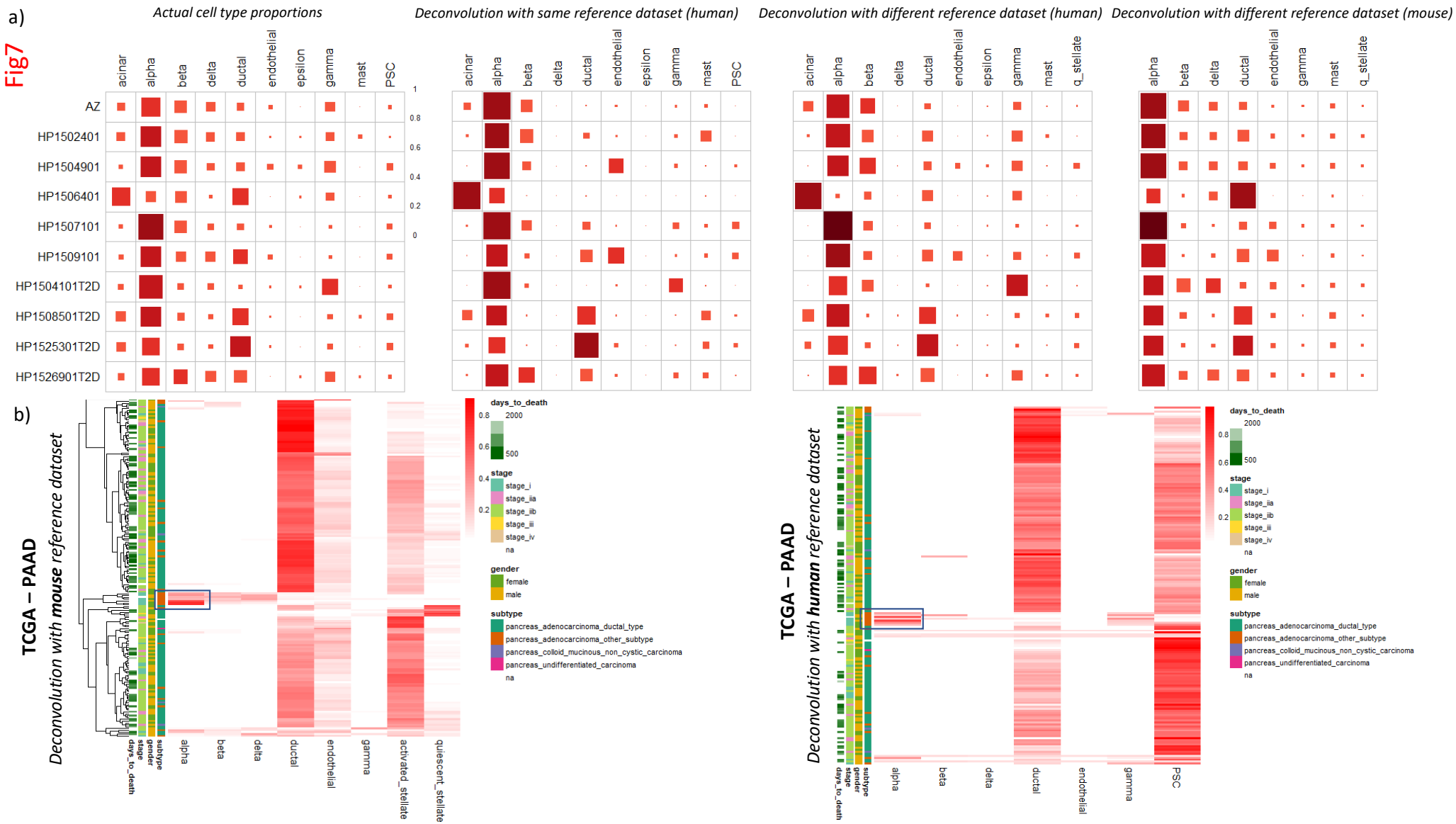
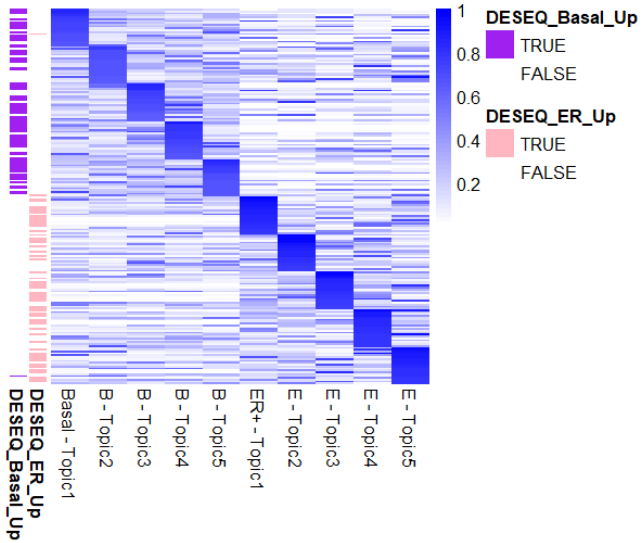


Fig7
a)
Actual cell type proportions
Deconvolution with same reference dataset (human)
Deconvolution with different reference dataset (human)
Deconvolution with different reference dataset (mouse)
acinar
alpha
beta
delta
ductal
endothelial
epsilon
gamma
mast
PSC
q_stellate
AZ
HP1502401
HP1504901
HP1506401
HP1507101
HP1509101
HP1504101T2D
HP1508501T2D
HP1525301T2D
HP1526901T2D
b)
TCGA – PAAD
Deconvolution with mouse reference dataset
Deconvolution with human reference dataset
days_to_death
stage
gender
subtype
stage_i
stage_iia
stage_iib
stage_iii
stage_iv
na
female
male
pancreas_adenocarcinoma_ductal_type
pancreas_adenocarcinoma_other_subtype
pancreas_colloid_mucinous_non_cystic_carcinoma
pancreas_undifferentiated_carcinoma
activated_stellate
quiescent_stellate

