UNIVERSITY OF CALIFORNIA
SANTA CRUZ

TRANSCRIPTIONAL SIGNATURES OF THE TUMOR AND THE TUMOR MICROENVIRONMENT PREDICT CANCER PATIENT OUTCOMES.

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by

Bianca Xue

June 2024

The Dissertation of Bianca Xue
is approved:

Professor Joshua M. Stuart, Chair

Professor Olena Morozova Vaske

Professor Christopher C. Benz

Professor Vanessa Jonsson

Professor David Haussler

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Transcriptional Signatures of the Tumor and the Tumor Microenvironment Predict Cancer Patient Outcomes.

by

Bianca Xue

The cellular components of tumors and their microenvironment play pivotal roles in tumor progression, patient survival, and the response to cancer treatments.

In my doctoral thesis, I describe a new way to extract transcriptional signatures from gene expression data of tumor components and microenvironments and access their influence on cancer patients.

Tumor immune infiltration has been studied for years for its high correlation with patients' survival. Many immune therapies are dependent on the detection and quantification of cell types present in bulk tumors. Bulk tumor microenvironment deconvolution has been largely limited by low number of cell types signatures. Leveraging cell type signatures derived from scRNA-seq data provides a broader range of cell types in detection and quantification of tumor infiltration, therefore helping in developing cancer immunotherapy and targeted cancer therapies.

I developed a new method called scBeacon, a novel tool that derives cell-type signatures by integrating and clustering multiple scRNA-seq datasets to extract signatures from public data consortiums while minimizing batch effects. I derived a comprehensive set of human cell-type signatures from Single Cell Expression Atlas and performed TCGA bulk tumors

deconvolution analysis using the cell-type signature profile. These cell type estimates enable the detection of a pan-cancer high-risk sample group that is not detected by traditional gene expression analysis.

Cancers are traditionally classified into types and subtypes by the organ- and cell-of-origin. However, there are tumor samples that consist of a mixture of cancer subtypes and raise challenges in characterizing the subtype profile for mixture samples. Inspired by the scBeacon deconvolution analysis, I used a similar approach to detect and quantify the subtypes in testicular germ cell cancer mixture samples. In addition, I used single-cell RNA-seq signatures to characterize the major cell types and their differential states in testicular germ cell cancer samples.

For glioblastoma, I used predefined cell state marker genes to deconvolute bulk glioblastoma tumors using a hierarchical deconvolution approach. It proved using hierarchical deconvolution addresses the nature of cell type differentiation, which could give a finer resolution for deconvolution, especially when some rare cell types come from a subpopulation of a very similar cell type. This approach is particularly useful for brain tissue deconvolution because of the complexity of cell type lineages in brain development.

Acknowledgments

I would like to thank the members of my thesis committee for their time and their leadership during my time at UCSC, especially my advisor Josh Stuart for his support and guidance. I would also like to thank every current and past member of the Stuart lab, it was a pleasure to work with every single one of you.

All chapters of this thesis are written about work I did in collaboration with other researchers. The project described in Chapter 1 was finished in collaboration with Hongxu Ding, now submitted to Cell Reports Methods.

The project described in Chapter 2 was in collaboration with Verena Friedl and is now submitted to Cell Reports Methods along with Chapter 1.

The work in Chapter 3 is a contribution to the Testicular Germ Cell Cancer (TGCT) Analysis Working Group (AWG) formed under the National Cancer Institutes Center for Cancer Genomics Genomic Data Analysis Network (GDAN), the research collaboration following the TCGA project. It will be published within the GDAN TGCT AWG under the project lead of Katherine A. Hoadley and Victoria Cortessis.

The projects described in Chapter 4 were conducted as part of the research collaboration with Treehouse Childhood Cancer Initiative, Analiz Rodriguez group at University of Arkansas for Medical Sciences and Vadim Le Joncour at University of Helsinki.

Chapter 1

Motivation and Introduction

Single-cell RNA sequencing (scRNA-seq), first described in 2009[44], has made revolutionary changes in biology. scRNA-seq provides higher resolution compared to traditional bulk RNA sequencing. In contrast to studying the average gene expression profiles of a mixture of diverse components of cells, scRNA-seq enables biologists to study pure cell types at a single cell level, identifying trajectories in organ development from cell differentiation. Since then, more and more cell types in the human body are discovered. As the technology advances, the cost of performing scRNA-seq experiments drops dramatically, leading to an explosion of scRNA-seq datasets deposition on public data consortiums. However, with the abundance of scRNA-seq data, most of the research labs still use a small proportion of total scRNA-seq datasets available to perform analysis, leaving out useful transcriptomic profiles in other datasets due to the limitation of computational resources. It is essential to build a computational efficient pipeline to process large amounts of scRNA-seq data while preserving the biological informa-

tion within transcriptomic data. Motivated by this idea, in my first aim, we built scBeacon, a novel tool that derives cell type signatures by integrating and clustering multiple scRNA-seq datasets to extract cell-type signatures. With scBeacon, I curated 217 cell-type signatures from Single Cell Expression Atlas and annotate cell types for scBeacon signatures with curated marker gene databases and using statistical methods (see Chapter 2).

The World Health Organization reports cancer as the second leading cause of death globally. This year, 2021, 1,898,160 new cases of cancer and 608,570 deaths from cancer are projected to occur in the United States alone [43]. Even though more and more therapies and medicines for cancer have become available in the past few years, the mortality rates for cancer are improving. However, there is still a long way to go to understand cancer biology so that more effective drugs can be developed. Cancer immunotherapy is a revolutionary approach in the field of oncology that harnesses the power of the body's immune system to recognize, attack, and destroy cancer cells. Unlike traditional cancer treatments such as chemotherapy and radiation therapy, which directly target cancer cells, immunotherapy works by boosting the body's natural defenses to fight cancer. Over the past decade, cancer immunotherapy has transformed the landscape of cancer treatment, leading to remarkable successes in treating a wide range of cancers, including melanoma, lung cancer, and certain types of leukemia and lymphoma. However, challenges remain, including identifying biomarkers to predict patient response, overcoming tumor resistance to immunotherapy. However, studies have shown that additional cell types and molecular characters beyond immune cells play an important role in tumor character and response to treatment and patient outcomes[2, 19, 34]. Therefore, it is important to detect and quantify a full profile of cell types to improve our understanding and

treatment of cancer. In Chapter 3, I applied the cell-type signatures developed in Chapter 1 to deconvolute bulk tumor RNA-seq samples in TCGA to construct a comprehensive tumor cell-type profile. Using the deconvolution results, I analyzed how the presence of each cell-types influences the survival probability of patients. Some results coincide with known cancer biology properties such as subtypes.

In Chapter 4, I worked with Testicular Germ Cell Cancer working group(TGCT) to understand testicular cancer histology. I contributed to the analysis of characterizing the subtypes in undefined mixture samples, and validated my work with the microscopy results from pathologists. I've also delved into deconvoluting cell-type compositions within testicular tumor samples, gaining a comprehensive understanding of how these ratios correlate with tumor subtypes and differentiation stages.

The effectiveness of deconvolution tools is often constrained by their capacity to precisely deconvolute hierarchical subpopulations within complex cell mixtures, since the parent and children cell types on the same developmental linage often share similar marker gene profile with very few marker genes to differentiate them. This is especially a challenge for characterizing brain and brain tumor cell-types. I developed a new approach that incorporates hierarchical marker genes to construct signature gene profile for deconvolution. In this way, tumor deconvolution tools gained the ability for precise identification of small cellular subsets that could potentially be useful in cancer therapy (see Chapter 5).

Chapter 2

Building comprehensive cell-type signatures from scRNA-seq datasets

The project described in this chapter was published at Cell Reports Methods. I worked with Hongxu Ding, who was a postdoc in Stuart lab, to come up with the idea, then developed and implemented this project, scBeacon, in an R package.

I tested and optimized the method by running it on 63 scRNA-seq datasets collected from Single Cell Expression Altas, visualized the results, and investigated if the cell-type signatures generated from scBeacon could potentially be used for bulk tumor deconvolution.

2.1 Introduction

Single-cell RNA sequencing (scRNA-Seq), first described in Tang,2009[44], has since transformed biological research. For the first time, it is now possible to determine gene expres-

sion separately for each cell in a biological sample. The technology offers the opportunity for a more detailed and more accurate definition of cell types and cell states. With the explosion of scRNA-seq datasets deposition on public data consortiums, it is essential to build a computational efficient pipeline to process large amounts of scRNA-seq data while preserving the biological information within transcriptomic data. Motivated by this idea, we built scBeacon, a scRNA-seq processing pipeline, to rapidly cluster and integrate datasets and build a comprehensive human cell type signature profile. We validated scBeacon is not only able to handle batch effect effectively, but also computational efficient while ingesting and integrate datasets, without losing biological variance.

Few computational resources exist that automatically extract cell type information from scRNA-seq repositories in an unsupervised manner. SCDC[12] leverages multiple scRNA-seq reference datasets by integrating the deconvolution results with optimized weights. UniCell[8] is one such recent approach that uses a deep learning model trained on hundreds of fully annotated scRNA-seq datasets representing 840 cell types for comprehensive cell types deconvolution. However, deep learning approaches can lack robustness and lead to "black box" solutions that are difficult to interpret and share. In contrast, Ecotyper[28] used linear gene expression vectors extracted from single cell RNA-seq clusters that extend the original LM22 signatures into 64 immune system-related cell types used to deconvolute TCGA samples. Similarly, TIMEx[51] extracted 37 immune-related cell-type signatures from a pan-cancer single cell RNA-seq data compendium and performed enrichment based deconvolution on TCGA bulk tumors.

I extracted 217 cell type signatures from Single Cell Expression Atlas, 602,359 sin-

gle cells in total. It extends the work of TIMEx and Ecotyper by including additional single cell datasets beyond cancer samples and incorporates additional cell types beyond malignant and immune system types. We introduce a novel non-parametric signature comparison metric that can detect related clusters across diverse datasets and merge them into a single cell type signature.

I also validated if the cell-type signatures generated from scBeacon pipeline could be potentially used for bulk tumor deconvolution using a variety of PBMC(peripheral blood mononuclear cell) scRNA-seq datasets. I proved scBeacon generated signature can not only used for deconvolution in CIBERSORT, but also have better performance compared to traditional count-based signatures, used in most of the deconvolution researches.

scBeacon was developed as part of a collaboration with Seagate technologies. In that project, the goal was to optimize the retrieval of cell signatures using fast relational queries. For that reason, only rank-based transformations were considered. While I do show these rank transformations, perhaps surprisingly, outperform standard representations (like TPM) for the purposes of deconvolution, an exploration of other normalization and batch correction techniques could reveal further improvements to combining cell signatures across multiple diverse datasets.

2.2 Methods

2.2.1 Data

1. The Single Cell Expression Atlas (SCEA), a part of EMBL-EBI's Expression Atlas, is a public single-cell RNA sequencing data consortium that hosts datasets from published studies for six species8. For this analysis, we downloaded the 62 homo sapiens single-cell RNA sequencing datasets available in February 2020. The datasets come from a wide range of healthy and diseased tissues, consisting of numerous cell types in the human body. The 62 datasets were sequenced using different single-cell RNA sequencing techniques, such as 10X Genomics platform, smart-seq, drop-seq etc.

2. PBMC scRNA-seq datasets:

   (a) Ding, Jiarui, et al. "Systematic comparison of single-cell and single-nucleus RNA-sequencing methods." Nature biotechnology 38.6 (2020): 737-746.

   (b) 10X-v1: 6k PBMCs from a Healthy Donor, Platform: 10XGenomics v1 chemistry, Single Cell Gene Expression Dataset by Cell Ranger 1.1.0, published on July 31, 2016

   (c) 10X-v2: 8k PBMCs from a Healthy Donor, Platform: 10XGenomics v2 chemistry, Single Cell Gene Expression Dataset by Cell Ranger 2.1.0, published on November 8, 2017

   (d) 10X-v3: 10k PBMCs from a Healthy Donor, Platform: 10XGenomics v3 chemistry, Single Cell Gene Expression Dataset by Cell Ranger 3.0.0, published on November

19, 2018

3. Gene sets used to annotate cell-type signatures are obtained from PanglaoDB[42], Harmonizome[40], and the cell type pathways from MSigDB (C8)

4. MOCA (Mouse Organogenesis Cell Atlas) dataset[7] contains hundreds of cell types and 56 trajectories from mouse embryos, staged between 9.5 and 13.5 days of gestation.

5. scRNA-seq datasets used to create synthetic bulk samples: human head and neck cancer dataset[36] (GSE103322), human melanoma dataset[45] (GSE72056), and bulk RNA-Seq data for each of 6 different melanoma cell lines[33].

2.2.2 Validation of Reciprocal Top-K Enrichment (RTKE) metric

The scBeacon workflow relies on exemplar signatures, i.e. gene expression profiles aggregated across many single cells similar enough to be clustered together, constructed from multiple clusters, from possibly multiple datasets, derived from several scRNA-seq platforms (Fig 2.1A; see Methods). To help mitigate possible batch effects, we use signatures based on ranking the gene expression data; i.e. each cluster's expression profile is rank transformed to form a rank centroid before it is compared to other clusters. We created a Reciprocal Top-K Enrichment (RTKE) metric to detect if the highest expressing genes of one cluster's rank centroid are among the top-ranking expressed genes in another cluster's rank centroid and vice versa. Using RTKE to link related individual clusters, the pipeline then attempts a second clustering aggregation step, treating the clusters as the items to be clustered, to identify exemplars from the metacluster groups of clusters that could represent possible cell types in particular cell states

encompassing various tissues, contexts, and developmental stages.

Rank-transformation of the expression vectors is good for harmonizing across datasets, but it could reduce the cell-type signal that is present in the expression centroids, negatively affecting their usefulness in deconvolution tasks. To address this issue, we compared the deconvolution performance of rank-transformed expression centroids to that of count-based expression centroids. First, to illustrate that ranking preserves a high degree of cell type information, we plotted the PBMC data from multiple datasets and platforms using either the original cluster centroids based on averaged gene expression counts (Fig 2.1B) or using the corresponding rank centroids (Fig 2.1C). Except for two cases from smart-seq2, the rank normalization centroids are aggregated by cell type instead of by sequencing platform. In addition, using the RTKE metric as the distance function enhances the distinction between the three major cell types (Fig 2.1D). Thus, ranking and RTKE complement to preserve cell type information useful for identifying cell types across scRNA-seq datasets of different platforms and batches.

2.2.3 In silico immune infiltration evaluation

We created different types of in silico cell type mixtures simulating immune infiltration in cancer tissue in order to validate the 217 cell type signatures for deconvolution. We created 200 in silico mixtures from scRNA-Seq data from a human head and neck cancer dataset58 (GSE103322), human melanoma dataset59 (GSE72056), and bulk RNA-Seq data for each of 6 different melanoma cell lines60. The centroid of all scRNA-Seq tumor cells or bulk RNA-Seq cell lines in each dataset was used to represent the tumor components of the mixture. The tumor components was randomly assigned a mixture percentage between 50 and 90%. The rest of the

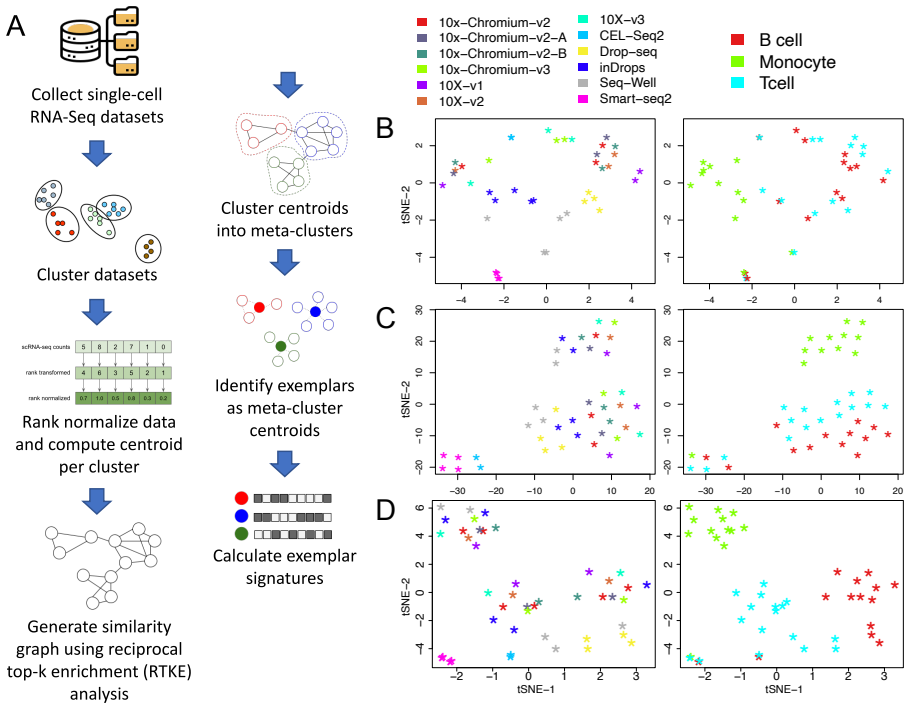

Figure 2.1: **scBeacon workflow and validation.** **(a)** scBeacon workflow. Individual scRNA-seq datasets are clustered using Louvain clustering. Cluster centroids are ranked and then compared to each other using a novel reciprocal top-k enrichment (RTKE) metric. High scoring cluster pairs that exceed an empirically determined threshold are retained as a graph for further clustering to identify exemplars from associated groups of metaclusters. Exemplar centroids are computed by averaging the cluster ranked centroids and recorded as exemplar signatures, assumed to be proxies of cell type signatures for downstream analysis (see Methods). B-D tSNSE plots of PBMC scRNA-seq centroids using different transformations of the count-based data or similarity calculations between centroids. Left panel in each plot shows cells colored by single-cell sequencing technology platform (10x version 1 or 2 chemistries, green; 10x version 3 chemistry, aqua; 10x Chromium version 3 chemistry, light green; 10x Chromium version 2A chemistry, red; 10x Chromium version 2B chemistry, orange; CEL-Seq2, lightblue; Drop-seq, medium blue; inDrops, dark blue; Seq-Well, purple; Smart-seq2, pink). Right panels in each plot show cells colored by cell type (T cells, lightblue; B-cells, red; Monocytes, green). **(b)** Centroids represent vectors of count-based data (transcripts per million reads, TPMs). **(c)** Same as B, but centroids were rank-normalized. **(d)** Same as C, but using the matrix of RTKE similarity metrics as input to tSNE.

mixture was randomly distributed between immune and microenvironment cell-type centroids in integer-valued percentages: B cells, dendritic cells, NK cells, endothelial cells, fibroblasts, macrophages, mast cells, myocytes, and T cells.

For the melanoma cell lines dataset, the immune cell types were purified from blood using marker genes in a vaccination study61. We take the average of the 2 patients at time point t0 (before vaccination) to represent pure cell type references. For both datasets, the expression data were reduced to the overlapping genes between the two datasets and quantile normalized to remove batch effects and enable mixing.

To validate this approach, we used scRNA-Seq PBMC (peripheral blood mononuclear cell) datasets from different sequencing platforms62. PBMCs consist mainly of monocytes, B cells, and T cells, with other minor fractions of dendritic cells, NK cells, and macrophages63. We created cell-type signatures from scRNA-Seq PBMC datasets62 from various single-cell sequencing technologies, e.g. 10X Chromium, CEL-Seq2, Drop-Seq, inDrops, Seq-Well. Additional PBMC datasets were downloaded from the 10X Genomics website, Chromium demonstration data64

We clustered each dataset using the Louvain algorithm and assigned three main clusters to monocytes, B cells, and T cells using the expression of established marker genes (CD3E for T cells, MS4A1 for B cells, and CD14 for monocytes). We calculated centroid for each cell type and generated a signature matrix for each dataset.

We performed three different deconvolution approaches with the signatures to determine if ranking the centroids and combining the signatures produced accurate deconvolution results. First, the log-transformed, count-based TPM (transcripts per million reads)-normalized

centroids from the 10X-v2 dataset alone were used as the signature matrix in deconvolution. Second, the rank-normalized centroids from the 10X-v2 dataset were used on their own as the signature matrix for deconvolution. Finally, the rank-normalized centroids were combined with all PBMC scRNA-Seq datasets and used as a combined signature matrix in deconvolution.

2.2.4 CIBERSORT deconvolution to identify cell types in bulk samples

The exemplar cell-type signatures generated from the scBeacon workflow were used for deconvolution of cancer bulk RNA-Seq data, in which each signature's contribution to the mixture was estimated. We used the Cibersort deconvolution method55, which performed well in the DREAM deconvolution competition56. We ran Cibersort with parameters: perm=100, QN=FALSE, absolute=TRUE, abs_method='no.sumto1'.

We used rank-normalized cell type signatures in CIBERSORT to deconvolute TCGA bulk tumors. Compared to cell type signatures derived from count-based expression values, rank-normalized signatures outperformed count-based signatures in bulk tumor deconvolution, which is commonly used in other deconvolution approaches. This was validated by our validation analysis using synthetic bulk samples.

In this study, we used the TCGA collection as the bulk tumor data for deconvolution. We downloaded the counts per tumor type data from Xena57, which represents The Cancer Genome Atlas (TCGA) gene expression HTSeq counts data originally provided by the NCI's Genomic Data Commons. We normalized the count data to TPM (transcripts per million reads).

2.2.5 scBeacon - Exemplar Signature Derivation

The scBeacon workflow is shown in Fig. 2.1A. Starting from a compendium of scRNA-Seq datasets, the cells in each dataset are clustered using the Louvain algorithm46. We found louvain clustering had the best performance regarding speed and memory efficiency on different environments (dgtMatrix in R, pandas data frame in python) compared to various other clustering algorithms, e.g. k-means and hierarchical clustering.

We used cell-wise rank normalization to reduce any possible batch effects that would be introduced by integrating clusters across different datasets. For each cell cluster, a centroid was computed by taking the average rank of a gene across all the cells in the cluster, resulting in a rank average for each gene. We found that rank normalized centroids were accurate and robust representations of single-cell clusters. First, we found that rank centroids accurately preserved biological information using the MOCA (Mouse Organogenesis Cell Atlas) dataset51 as a test case. Centroids "islands" in different colors were found to represent unique cell types in MOCA (Supplementary Fig. S15A) and the developmental trajectory was well preserved according to the annotated murine developmental stages (Supplementary Fig. S15B). Second, we found that generating rank normalized centroids from 50 cells is robust to represent a cluster based on subsampling cells and finding that the Spearman correlation between the centroid derived from a random subset and its corresponding centroid from the complete data saturates at 50 (Supplementary Fig. S16).

In order to group centroids by unique cell types, we used the reciprocal top-k enrichment (RTKE) method introduced in the Biological Process Activity manuscript by Ding et

al.52. After rank-normalization, the top 10 percent of the genes in a centroid were used to perform RTKE and the enrichment scores were used as similarity scores to compare all centroids to each other. We found that other choices for the top k genes, ranging from 5% up to 40%, yielded highly similar results as using the top 10% of genes (Supplementary Fig. S17A).

To cluster the cluster centroids into meta-clusters that include similar cell types, we compute the empirical distribution of similarity scores. We set a threshold for centroids as 0.006 upper quantile of the empirical distribution and then use the louvain clustering to define meta clusters. Each meta-cluster is considered to represent a cell type, and it can be made of multiple centroids from one or multiple datasets or be a unique cluster from just one dataset. The 0.006 top quantile is selected by screening through thresholds ranging from 0 up to 0.1, the clustering results are evaluated using Silhouette scores, when threshold is 0.006, Louvain clustering reaches the highest Silhouette score (Supplementary Fig. S17B).

The meta-cluster centroids, also called exemplars, are used as cell-type signatures. To obtain cell-type signatures for tumor deconvolution, we first constructed a differential gene expression matrix. For each cell type, we identify a unique set of genes that distinguishes it from other signatures. First, we compute the average expression of each gene in the 217 cell types. For each gene, we subtract the average expression value of the highest-expressing cell type and the second-highest expressing cell type. This strategy ensures that only genes expressed distinctly high in each cell type are included in the signature matrix, which is key for subsequent analyses as overlap in gene signatures between cell types can complicate deconvolution results. The 20% most differentially expressed genes are chosen as signature genes and this subset matrix was used as the signature matrix as the input for CIBERSORT. Supplementary Fig. 9

shows a heatmap of the 217 cell type signatures. We used this signature matrix in bulk tumor deconvolution.

2.2.6 Annotating the SCEA signatures using Pathway enrichment

To better understand the biological features of SCEA signatures, we use GSEA enrichment analysis to test for both enriched cell types and pathways by using a combination of gene sets from PanglaoDB53, Harmonizome54, and the cell type pathways from MSigDB (C8). To maintain specificity as well as robustness for the enrichment analysis, we retained gene sets that had more than 50 genes and less than 100 genes. This resulted in a collection of 5398 gene sets in total – 178 from PanglaoDB, 84 from Harmonizome, 4436 from MSigDB GO genesets and 700 from MSigDB cell type genesets(Supplemental Table S6). For each signature, we used GSEA to score and rank all of the gene sets in the collection. The top five ranking gene sets for each cluster was recorded in an annotation table (Supplemental Table S2). We also used cell-level annotations published in the manuscripts that described the dataset from which a cluster was derived and prioritized using these author-provided annotations to label a cluster centroid wherever it was available. If multiple annotations were present among the cells in a cluster, a summary annotation "short name" was created. Manual inspection of the cases where author annotations and PangloDB-inferred annotations were both available revealed a high concordance between the independently derived annotations (see Table S2). In the absence of an author-derived annotation, a "short name" was created by summarizing the top ranking gene sets for the associated signature.