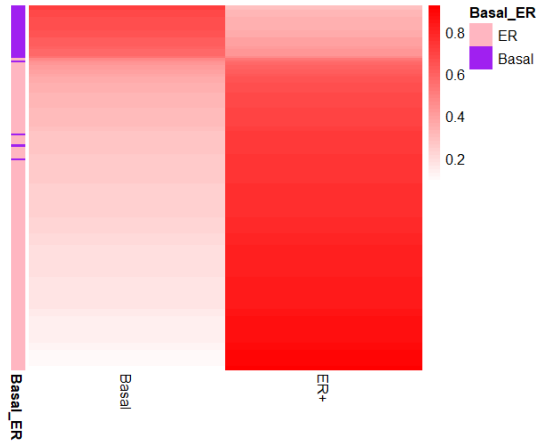
Fig8

Breast Cancer (BRCA) – topics modelled on Basal vs. ER+ phenotypes

a)

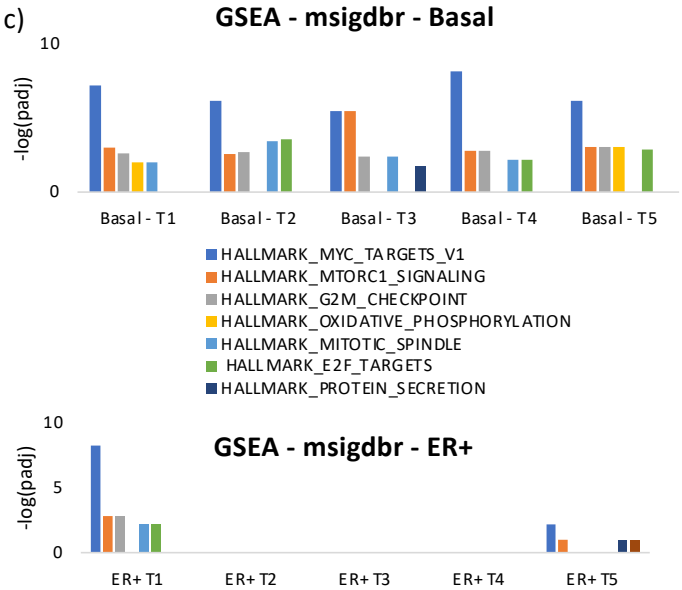


b)



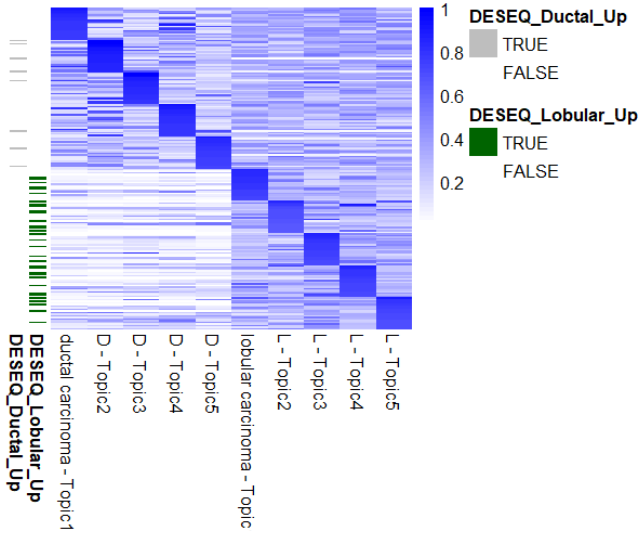
Classifies the 2 phenotypes with 97% accuracy on the held-out dataset (n=160); Using DESEQ genes, accuracy is 83%

c)

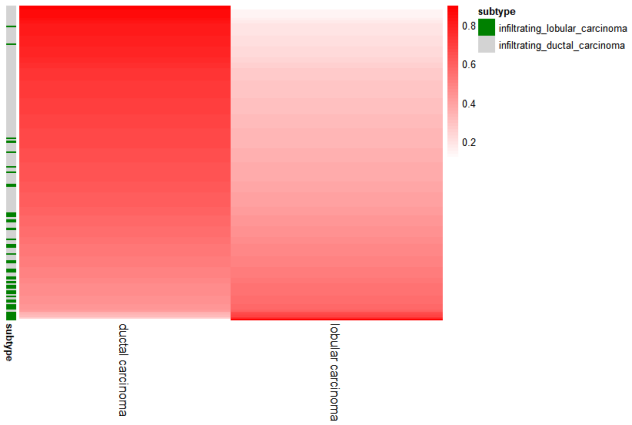


Breast Cancer (BRCA) – topics modelled on histological types

d)

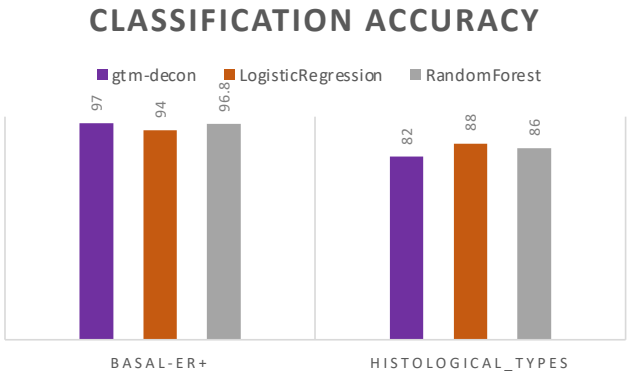


e)



Classifies the 2 phenotypes with 82% accuracy on the held-out dataset (n=177), Using DESEQ genes, accuracy is 75%

f)