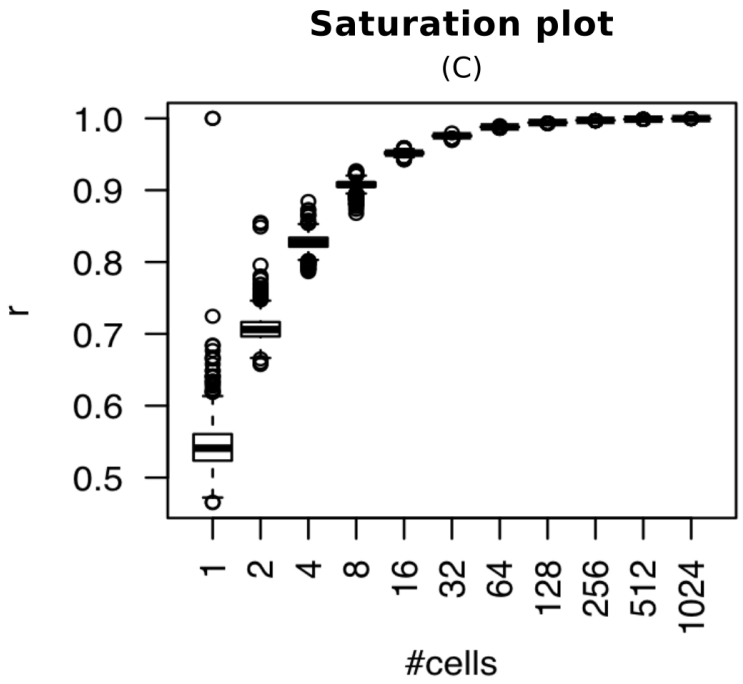

Figure 2.2: **Number of cells needed to reliably estimate cluster centroid** Y-axis shows Spearman correlation coefficient between the centroid from all the cells of a cluster with centroid from sampled cell, number of cells sampled shows on x-axis. Correlation coefficient saturates when samples of over 50 cells are included in the analysis.

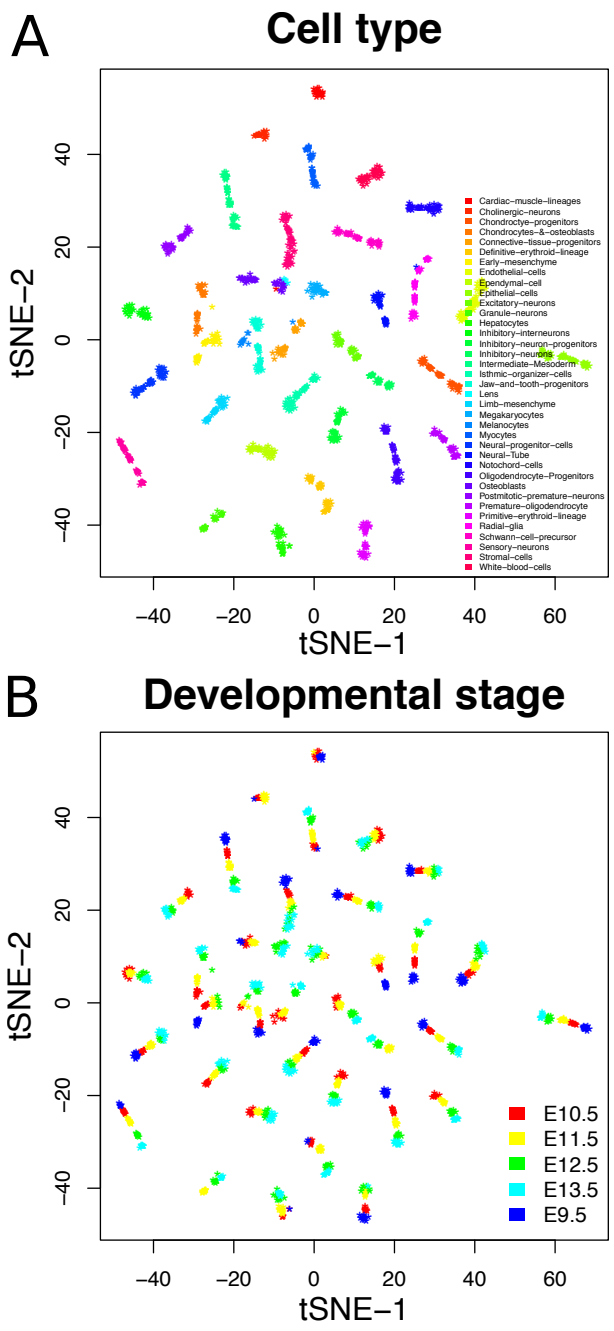

Figure 2.3: **tSNE plot of Mouse Organogenesis Cell Atlas (MOCA) dataset after scBeacon pipeline, clustering solution is curated by cell type annotation in MOCA. Data points are centroids (a)** tSNE plot of MOCA centroids, colored by cell types. **(b)** same as A, colored by mouse developmental stages from E9.5 to E13.5.

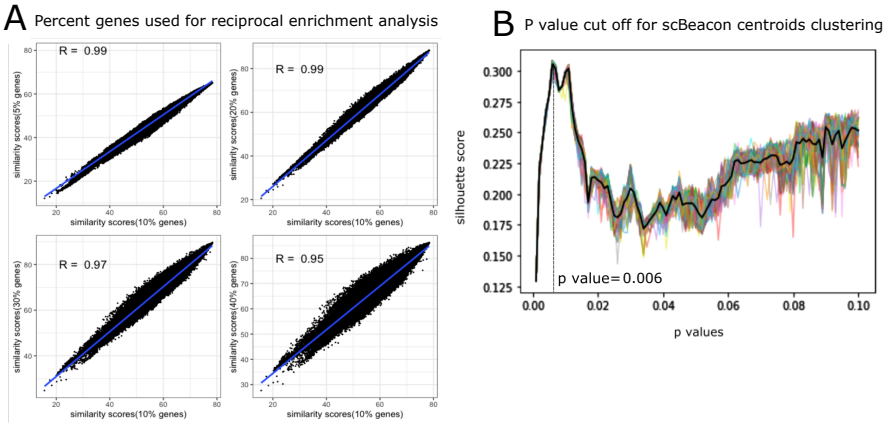

Figure 2.4: **Clustering solution and percent genes for reciprocal enrichment analysis in scBeacon (a)** We chose 10% top ranked genes to perform reciprocal enrichment analysis, here we compared the similarity score generated using top 5%, 20% 30% and 40% genes using correlation plot, with correlation coefficient shown in the plots. **(b)** Use silhouette score to decide p-value cut-off to build scBeacon centroids adjacency matrix for louvain clustering (resolution=0.7). To decide scBeacon centroids clustering solution, we evenly spaced 100 p values from 0 to 0.1 and calculated silhouette scores for louvain clustering. For each p-value, silhouette scores were calculated under 50 different random seeds, the black line is the average silhouette score of the 50 silhouette scores calculated. From the plot, the highest silhouette score is reached at p value=0.006, silhouette score=0.3, generating 217 clusters.

2.3 Results

2.3.1 Validation of ranked cell-type signatures for deconvolution

Next, we measured the effectiveness of rank centroids for their use as exemplar signatures for deconvolving in silico mixtures. To this end, we created in silico mixtures from single-cell as well as bulk RNA-Seq data that simulate immune infiltration into tumor tissue. We created in silico mixtures by combining several PBMC-related expression signatures together at known mixing proportions. The expression signatures were generated by taking the average of single-cell transcriptomes sampled from pre-established clusters either from the PBMC dataset or a published scRNA-seq cancer dataset. Next, we measured the accuracy of CIBERSORT deconvolution for identifying and quantifying the PBMC cell types at the prescribed mixing proportions.

We compared the use of count-based signatures to rank-based signatures for deconvolution and found that rank-based signatures provided slightly more accurate estimates of cell proportions. We used both the Pearson correlation and the Root Mean Square Error (RMSE) to measure the concordance between the known to predicted levels. The Pearson correlation measures if the estimates track with one another in a relative sense while the RMSE measures how close each estimate is to their actual known levels. For the count-based signatures, we used an immune cell-type signature matrix derived from a scRNA-Seq PBMC dataset with TPM count-based expression values to deconvolute a synthetic bulk melanoma single-cell dataset containing infiltrating immune cells (Fig 2.2A). For the rank-based signatures, we formed exemplar rank centroids by averaging the rank centroids of clusters found in multiple PBMC

datasets (Fig 2.2B). While there is not a consistent trend over all three immune cell types, the deconvolution estimates using the scBeacon-derived signature matrix are generally closer to the mixed-in proportion resulting in a lower RMSE. For example, CIBERSORT tends to overestimate T cell populations when count-based signatures are used compared to rank-based. Rank- and count-based centroids provide comparable estimates for all cell-types, with higher correlations in T- and B-cells, and slightly lower correlation for monocytes. Rank-based signatures produced more accurate cell proportion estimates in a second evaluation in which we used scRNA-seq to construct synthetic bulk head-and-neck tumors, with B-cells and monocytes very poorly estimated by count-based signatures and recovered with much higher correlations and lower RMSEs by the rank-based signatures (Fig 2.3). In addition, we tested the use of rank-based centroids derived from the multiple PBMC datasets, in place of the lymphocyte-related LM22 signatures originally published with CIBERSORT, for deconvoluting synthetic PBMC mixtures. In the case of the LM22 comparison, ranked centroids performed comparably with count-based versions and constructing ranked signatures using multiple datasets did not influence the usefulness of the signatures in deconvolution in this case (Supplementary Fig. S2). Deconvolution for all cell lines used to construct the synthetic bulk were comparable (Supplementary Fig. S3). We observe that T-cells were overestimated using the count-based single PBMC signature matrix while T-cell estimation was slightly underestimated using a rank-based signature matrix with improved RMSE (Fig. S2-S3B-D). The plots suggest some systematic biases for over-estimating some cell types (e.g. monocytes in the melanoma sample) at the expense of under-estimating others (e.g. B-cells in the same sample) likely a result of some mismatch in PBMC vs infiltrating immune transcriptional signatures. Overall, these tests illus-

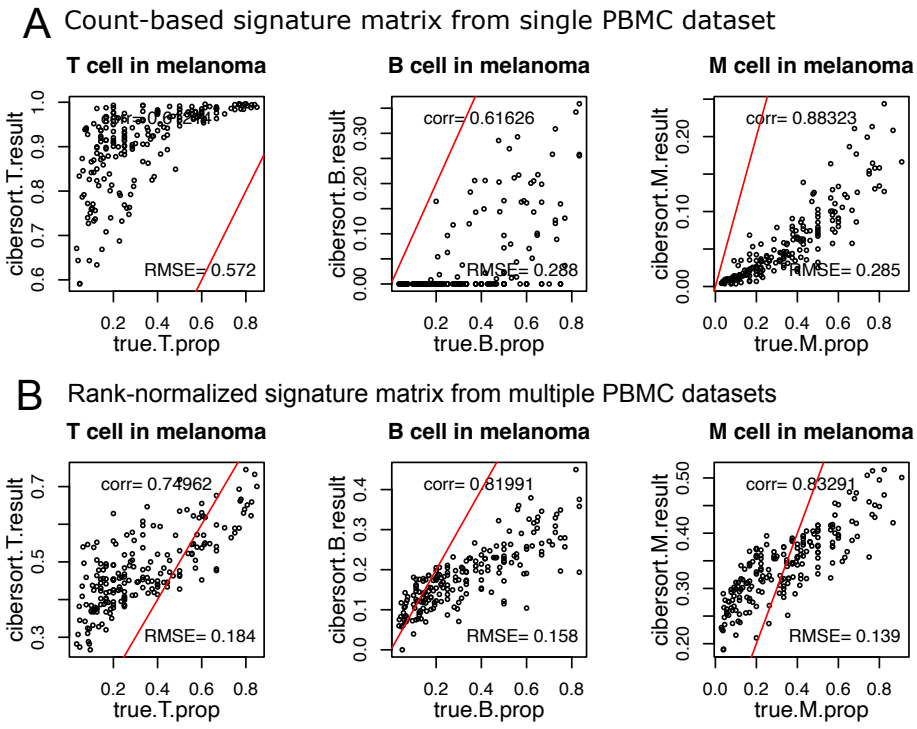

Figure 2.5: **Validation of scBeacon workflow in synthetic mixtures of a scRNA-Seq melanoma dataset (a)** Correlation between the true mixture proportion of in silico mixtures from a scRNA-Seq melanoma dataset to the deconvolution estimates of using a count-based signature matrix from a single PBMC scRNA-Seq dataset (10X-v2). Red line marks the correct estimate (x=y). Cell type ratios are normalized to sum up to 1. **(b)** Same as A, but using a rank-normalized signature matrix from the combination of multiple PBMC scRNA-Seq datasets: all PBMC datasets from Fig. 1B-D, except Smart-seq2: 10X chemistry v1-v3, CEL-Seq2, Drop-Seq, inDrops, Seq-Well. (RMSE = Root Mean Square Error, corr = pearson correlation).

trate that using ranked centroids to derive signatures for deconvolution provides much lower RMSEs and maintains high correlations between predicted and known cell-type proportions.

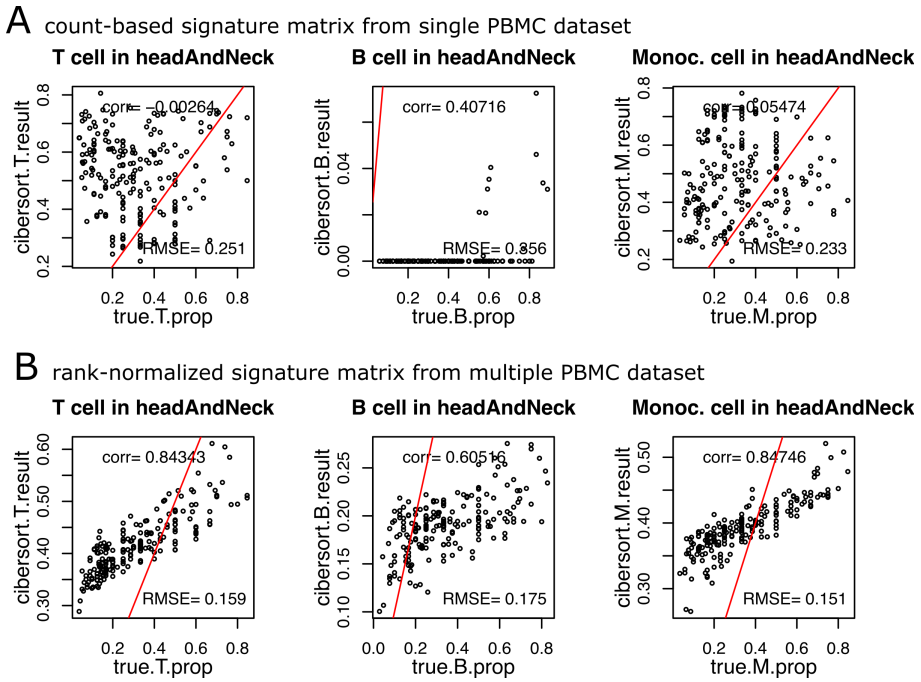

Figure 2.6: **Validation of scBeacon workflow in synthetic mixtures of a scRNA-Seq head and neck cancer dataset (a)** Correlation between the true mixture proportion in synthetic mixtures from a scRNA-Seq head and neck cancer dataset and the deconvolution results of using a count-based signature matrix from a single PBMC scRNA-Seq dataset (10X-v2). Red line marks the correct estimate (x=y). Cell type ratios are normalized to sum up to 1. **(b)** Same as A, but using a rank-normalized signature matrix from the combination of multiple PBMC scRNA-Seq datasets (all PBMC datasets from Fig. 1B-D, except Smart-seq2: 10X chemistry v1-v3, CEL-Seq2, Drop-Seq, inDrops, Seq-Well).

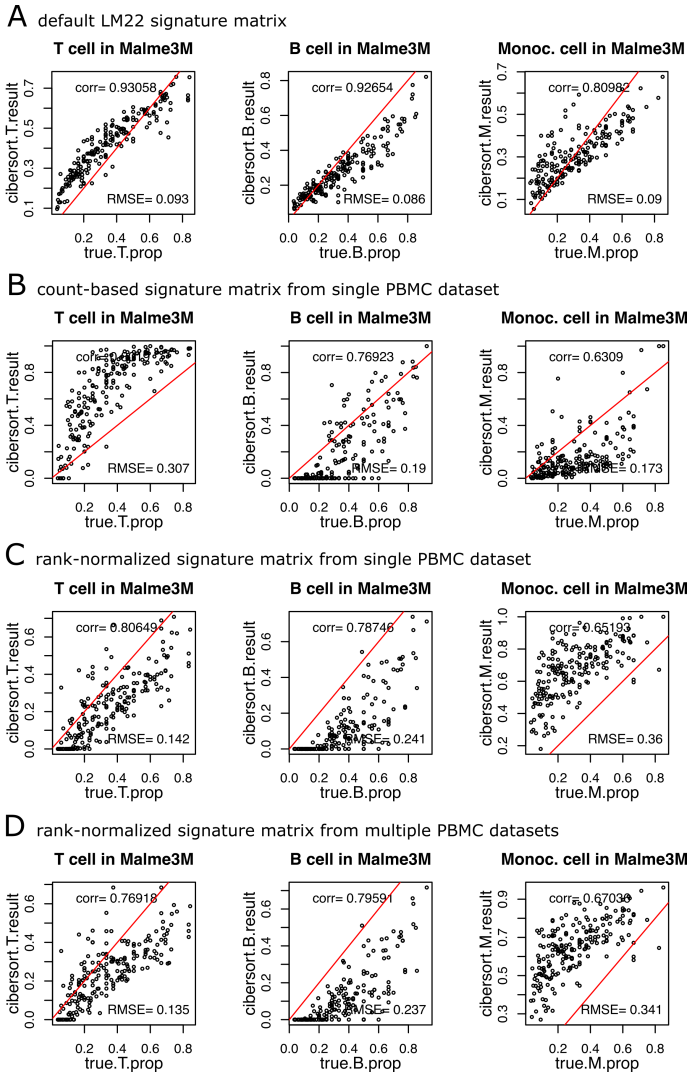

Figure 2.7: **Validation of scBeacon workflow in synthetic mixtures from one sorted bulk RNA-Seq datasets (a)** Correlation between the true mixture proportion in synthetic mixtures from bulk RNA-Seq and the deconvolution results of using Cibersort's default signature matrix, LM22, and summing more specific deconvolution results into T cells (T), B cells (B), and Monocytes (M). The synthetic mixtures are built from bulk RNA-Seq of the cancer cell line Malme3M and bulk RNA-Seq of purified immune cells. Red line marks the correct estimate (x=y). Cell type ratios are normalized to sum up to 1. **(b)** Same as A, but using a count-based signature matrix from a single PBMC scRNA-Seq data set (10X-v2). (C) Same as A, but using a rank-normalized signature matrix from a single PBMC scRNA-Seq data set (10X-v2). (D) Same as A, but using a rank-normalized signature matrix from the combination of multiple PBMC scRNA-Seq data sets (all PBMC data sets from Fig. 1B-D, except Smart-seq2: 10X chemistry v1-v3, CEL-Seq2, Drop-Seq, inDrops, Seq-Well).

2.3.2 scBeacon clusters and signatures from EBI's Single-Cell Expression Atlas (SCEA): Building a comprehensive single-cell derived cell type signature library

EBI's Single Cell Expression Atlas (SCEA) is a public single-cell RNA sequencing data consortium that hosts datasets from published studies for six different species8. For this analysis, we downloaded 62 homo sapiens scRNA-Seq datasets available in February 2020. The datasets cover a wide range of healthy and diseased tissues, consisting of numerous cell types in the human body, and were processed with different single-cell sequencing technologies. SCEA serves as the fundamental data resource for this project to build a comprehensive collection of human cell-type signatures, which is then used for bulk tumor deconvolution.

Clusters were extracted for each SCEA dataset, producing a total of 585 clusters. Centroids and rank centroids were calculated for each of these and used as the clusters' signatures. Clusters were linked if their RTKE metric was above 77.39 (top 10% percentile) and then clustered into metaclusters using the Louvain algorithm with default Seurat settings. The RTKE threshold and Louvain method were found to obtain the highest Silhouette scores out of a series of thresholds and several clustering methods (including K-means, Hierarchical, and a graph-based iGraph methods). Louvain clustering produced 217 metaclusters from which exemplars were defined. Exemplar signatures were created from the average rank signatures of clusters assigned to a metacluster and using only the top 20% of differentially ranked genes (see Methods; Fig. 3A). The 217 metaclusters were annotated using both author's published annotations and marker genes based enrichment test, the full annotation is available.

We found several examples in which multiple datasets contributed to the definitions of a single exemplar. Overall, 16 (7.4%) of the exemplars were implicated by two or more datasets. Even so, the map contains a majority of singleton exemplars 141 (65.0%) – those metaclusters containing a single cluster from a single dataset. Altering the metaclustering parameters would give different numbers of clusters and singletons. However, we found the chosen setting to maximize a Silhouette score struck a good balance as even singletons that were "close" to one another in the map had deconvolution results across the TCGA that were just as distinct from each other as those that were "far" apart, justifying maintaining them as separate signatures for our use. We also note that a few centroids combine clusters from different datasets that probed distinctly different human tissues. These centroids could represent a common cell type found in many tissues, as is the case with immune cell types.

We queried the scBeacon collection of exemplars to determine the extent to which they reflected distinct cell types. First, we investigated the distribution of cell types expected to be highly similar based on the expression of a particular known tissue-specific marker gene. To that end, we queried the map for all centroids with high expression of the insulin gene to identify pancreatic-associated clusters. Meta-cluster X85 contains several such pancreatic clusters (Fig. 3B) that were derived from three different datasets that all assayed different states of pancreatic tissue (Fig. 3C). We also queried three immune cells using marker genes, CD3E for T cells, MS4A1 for B cells and CD14 for monocyte (Supplementary Fig. S8).

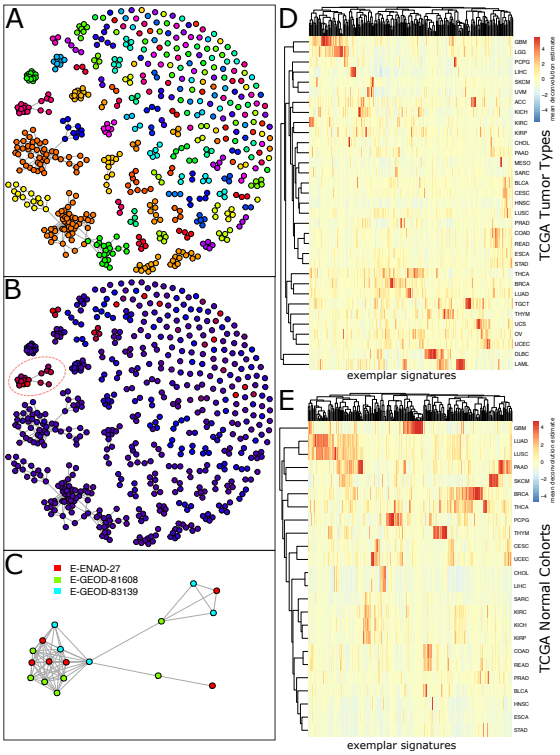

Figure 2.8: **217 exemplars of cell types and states identified from single-cell RNA-seq datasets in the Single Cell Expression Atlas (SCEA) (A)** Distinct cell types were identified by comparing clusters of single cells with similar expression profiles across multiple datasets. scBeacon clusters are colored by grouping into exemplars representing distinct cell types/states. Nodes represent 585 clusters of single cells derived from clustering individual datasets in the SCEA collection. To determine a non-redundant set of cell states/types from these dataset-derived clusters, clusters were connected to each other, linking clusters found in possibly separate datasets. **(B)** To reveal pancreas-related cell-type clusters, clusters in A are colored based on INS (insulin marker gene) gene expression (low expression, blue; high expression, red). Exemplar X85's centroid (circled) had a high level of INS expression, implicating an insulin system role for its represented cell type. **(C)** Detailed view of the X85 exemplar illustrating it was derived from 18 different clusters (nodes) contributed by three different pancreas-related SCEA datasets (colors of the nodes), 12 clusters of which are highly mutually similar and make up the core of the exemplar. **(D)** CIBERSORT estimation of 217 exemplars on TCGA bulk tumor samples. Columns: 217 exemplar deconvolution estimation. Row: averaged across the samples within each of the 33 TCGA cancer types. E. Same as part D but for CIBERSORT deconvolution of TCGA normals using the same set of 217 exemplars.