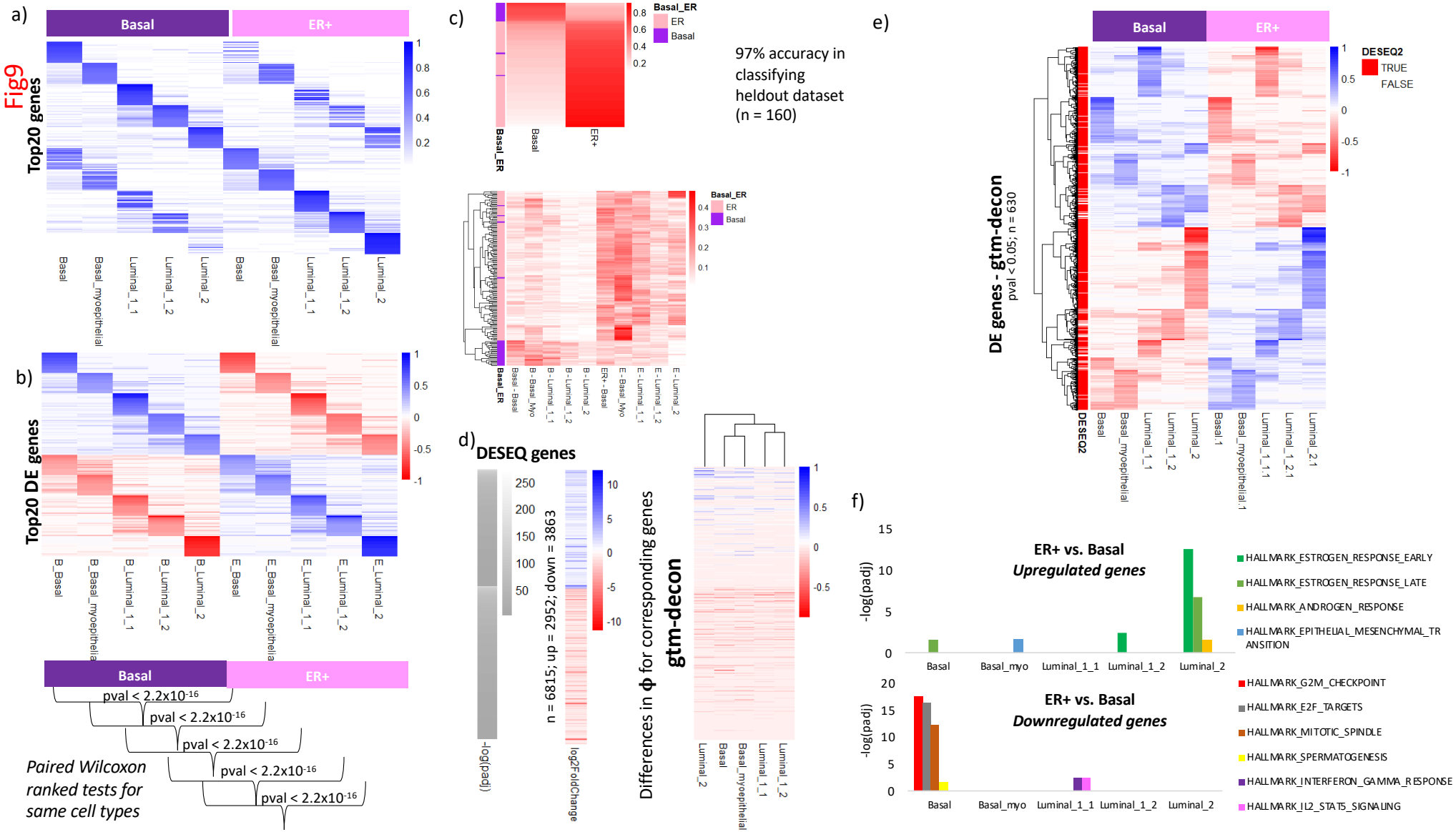


Fig9
a)
Basal
ER+
Top20 genes
Basal
Basal_myoepithelial
Luminal_1_1
Luminal_1_2
Luminal_2
b)
Top20 DE genes
B_Basal
B_Basal_myoepithelial
B_Luminal_1_1
B_Luminal_1_2
B_Luminal_2
E_Basal
E_Basal_myoepithelial
E_Luminal_1_1
E_Luminal_1_2
E_Luminal_2
pval < 2.2x10-16
Paired Wilcoxon ranked tests for same cell types
c)
Basal_ER
ER
Basal
97% accuracy in classifying heldout dataset (n = 160)
d)
DESEQ genes
n = 6815; up = 2952; down = 3863
-log(padj)
log2FoldChange
Differences in ϕ for corresponding genes
gtm-decon
e)
DE genes - gtm-decon
pval < 0.05; n = 630
DESEQ2
TRUE
FALSE
f)
ER+ vs. Basal
Upregulated genes
Downregulated genes
-log(padj)
Basal
Basal_myo
Luminal_1_1
Luminal_1_2
Luminal_2
HALLMARK_ESTROGEN_RESPONSE_EARLY
HALLMARK_ESTROGEN_RESPONSE_LATE
HALLMARK_ANDROGEN_RESPONSE
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
HALLMARK_G2M_CHECKPOINT
HALLMARK_E2F_TARGETS
HALLMARK_MITOTIC_SPINDLE
HALLMARK_SPERMATOGENESIS
HALLMARK_INTERFERON_GAMMA_RESPONSE
HALLMARK_IL2_STAT5_SIGNALING