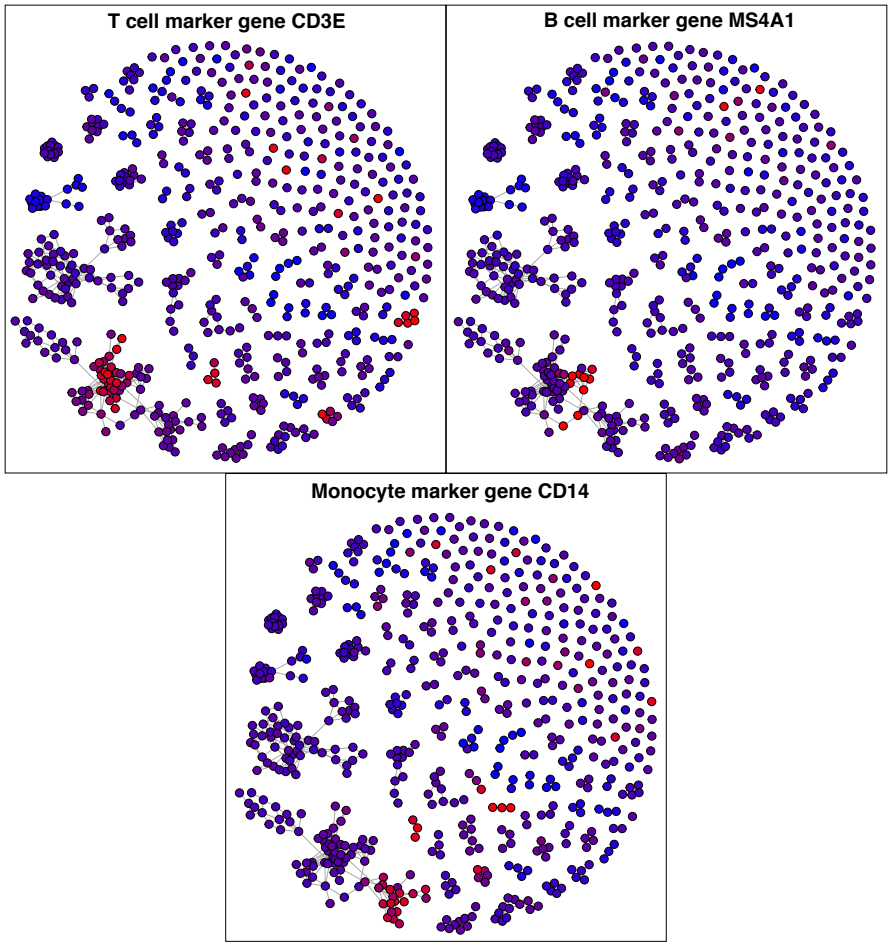

Figure 2.9: **Immunomarkers in scBeacon clusters** scBeacon centroids(colored circles) colored by T cell(CD3E), B cell(MS4A1) and monocyte(CD14) marker genes (low expression, blue; high expression, red).

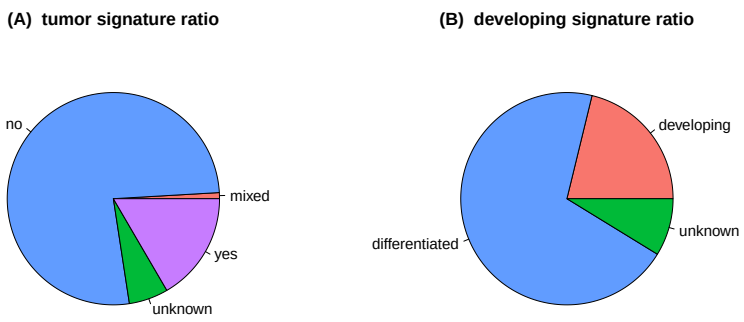

Figure 2.10: **Tumor and developing signature ratios in scBeacon. (A)** Tumor signatures number and ratio(36, 16.6%), non-tumor signatures(166, 76.5%), mixed with tumor and none-tumor signatures(2, 0.92%), unknown(13, 6.0%), non-tumor signatures could come from datasets of other diseases other than tumor. **B** Developing signatures number and ratio(46, 21.1%), differentiated signatures(152, 70.0%), unknown(19, 8.8%).

2.4 Discussion

There is ever-growing evidence that the cell types present in a tumor's microenvironment influences the outcome of a cancer patient1. In recent years since single-cell sequencing became available, the characterization of various cell types in the human body has improved immensely38–40. A growing number of public single-cell sequencing datasets provides a more accurate and comprehensive definition of the human cell type repertoire. However, there are still challenges to efficiently integrate and analyze those datasets together. First, due to the high level of technical noise and systematic differences between sequencing platforms, simple concatenation could result in batch effects that become the dominant variance rather than biology. Batch effects have been shown to cause an increased number of false positives in downstream analyses41. To reduce the chance of false discoveries, integration of multiple datasets must eliminate

batch effects42. Whole reference atlas initiatives such as the Human Cell Atlas (HCA) started collaborative projects to integrate as many datasets as possible to create a whole human cell type map, the data integration process for this task should not only be able to handle batch effects well, but also be computationally efficient and fast while ingesting and integrating datasets.

We introduced a single cell RNAseq pipeline called scBeacon that clusters and integrates datasets to identify single cell signatures useful for the deconvolution of bulk cancer samples. Unsupervised clustering of full transcriptome data has been used to identify subsets of related samples or genes for years since the establishment of DNA microarrays43,44. Since then, clustering has only increased in importance for the analysis of bulk and later scRNA-seq datasets45. Computational algorithms leverage an ever increasing number of samples of scRNA-seq datasets using approaches like community detection46 and later deep-learning autoencoders47. In our approach, we assume many of the clusters represent a collection of cells with highly similar transcriptomes that concentrate distinct cell types. Given this assumption, "marker genes" of a cell type/state or lineage may be approximated with the cluster centroids. Our pipeline infers cell types using multiple datasets by using a novel enrichment-based test to determine when clusters from different scRNAseq datasets are highly similar.

We validated scBeacon's deconvolution using in silico mixtures from single-cell and sorted bulk RNA-Seq data. We used the EBI Single-Cell Expression Atlas (SCEA) as a database to create a comprehensive set of cell-type signatures with a new enrichment-based similarity test, the reciprocal top-k enrichment (RTKE) test.

Several methods have been created to help biologists search these collections to find cell types of interest. Scmap48 implemented a fast approximate k-nearest-neighbor search with

cosine distance to project cells in scRNA-seq datasets to reference databases. CellBlast49 built a robust data query method in scRNA-seq database based on a neural network-based generative model and a customized cell-to-cell similarity metric. CellAtlasSearch50 used locality-sensitive hashing (LSH) Hamming distance for bulk and single cell RNA-seq data processing and query. We found the RTKE test to be robust to the comparison of cluster centroids across datasets and scRNA-seq platforms.

We found that the use of rank-based cell type signatures for the deconvolution of bulk RNA-Seq data compared to count-based cell type signatures is effective for forming signatures from multiple data sources. The rank-normalization and combination of multiple datasets did not impact the accuracy of deconvolution and sometimes even improves the inference. Thus, our rank-based approach offers a promising and simple strategy for the ongoing derivation of a comprehensive set of cell type signatures from an expanding collection of scRNAseq datasets.

Chapter 3

Single-cell signatures identify microenvironment factors in tumors associated with patient outcomes

The project described in this chapter was submitted to Cell Reports Methods with Chapter 2. I worked with Verena Friedl to develop and implement this project.

3.1 Introduction

Cancer is a disease involving the interplay of many cell types[21]. Tumor cells are surrounded by a microenvironment of various types of cells, such as stromal and blood cells. Characterizing the composition and spatial arrangement of human cell types embedded in the tumor microenvironment is a relatively new direction in cancer biology research. Most no-

tably, immune infiltration has been a focus in recent years for the implications of emerging and promising immunotherapies that have been shown to depend on the presence of certain immune cell types and states[53]. However, studies have shown that additional cell types and molecular characters beyond immune cells play an important role in tumor character and response to treatment and patient outcomes[2, 19, 34]. Therefore, it is important to detect and quantify a full profile of cell types to improve our understanding and treatment of cancer.

Characterizing the tumor cell types has been largely limited by the low number of known cell type signatures. Most studies have focused nearly exclusively on immune-associated cell types. Leveraging cell type signatures derived from newly available single-cell RNA sequencing (scRNA-seq) presents an opportunity to broaden the detection of cell types. scRNA-Seq[44] has transformed biological research by making it possible to determine gene expression separately for each cell in a biological sample. The technology provides a higher definition of cell types and cell states and has already expanded the catalog of known cell types[47]. Advances in sequencing technology have facilitated an explosion of the availability of scRNA-seq datasets supported by databases such as the Single Cell Expression Atlas[31] and the Human Cell Atlas[39]. Those large databases are great resources of cell-type transcriptomes.

Over the years there have been several bioinformatics tools developed to deconvolute bulk tumors with cell type specific gene expression profiles derived from scRNA-seq data. CIBERSORT(X) is the most widely used cell type deconvolution tool based on support vector regression. BSeq-SC applied scRNA-seq derived cell type signatures to deconvolute bulk tissues using CIBERSORT, and discovered subpopulations and heterogeneity within pancreatic cell types[4]. MuSiC deoconvolutes bulk RNA-seq samples using cell type reference generated

from hierarchical clustering on multi-subject scRNA-seq data using weighted non-negative least squares(NNLS)[49], along with DeconvSeq utilizes a generalized linear model for cell type ratio estimation[13], Bisque uses NNLS regression[24], and BayesPrism[10], BLADE[1] implements probabilistic model (multinomial) to deconvolute bulk RNA-seq data using scRNA-seq derived gene expression profile. These methods rely on a cell type signature matrix from only one scRNA-seq dataset that has been pre-annotated, which limits the number of datasets used for bulk tissue deconvolution. With the increasing number of scRNA-seq datasets available and large scRNA-seq consortiums being built, strictly supervised deconvolution approaches could limit the opportunity to discover new cell types and a comprehensive characterization of bulk tissues.

I used 217 cell-type signatures using scBeacon described in Chapter 1 and used them to quantify cell types in bulk tumor specimens from the TCGA RNA-Seq compendium. We find dozens of expected and novel associations between cell types and tumor types in the TCGA collection, with implications for synergistic and antagonist interactions between cell types based on the co-occurrence or mutual exclusivity of cell type groups. Some cell type signatures were found to be significantly associated with patient outcomes in several tested tumor types, many of which are independent of published cancer subtypes and thus provide a new independent measure of disease state.

To provide a comprehensive view of the relationship of all TCGA samples to each other based on their inferred microenvironment contents, we developed an interactive tumor cell-type (TCT) map that uses the inferred exemplar estimates to arrange the samples in one layout. The two dimensional projection of TCGA samples on the Tumor Map20 revealed sev-

eral unexpected cluster associations, several with implications about patient survival.

3.2 Methods

3.2.1 Deconvolution to identify cell types in bulk tumors

The exemplar cell-type signatures generated from the scBeacon workflow were used for deconvolution of cancer bulk RNA-Seq data, in which each signature's contribution to the mixture was estimated. We used the Cibersort deconvolution method55, which performed well in the DREAM deconvolution competition56. We ran Cibersort with parameters: perm=100, QN=FALSE, absolute=TRUE, abs_method='no.sumto1'.

We used rank-normalized cell type signatures in CIBERSORT to deconvolute TCGA bulk tumors. Compared to cell type signatures derived from count-based expression values, rank-normalized signatures outperformed count-based signatures in bulk tumor deconvolution, which is commonly used in other deconvolution approaches. This was validated by our validation analysis using synthetic bulk samples.

In this study, we used the TCGA collection as the bulk tumor data for deconvolution. We downloaded the counts per tumor type data from Xena57, which represents The Cancer Genome Atlas (TCGA) gene expression HTSeq counts data originally provided by the NCI's Genomic Data Commons. We normalized the count data to TPM (transcripts per million reads).

3.2.2 Bimodality test to associate a cell type signature with a patient cohort

After obtaining the CIBERSORT deconvolution results on TCGA cancer samples, we analyze if the presence of the cell-type signatures in tumors correlates with the survival outcomes of patients. First, we define patient groups based on how much a signature is detected in the patients' tumor samples. For each signature in each tumor type, samples that have a relatively high proportion of the signature detected are defined as "patients-up group" and samples that have a relatively lower proportion of the signature detected are defined as "patients-down group".

To formalize this separation of samples in the deconvolution results, we applied a bimodality test for each signature, based on the student-t distribution65 implemented in the t-Student Mixture Models Module (SMM) library66 in python. It models data by a mixture of t-Student distributions, estimating the parameters with Expectation-Maximization, and uses the Bayesian information criterion(BIC) to decide whether the current model fits the proposed data. Signatures that fit the student-t bimodal distribution are kept for survival analysis since they represent a meaningful separation between patient groups. From the two distributions identified in the model, we define sample groups: samples that have a cell type estimate higher than the upper mean are labeled as "patients-up", samples that have a cell type estimate lower than the mean of the lower distribution are labeled as "patients-down" (Fig. 5B). For signatures that don't fit the student-t bimodal distribution, the patients are separated by the median. However, in cases where a signature had estimates of zero in more than 50% of the tumor-type samples, all samples with an estimate of zero were assigned to the "patients-down" group and all samples

with an estimate above zero were assigned to the “patients-up” group. A signature was excluded from survival analysis in a tumor type if less than 10 samples had an estimate above zero.

3.2.3 Survival Analysis to Associate Cell Type Signatures with Patient Outcomes

TCGA survival information was downloaded from the Xena portal67 . We used progression-free interval (PFI) to measure disease progression, except for Acute Myeloid Leukemia (LAML) patients, which only have Overall Survival (OS) available.

To measure the separation between the two sample groups, we used the R package “survminer” for Kaplan-Meier survival analysis, and applied Cox proportional hazards (CoxPH) model68 by using R package “survival”69. Reported hazard ratios (HR) were extracted from the CoxPH model and all p-values for survival analysis were, unless stated otherwise, p-values of the log-rank test. We report the results for a ’naive’ signature outcome separation (SOS), which is a univariate CoxPH model.

In Supplementary Table 3 we curated subtype annotations for all TCGA tumor types, mostly from the TCGA PanCanAtlas project21 and TCGAbiolinks70, except for DLBCL (diffuse large B-cell lymphoma), which had no subtype information available. Subtype information was used as a covariate in multivariate CoxPH models per tumor type in order to correct a potential imbalance in subtypes, and avoid recapitulating known cancer subtypes by the separation of the patients groups.

To understand how the 217 cell type signatures separate the patients survival, we used Benjamini Hochberg multi-test corrected p values from the survival analysis, and focused on the ones that have corrected p value lower than 0.05. We also extracted hazard ratio from the

models, the hazard ratio greater than 1 indicates the event hazard increases and thus the length of survival decreases. When hazard ratio smaller than 1 indicates the cell type variant positively influences the patients' survival length.

3.2.4 Tumor Cell-Type (TCT) Map to Identify New Pancancer Connections

(Building the Map) The two-dimensional layout of the Tumor Microenvironment (TCT) map was created by providing the matrix of the 217 exemplar CIBERSORT estimates for each of the 11,057 TCGA samples to the DrL layout engine of the UCSC TumorMap tool20. The interactive TCT map is available online (bit.ly/TCTmap_217exemplars). The interactive map includes attributes for browsing various results of our analysis including exemplar estimates, TCGA disease categories, TCGA disease subtype categories (Table S5), and TCGA PancanAtlas clustering solutions21.

(Clustering the Samples on the TCT Map) The samples on the TCT map were clustered by their two-dimensional coordinates using hdbscan, a spatial hierarchical clustering method33, with a minimum cluster size of 20. This resulted in 49 sample clusters (Supplementary Fig. S20). Additionally, 1,277 samples were not assigned a TCT map cluster.

To measure the novelty of the resulting clustering solution, we first measured the similarity between the spatial TCT map clusters and the grouping by disease subtype using the adjusted rand index. Additionally, we measured the similarity to the PancanAtlas mRNA-based TumorMap21. This TumorMap provides a similar comprehensive look at the same set of TCGA samples, and it is based on mRNA data, which is also the basis of our exemplar estimates. Therefore, we can now determine if any grouping we find on the TCT map is only a

recapitulation of known subtype biology or gene expression, or if it is newly determined by our exemplar estimates.

We applied the same spatial hdbscan clustering method to the PancanAtlas mRNA TumorMap with a minimum cluster size of 50 samples in order to reach a similar number of resulting clusters (Supplementary Fig.21). The samples on the PancanAtlas TumorMap were assigned to 41 clusters and 1,123 samples were not assigned a cluster. We then measured the similarity of the two spatial clustering solutions using the adjusted rand index.

(Finding Survival Differences on the TCT Map) We applied survival analysis on TCT map cluster groupings of the TCGA samples analogous to the approach we described previously. First, we analyzed survival in the context of each disease. We defined the main clusters of each disease as any cluster containing 5 or more samples of that disease. Then, we applied CoxPH models between pairs of sample clusters of the same disease, comparing each cluster to the largest cluster, i.e. the cluster with the most samples of that disease. We again provided the disease subtype information curated in Table S3 to the CoxPH models as a covariate in order to correct for a potential imbalance in subtypes. Second, we repeated the same survival analysis in each disease subtype, eliminating the need to provide a subtype covariate and determining which subtypes contributed to the overall findings per disease. The disease and subtype level results are presented in Table S4. Additionally, Table S4 lists the most differential exemplars between each cluster and the largest cluster in each disease and each subtype. We determined the three highest, and the three lowest exemplars in each comparison using a student's t-test.

3.3 Results

3.3.1 Deconvolution of TCGA samples using scBeacon signatures

The meta-clusters from the human SCEA were further processed with the scBeacon workflow (Fig.1; see Methods) to create a signature matrix for use in deconvolution (Supplementary Fig. S9). We used CIBERSORT to deconvolute the bulk RNA-seq samples available for 33 different tumor types from The Cancer Genome Atlas (TCGA)21 using the signatures matrix derived from the 217 cell-type exemplars (Fig. 3D-E). As expected, many cell-type signatures are undetected within most tumor samples, reflecting a degree of specificity to the signatures and their use in deconvolution. Assuming that a signature was "detected" in a sample if it had a CIBERSORT score of 0.01 or greater (i.e. it was estimated to account for 1% of the expression among all detected signatures for a particular sample after 0 - 1 normalization), then 83.4% of the signatures (n=181) were detected in at least one sample but less than 50% of all samples. On the other hand, a small number of signatures (n=2) were detected in over 90% of the samples. Finally, 10 signatures were detected at levels of 1% or less in any of the samples. These lowly-estimated signatures could represent cell-types absent from the current TCGA collection among several possibilities. Still, the vast majority of the SCEA signatures (207, 95%) were detectable in at least some of the samples.

Tumor types that arise in similar tissues of the body had similar deconvolution profiles (Figure 3D). For example, the estimated cell-type profile for COAD (colon adenocarcinoma) is most similar to the estimated cell-type profile of READ (rectum adenocarcinoma). Likewise, LIHC (liver hepatocellular carcinoma) and CHOL (cholangiocarcinoma) clustered together, as

well as GBM (glioblastoma multiforme) and LGG (brain lower grade glioma), and a group of squamous cell carcinomas (HNSC = head and neck squamous cell carcinoma, LUSC = lung squamous cell carcinoma, BLCA = bladder urothelial carcinoma), CESC = cervical squamous cell carcinoma and endocervical adenocarcinoma). These results suggest that tumors arising from related tissues in the body share a similar microenvironment makeup compared to tumors arising from different tissues. Indeed, when we repeat the cell type analysis using deconvolution on normal tissue (using the TCGA matched normal samples), we again find that tissues cluster together based on their cell type profiles (Figure 3E).

To confirm this result and to validate the scBEACON procedure for identifying exemplars using a positive control test case, we repeated the entire analysis using normal samples from the GTEx consortium from which exemplars were derived from the published single-nucleus RNA sequencing (snRNA-seq) dataset 22 and deconvolution was performed on samples from the bulk GTEx RNA-seq dataset 23. We found that similar tissues of related organ systems clustered together based on their GTEx exemplar deconvolution scores; e.g. brain cerebellum with another cerebellum, colon with small intestine with stomach, several arteries clustered together, and so on (Supplementary Fig. S10). Expected cell-types were again found with high deconvolution scores in GTEx tissues (Supplementary Fig. S11C-E). Slightly more than half of the signatures in GTEx (19 out of 35; 54%) had high correlations (Pearson ¿ 0.5) with at least one signature in scBeacon's SCEA-derived set. Thus, we estimate another 16 signatures from GTEx could have been included to the collection of the EBI 217, consisting of a marginal increase in cell-type representation (7.4%). On the other hand, the EBI collection captured many signatures not represented in GTEx (164 out of the 217 had correlations below 0.50 for any-

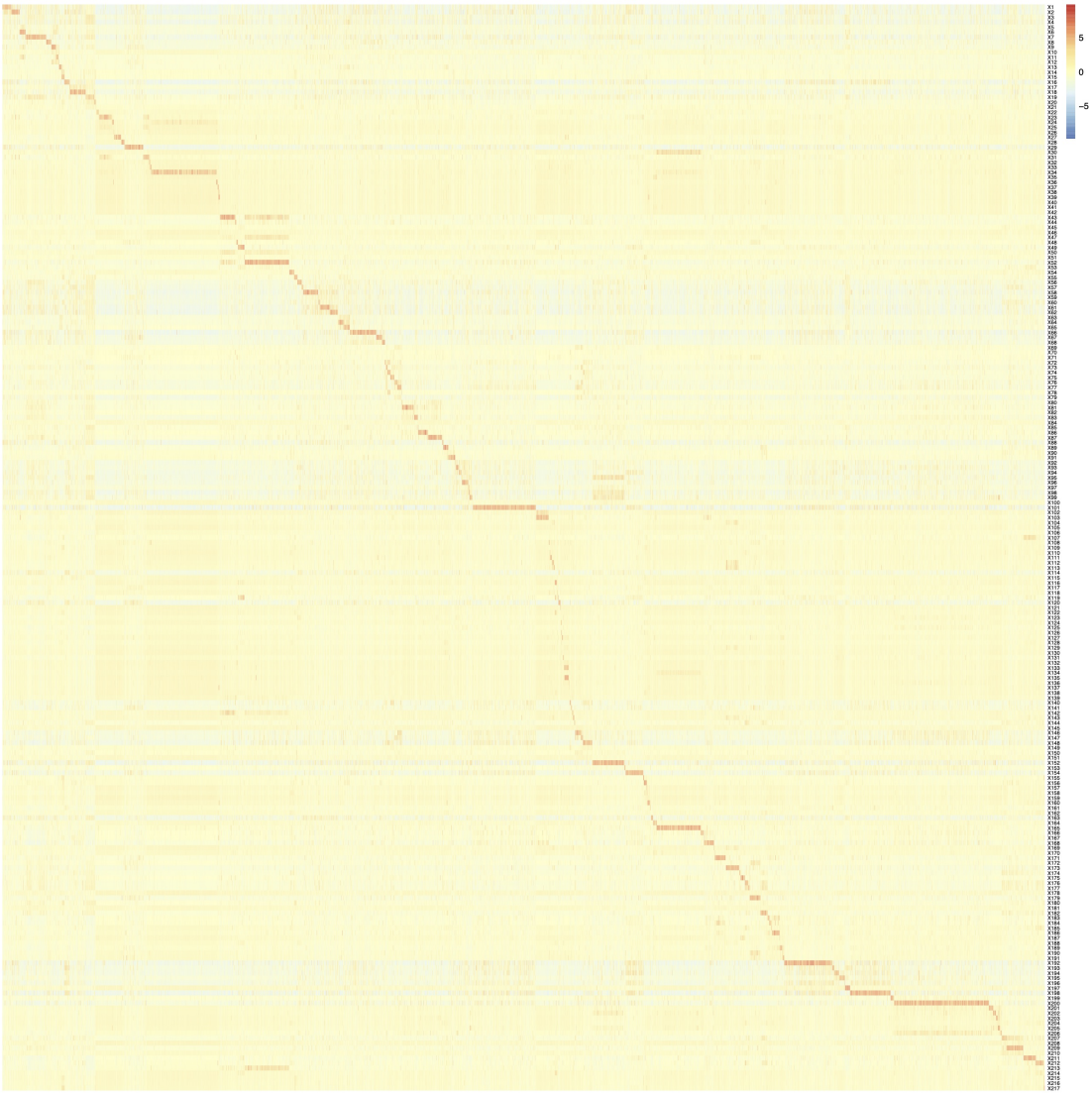

Figure 3.1: **Signature matrix for meta-cluster exemplars** The heatmap shows the average gene expression in the selected 3988 signature genes. The middle redline shows the differential expressed genes for each exemplar that separate it from others.