Figure S1

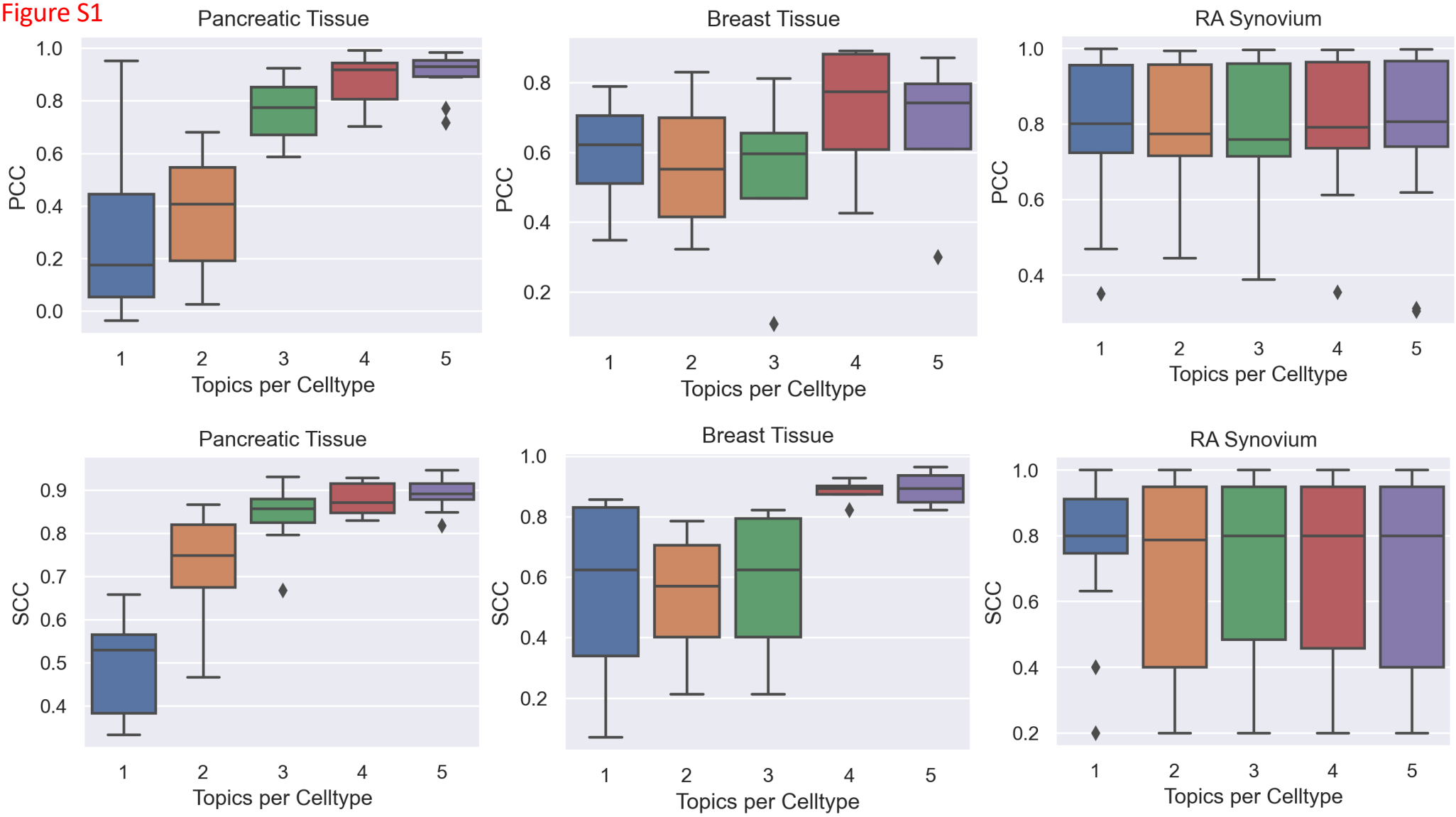

Pancreatic Tissue
Breast Tissue
RA Synovium
PCC
SCC
Topics per Celltype
1
2
3
4
5
0.0
0.2
0.4
0.5
0.6
0.7
0.8
0.9
1.0

Figure S2

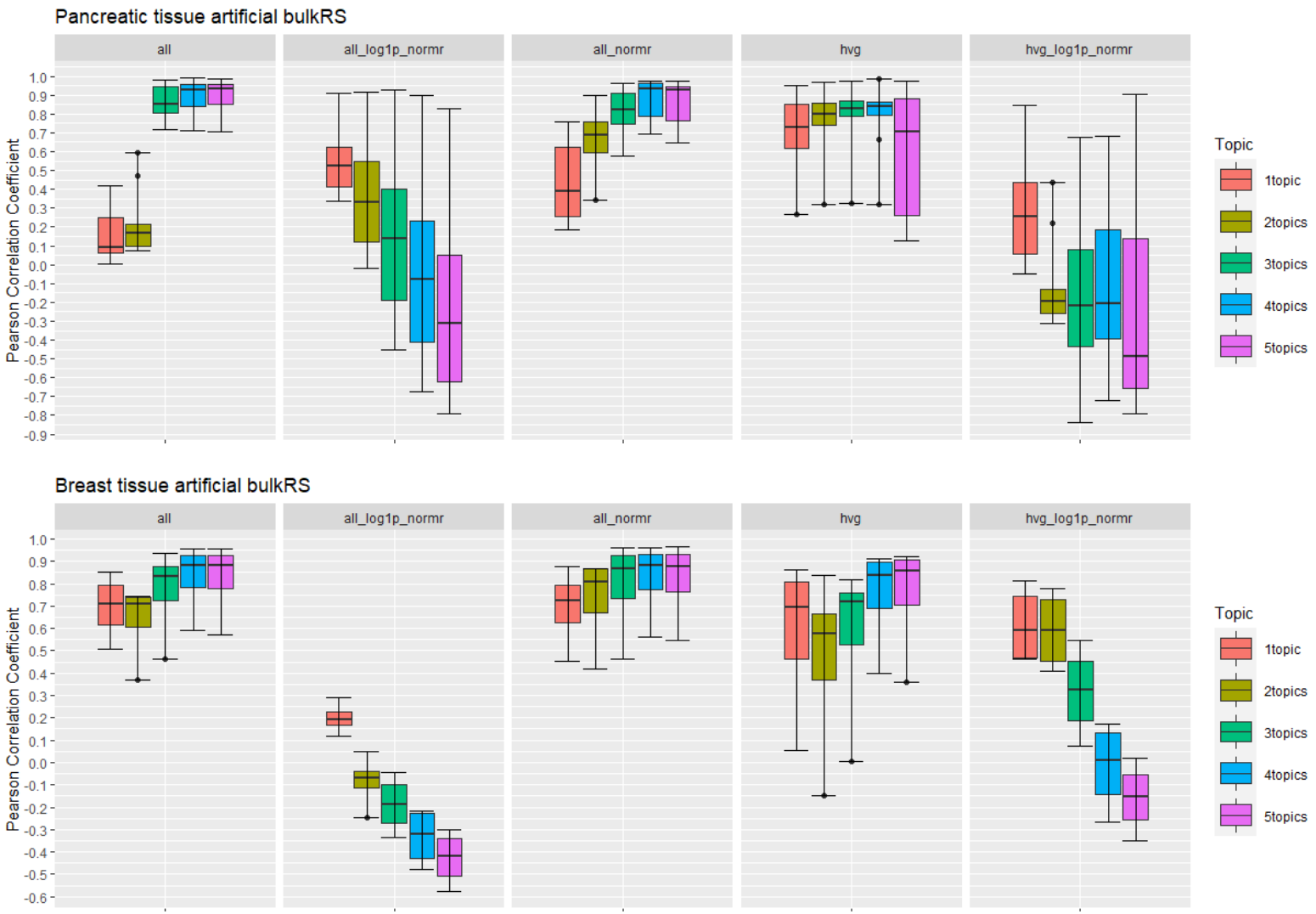

Pancreatic tissue artificial bulkRS
all
all_log1p_normr
all_normr
hvg
hvg_log1p_normr
Pearson Correlation Coefficient
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
Topic
1topic
2topics
3topics
4topics
5topics
Breast tissue artificial bulkRS
all
all_log1p_normr
all_normr
hvg
hvg_log1p_normr
Pearson Correlation Coefficient
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
Topic
1topic
2topics
3topics
4topics
5topics