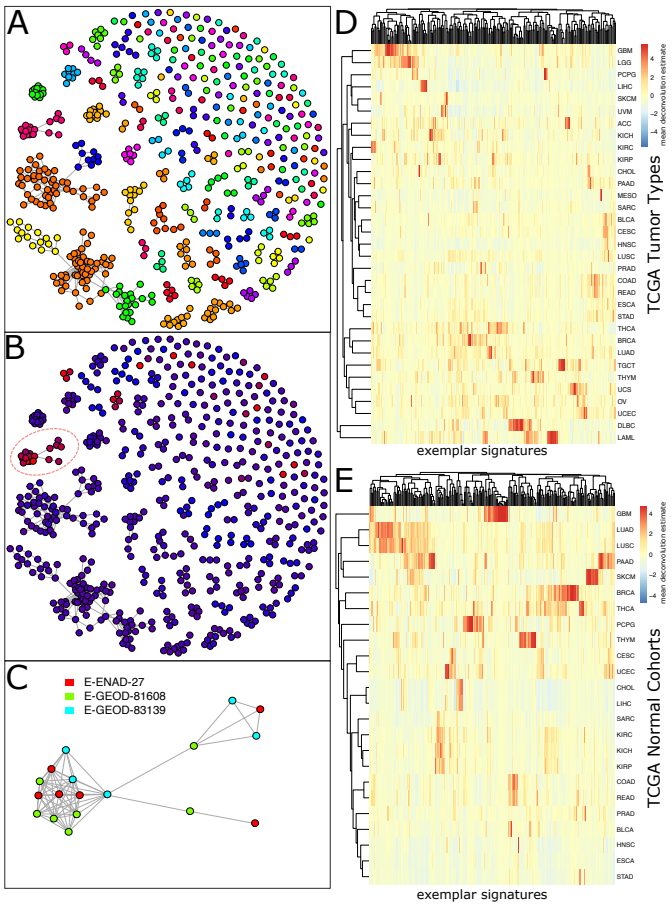

Figure 3.2: **217 exemplars of cell types and states identified from single-cell RNA-seq datasets in the Single Cell Expression Atlas (SCEA) (A)** Distinct cell types were identified by comparing clusters of single cells with similar expression profiles across multiple datasets. scBeacon clusters are colored by their grouping into exemplars representing distinct cell types/states. Nodes represent 585 clusters of single-cells derived from clustering individual datasets found in the SCEA collection. To determine a non-redundant set of cell states/types from these dataset-derived clusters, clusters were connected to each other, linking clusters found in possibly separate datasets. **(B)** To reveal pancreas-related cell-type clusters, clusters in A are colored based on INS (insulin marker gene) gene expression (low expression, blue; high expression, red). Exemplar X85's centroid (circled) had a high level of INS expression, implicating an insulin system role for its represented cell type. **(C)** Detailed view of the X85 exemplar illustrating it was derived from 18 different clusters (nodes) contributed by three different pancreas-related SCEA datasets (colors of the nodes), 12 clusters of which are highly mutually similar and make up the core of the exemplar. **(D)** CIBERSORT estimation of 217 exemplars on TCGA bulk tumor samples. Columns: 217 exemplar deconvolution estimation. Row: averaged across the samples within each of the 33 TCGA cancer types. E. Same as part D but for CIBERSORT deconvolution of TCGA normals using the same set of 217 exemplars.

thing present among the GTEx signatures) and thus provides a 3.5-fold increase over what is represented in the GTEx collection. In summary, the metaclustering procedure for identifying exemplars from scRNA-seq cluster signatures, as well as their use to identify them in bulk samples via deconvolution, was reproducible using a completely orthogonal dataset in a scenario where the signatures and deconvolution results were well-annotated. In addition, the resulting GTEx signatures compared well to what was found and represented in the scBeacon collection based on SCEA, even though the GTEx signatures were derived from nuclei transcriptomes.

3.3.2 Single cell exemplar signatures deconvolve appropriate bulk tumors but with lower scores compared to their normal counterparts

We measured the degree to which deconvolution with exemplars, derived from a particular tissue, could "detect" the presence of a cell type in bulk tumor (or normal) samples from TCGA when using a tumor type of that same tissue. To quantify and visualize exemplar specificity, we used the CIBERSORT deconvolution results that considered all 217 exemplars to compare the estimates obtained in related to unrelated tissues. We selected three tissues – breast, lung, and brain – for which exemplars were annotated as either derived from normal or cancerous tissue. We collected the CIBERSORT estimates and aggregated them as either related to the exemplar's tissue or unrelated. For example, X10 (myoepithelial cell of mammary gland) was used as the normal breast exemplar while X62(B cells from lymph node in breast carcinoma patients) was used as the cancerous breast adenocarcinoma (BRCA) exemplar since these signatures had the highest CIBERSORT scores in normal breast and cancerous breast, respectively, among all other signatures annotated as breast-related (Supplementary Fig.

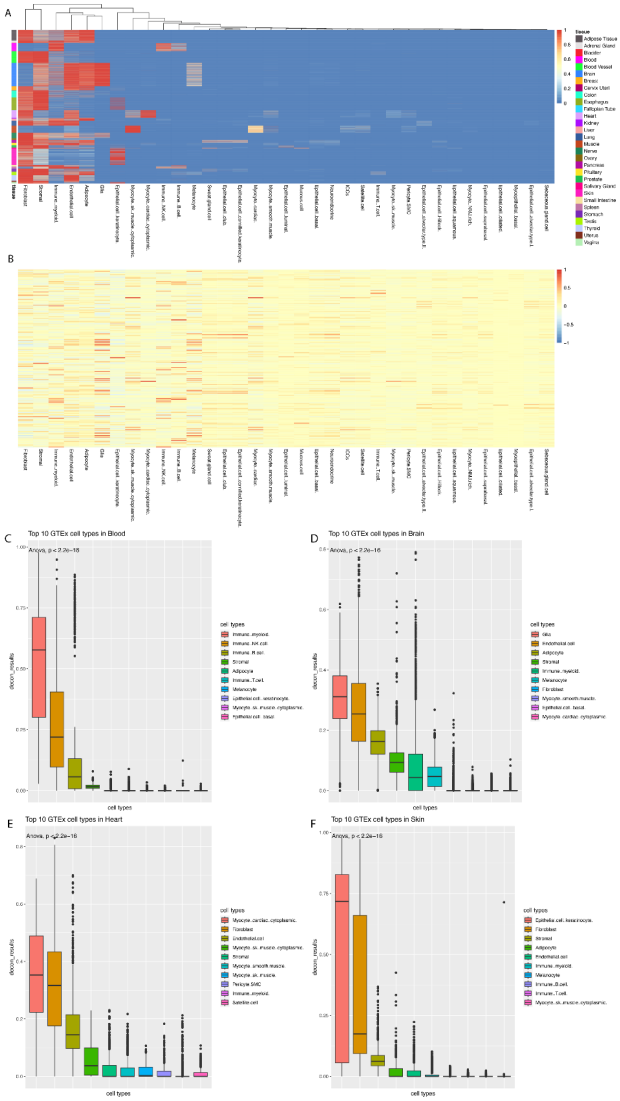

Figure 3.3: **scBeacon deconvolution using GTEx snRNA-seq and bulk RNA-seq data, compared to SCEA-derived 217 signatures GTEx deconvolution results (A)** Overview of deconvolution of bulk samples from the GTEx Consortium dataset 23 in which cell type signatures were derived from the GTEx single-nucleus RNA sequencing dataset 22, cell type annotation for single cells, and tissue annotation for bulk samples were taken from curated metadata. **(B)** Pearson correlation between GTEx snRNA-seq deconvolution results and SCEA-derived 217 deconvolution results on GTEx bulk samples. With rows as SCEA-217 signatures and columns as GTEx snRNA-seq cell type signatures. **(C,D,E,F)** The top 10 GTEx cell types in blood, brain, heart, and skin.

S12A-B). Exemplars for the other two tissues were chosen using the same criteria (Supplementary Fig. S12C-F). The 113 normal samples of the TCGA BRCA cohort showed significantly higher CIBERSORT scores for X10 than normal samples in other TCGA cohorts (P ¡ 2.2e-16, Kruskal-Wallis test; Fig.4A, left panel). Similarly, the 1104 tumor samples of the TCGA BRCA cohort showed higher scores for X62 than tumor samples in other cohorts (P¡2.2e-16, Kruskal-Wallis test; Fig.4A, left panel). The same trends were found for both the normal and tumor signatures when the comparisons were repeated in lung (Fig. 4A, center panel) and brain (Fig. 4A, right panel). Thus, exemplars annotated as derived from a specific tissue, and that have the highest match to a particular tissue in TCGA among all other exemplars annotated as derived from that tissue, also were found to be relatively specific for deconvolving that tissue (i.e. they receive the highest CIBERSORT scores among all other tissues). In summary, even when used together with signatures derived from many cell types, deconvolution of TCGA samples using the exemplars results in scores that are consistent at the tissue level.

Cancer signatures had lower CIBERSORT scores than their corresponding normal counterparts for all three tissue types tested (cancer box plots in Fig. 4A). This suggested that cancer signatures reflect a quantitatively lower degree of tissue specificity compared to their normal counterparts. This could be due to patient-specific factors or loss of differentiation fidelity, among other possibilities. To further investigate, we plotted the CIBERSORT scores of both the normal and cancer samples summarized at the TCGA tumor type level (Fig. 4B) and at the level of tumor subtypes (Fig. 4C). The radar plots of all three tissue types investigated reveal that, compared to the normal signatures (Fig. 4B-C, blue radar areas) the cancer signatures (Fig. 4B-C, yellow radar areas) have a reduced relative match to their expected tissues. For

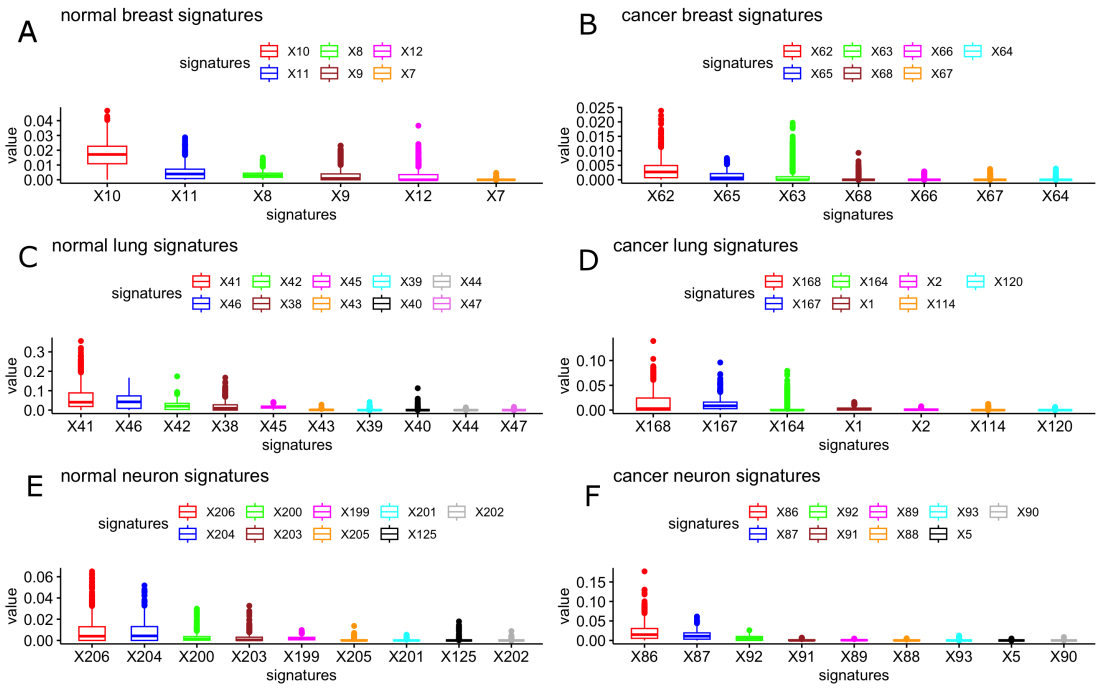

Figure 3.4: **Tissue specific signatures estimation in matching cancer and normal samples** The barplot shows the average CIBERSORT estimation in signatures derived from normal (left) and cancer (right) tissues. **(A)** normal breast signatures in normal breast samples. **(B)** cancer breast signatures in breast cancer samples. **(C)** normal lung signatures in normal lung samples. **(D)** cancer lung signatures in lung cancer samples. **(E)** normal neuron signatures in normal brain samples. **(F)** cancer neuron signatures in brain tumor samples. The signatures in each barplot are ordered by the average CIBERSORT estimation values, from highest to lowest.

the breast and lung signatures, matches apparently similar to cell types in other tissues may explain the relative lower scores; whereas for the GBM signature, the similarity to cell types in brain-related tissue is lower without a concomitant increase in scores to cell types in non-brain tissues. In the case of the breast signature, strong matches to prostate (PRAD) cancer samples appeared to provide a better match than to breast samples when the scores were averaged. However, when the scores were averaged at the subtype level instead of at the cohort level (Fig. 4C), the highest average score matched a HER2 amplified subtype of breast cancer, which represents a minor proportion of the overall BRCA samples, even though matches to several PRAD subtypes also had high scores, the match to HER2 has a higher score than any of the PRAD matching score. The breast cancer signature comes from a mix of two breast cancer subtypes, with 62% HER2 samples and 38% triple-negative samples. In Fig. 4C, the highest average score matched a HER2 amplified subtype of breast cancer, this could be caused by the majority of the samples of the breast cancer signature coming from the HER2 subtype. Alternatively, it could be caused by the different developmental states of luminal versus basal cell types. Basal cells exhibit heightened cancer stem cell activity compared to luminal cells (HER2). Given that triple-negative breast cancer samples primarily comprise basal cells[37] while the HER2 subtype predominantly features luminal cells[35], the HER2 subtype emerges as a more differentiated breast cancer subtype in comparison to others. This might explain why the HER2 subtype is the dominant subtype showing breast tissue-specific signal shown in Fig.4C.