Figure S3

Gene set enrichment analysis (GSEA) results using fgsea (R package)

Pancreatic dataset - Segerstolpe

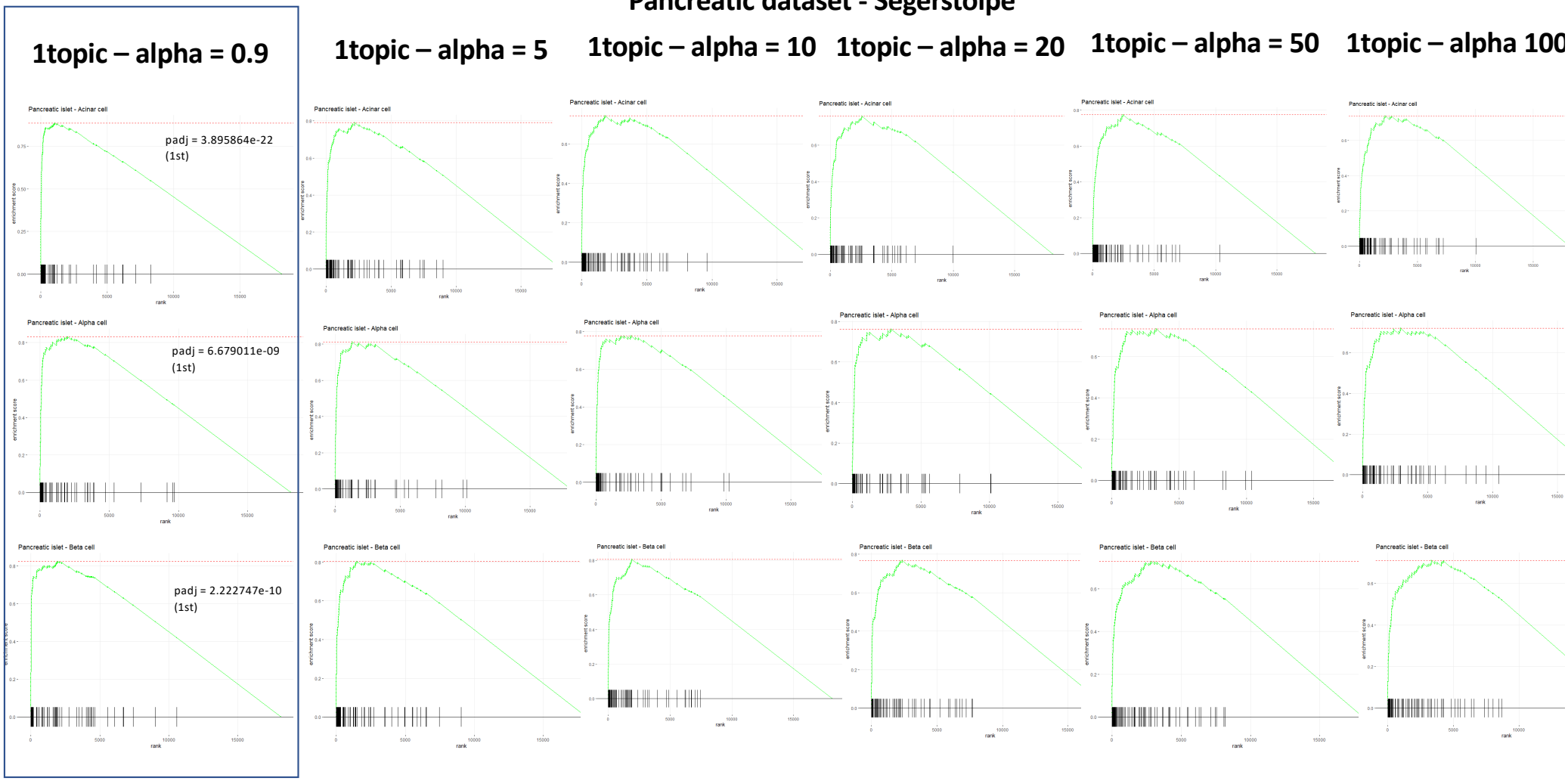


Figure S4

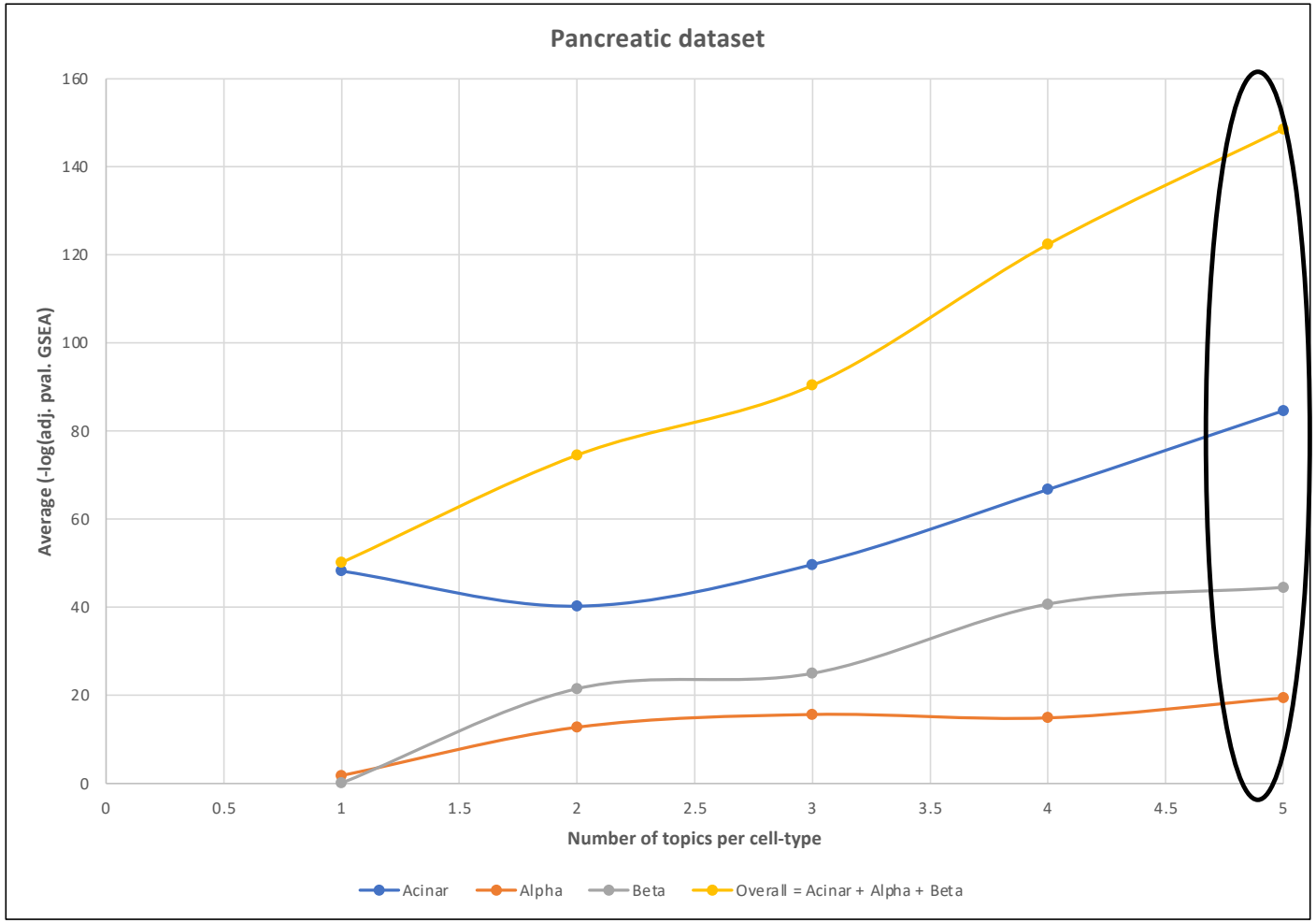

Pancreatic dataset
Average (-log(adj. pval. GSEA)
160
140
120
100
80
60
40
20
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Number of topics per cell-type
Acinar
Alpha
Beta
Overall = Acinar + Alpha + Beta

Figure S5

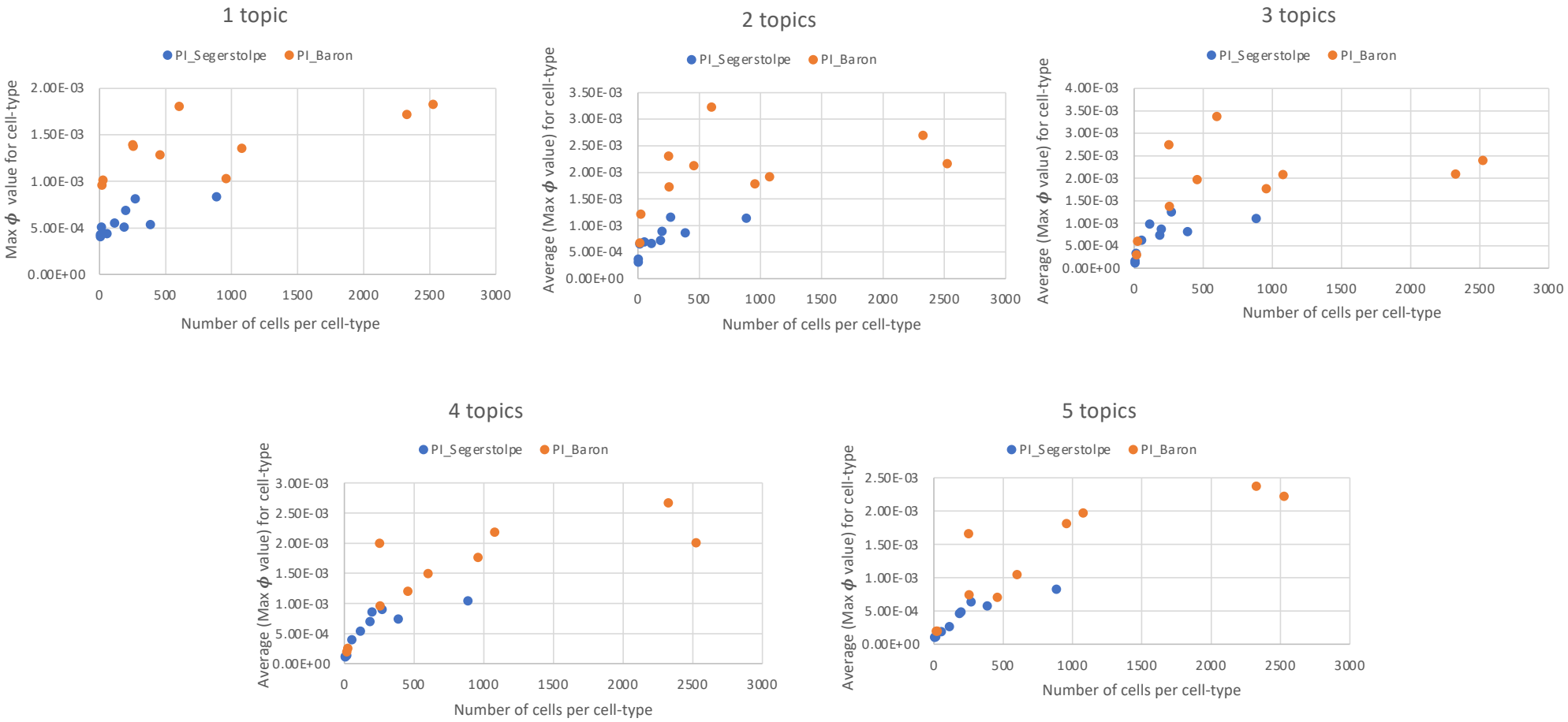
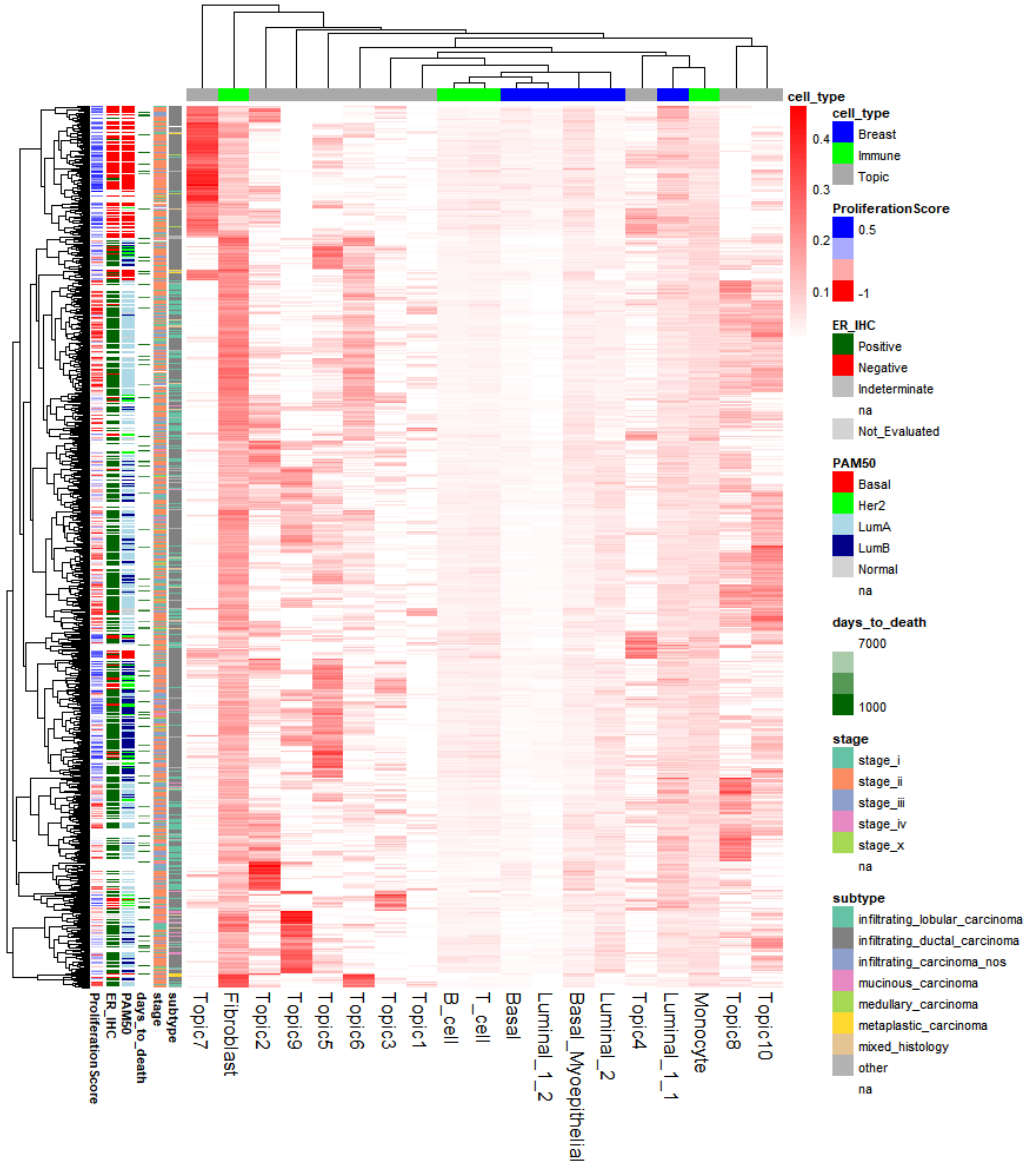

1 topic
PI_Segerstolpe
PI_Baron
Max ϕ value for cell-type
2.00E-03
1.50E-03
1.00E-03
5.00E-04
0.00E+00
0
500
1000
1500
2000
2500
3000
Number of cells per cell-type
2 topics
PI_Segerstolpe
PI_Baron
Average (Max ϕ value) for cell-type
3.50E-03
3.00E-03
2.50E-03
2.00E-03
1.50E-03
1.00E-03
5.00E-04
0.00E+00
0
500
1000
1500
2000
2500
3000
Number of cells per cell-type
3 topics
PI_Segerstolpe
PI_Baron
Average (Max ϕ value) for cell-type
4.00E-03
3.50E-03
3.00E-03
2.50E-03
2.00E-03
1.50E-03
1.00E-03
5.00E-04
0.00E+00
0
500
1000
1500
2000
2500
3000
Number of cells per cell-type
4 topics
PI_Segerstolpe
PI_Baron
Average (Max ϕ value) for cell-type
3.00E-03
2.50E-03
2.00E-03
1.50E-03
1.00E-03
5.00E-04
0.00E+00
0
500
1000
1500
2000
2500
3000
Number of cells per cell-type
5 topics
PI_Segerstolpe
PI_Baron
Average (Max ϕ value) for cell-type
2.50E-03
2.00E-03
1.50E-03
1.00E-03
5.00E-04
0.00E+00
0
500
1000
1500
2000
2500
3000
Number of cells per cell-type

Figure S6


cell_type
0.4
0.3
0.2
0.1
cell_type
Breast
Immune
Topic
ProliferationScore
0.5
-1
ER_IHC
Positive
Negative
Indeterminate
na
Not_Evaluated
PAM50
Basal
Her2
LumA
LumB
Normal
na
days_to_death
7000
1000
stage
stage_i
stage_ii
stage_iii
stage_iv
stage_x
na
subtype
infiltrating_lobular_carcinoma
infiltrating_ductal_carcinoma
infiltrating_carcinoma_nos
mucinous_carcinoma
medullary_carcinoma
metaplastic_carcinoma
mixed_histology
other
na
Topic7
Fibroblast
Topic2
Topic9
Topic5
Topic6
Topic3
Topic1
B_cell
T_cell
Basal
Luminal_1_2
Basal_Myoepithelial
Luminal_2
Topic4
Luminal_1_1
Monocyte
Topic8
Topic10
ProliferationScore
ER_IHC
PAM50
days_to_death
stage
subtype

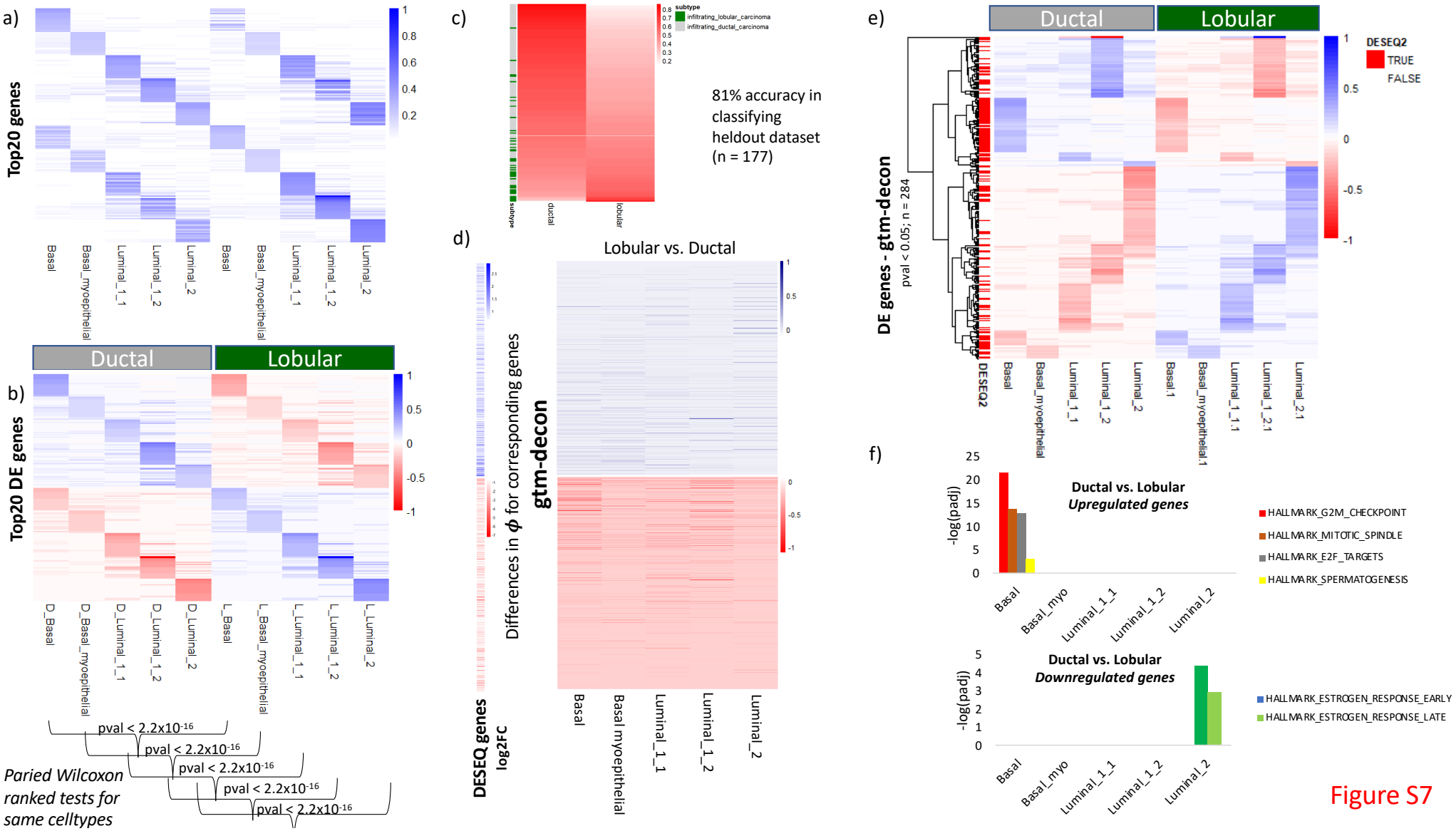

a)
Top20 genes
Basal
Basal_myoepithelial
Luminal_1_1
Luminal_1_2
Luminal_2
b)
Top20 DE genes
Ductal
Lobular
D_Basal
D_Basal_myoepithelial
D_Luminal_1_1
D_Luminal_1_2
D_Luminal_2
L_Basal
L_Basal_myoepithelial
L_Luminal_1_1
L_Luminal_1_2
L_Luminal_2
pval < 2.2x10-16
Paried Wilcoxon ranked tests for same celltypes
c)
subtype
infiltrating_lobular_carcinoma
infiltrating_ductal_carcinoma
ductal
lobular
81% accuracy in classifying heldout dataset (n = 177)
d)
Lobular vs. Ductal
Differences in ϕ for corresponding genes
gtm-decon
DESEQ genes
log2FC
Basal
Basal myoepithelial
Luminal_1_1
Luminal_1_2
Luminal_2
e)
DE genes - gtm-decon
pval < 0.05; n = 284
DESEQ2
TRUE
FALSE
Basal.1
Basal_myoepithelial.1
Luminal_1_1.1
Luminal_1_2.1
Luminal_2.1
f)
Ductal vs. Lobular
Upregulated genes
-log(padj)
HALLMARK_G2M_CHECKPOINT
HALLMARK_MITOTIC_SPINDLE
HALLMARK_E2F_TARGETS
HALLMARK_SPERMATOGENESIS
Basal_myo
Ductal vs. Lobular
Downregulated genes
HALLMARK_ESTROGEN_RESPONSE_EARLY
HALLMARK_ESTROGEN_RESPONSE_LATE

Figure S7