Taken together, exemplar signatures had their highest relative matches in TCGA to samples obtained from the same tissues as the exemplar signatures were obtained. In addition,

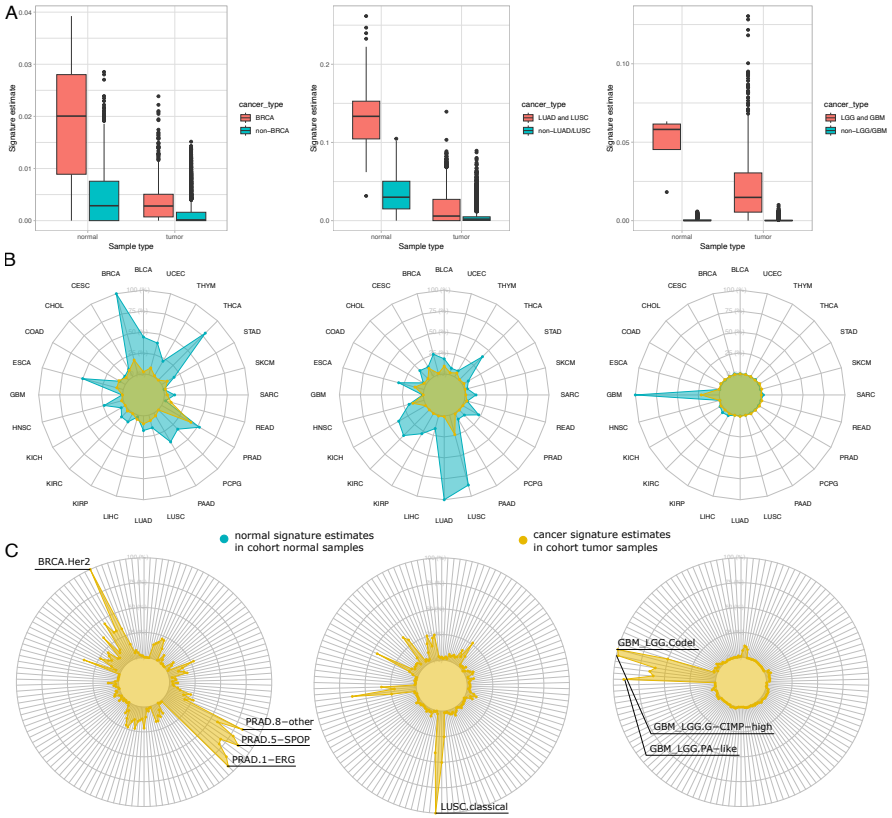

Figure 3.5: **Cell type exemplar signatures are specific to their tissue type for tumor deconvolution (A)** Exemplars were selected if annotated as derived from a tissue common to a TCGA cohort. Normal and cancer exemplars were selected, either from a normal or cancer-derived cluster. Both types of exemplars show specificity to the matching tissue-type in TCGA for all three tumor types inspected including breast cancer (BRCA, left panel, X10 for normal breast, X62 for breast cancer), lung cancer (LUAD and LUSC cohorts, middle panel, X41 for normal lung, X168 for lung cancer) and brain cancer (LGG and GBM, right panel, X206 for normal brain, X86 for brain cancer). The distribution of CIBERSORT estimation scores for samples within the tissue type (pink/left box in each panel) was compared to all estimations for samples outside the tissue type (blue/right box in each panel). **(B)** Radar plots illustrate more detail of the exemplar CIBERSORT deconvolution results in distinct tumor subsets (higher estimates correspond to outer rings) for the same cohorts as in part A (breast cancer BRCA, left panel; lung cancers of LUSC and LUAD, middle panel; brain cancers of GBM and LGG, right panel). Each radar level shows the average CIBERSORT estimate of a cancer-related exemplar for that cancer type (yellow area) or a normal-tissue-specific exemplar for the cancer type (blue area) averaged across all TCGA samples within one of the 33 tumor types. **(C)** Similar to (B) but the CIBERSORT estimates of each exemplar are averaged for 132 different cancer subtypes (spokes around the circle), which group tumors based on shared molecular properties within each of the 33 tumor types.

cancer exemplar signatures exhibited lower relative scores to their tissues on average compared to normal signatures from the same tissue. These findings suggest CIBERSORT maintains its ability to identify the presence of a cell type in a bulk RNA-seq sample using the ranked exemplar signature together with 217 total signatures. Moreover, the results indicate cancer tissue signatures may lose some of the strength of their match relative to normal tissues, which may reflect a loss of differentiation fidelity.

3.3.3 Survival analysis based on deconvolution results: Some cell-type signatures align with patient outcomes in in some tumor types

We next asked whether any of the examplars represented microenvironment determinants that indicate either better or worse outcomes for patients. To that end, we performed survival analysis separately for each cancer cohort using each of the exemplar signatures (see Methods). In total, 6944 exemplar-cohort pairs were tested, formed from the 217 exemplars tested against 32 cancer cohorts. For each exemplar-cohort pair, we grouped the patients in the cohort as either scoring high or low using the CIBERSORT estimates of the exemplar's deconvolution proportion for each patient's bulk tumor sample. We determined if the patient scores reflected a natural bimodal distribution (see Fig. 5B for an example with signature X164 in PRAD; see Methods). 2801 exemplar-cohort pairs passed the bimodality test (n=2801). In each of these cases, the two modes were detected and a cutoff was determined that was equidistant between the modes, dividing the samples into high- and low-scoring groups. 4143 exemplar-cohort pairs failed the bimodality test. For these cases, the patient samples were split into two groups using the median of the score distribution as the cutoff (see Fig. 5C for an example with

signature X58 in PRAD).

Once two groups were determined, we asked if the presence versus absence of an exemplar's signature implicated a difference in patient outcomes for a particular type of cancer. To that end, we calculated a signature outcome separation (SOS) measure for an exemplar applied to a TCGA cohort by fitting a Cox proportional hazards (CoPH) model using the covariate of high-/low-scoring patient group (see Fig. 5D-E for Kaplan-Meier plots illustrating SOS for signatures X164 and X58). The significance (-log base 10) of the SOS measure was recorded as the fit of the model. Both univariate CoPH –in which only the signature was used as the predictor of outcome– and multivariate –in which an additional covariate was used that represented the previously published subtype groupings of the samples- tests were calculated. In this latter multivariate case, we refer to the SOS as the subtype-corrected SOS. A significant subtype-corrected separation would indicate an exemplar's deconvolution score separates the patients into groups that are distinct from the established cancer subtypes, or that further separate patients within a subtype, and may be of particular biological and clinical interest.

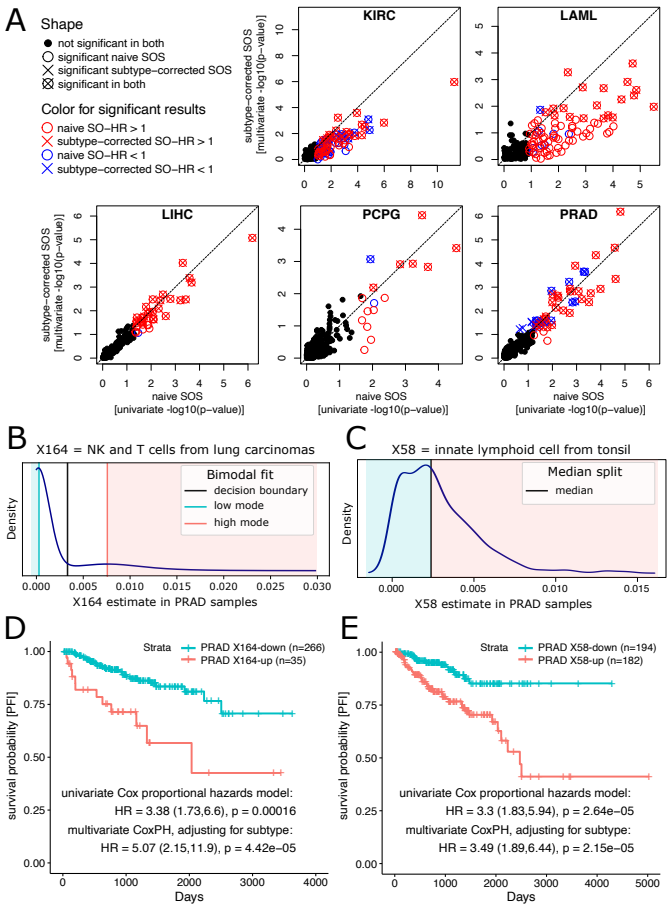

Figure 3.6: **Single-cell exemplar signatures stratify patients into high- and low-risk groups in several types of cancer (A)** CIBERSORT estimates for each of the 217 exemplars (circles, crosses, boxes in plots) were used to stratify patients in each cohort from low-scoring to high-scoring. Five cohorts had at least one signature with a significant separation (FDR¡0.25 on CoxPH). The CoxPH results of the survival separation using only the exemplar signature was plotted ("naive SOS", x-axis, log10 of univariate P-value) or combined with a covariate to account for a tumor type's published subtypes ("subtype corrected SOS", y-axis, log10 of multivariate P-value) to show those that are significant on their one (open circles), with subtype-correction (crossed), or both (crossed boxes) and colored if the presence of the signature indicates a significant separation in patient outcomes (FDR ¡ 0.25) that are either poorer (red, hazard ratio ¿ 1) or better (blue, hazard ratio ¡ 1) (full results in Table S3 and plotted in Supplementary Fig. S18-19). **(B)** For each exemplar signature in each cohort, two groups of patient samples were determined as the high and low category of the score distribution if it matched a bimodal distribution. An example of such a case is shown for exemplar X164 estimated in PRAD samples where a "down group" (blue-shaded area) was distinguished from an "up group" (red-shaded area).

(C) In the case where the bimodal test failed, samples were grouped into the top and bottom half using the median of an exemplar's score. An example is shown for exemplar X58 in PRAD samples with samples below the median score defined as the "down group" (blue-shaded area) and those above as the "up group" (red-shaded area). **(D)** The significance by which each cell type exemplar in each cohort separated the outcomes of the patients was measured using a Cox proportional hazards model (CoxPH) that used either the exemplar signature alone (univariate CoxPH) or combined with a covariate to account for published patient subtypes (multivariate CoxPH). The survival separation is illustrated for exemplar X164 in PRAD using a Kaplan-Meier survival plot to show that samples with estimated higher levels of the cell type represented by X164 have associated poorer outcomes. **(E)** Same as part D but for a different cell type exemplar X58 that also shows poorer outcomes when the exemplar signature is present.

We tested all exemplar-cohort pairs to determine if an exemplar's signature separated the patients by their outcomes using a subtype-corrected and FDR-adjusted test (Supplementary Fig. S13). We calculated SOS and subtype-corrected SOS only for pairs that had at least 10 non-zero samples in samples classified into the high-scoring category (5931 out of 6944). Of these, we found 5730 cases that did not separate by outcome, across all 217 exemplars and 32 tumor types. 89 exemplars produced no outcome separation on any of the tumor types; likewise, for 27 tumor types no exemplars were found that could separate the outcomes after accounting for the published subtypes. For example, there were 163 exemplar-tumor type pairs in which the subtype correction in the multivariate model eliminated the outcome separation detected by the univariate model. In these cases, it may be informative to investigate whether unanticipated microenvironment factors correlate with the published subtypes. However, we chose to focus on cases in which an exemplar had a clear implication on patient outcomes and that were independent of the published subtypes that we discuss next.

We found 38 exemplar-cohort pairs that had a significant subtype-corrected SOS for at least one exemplar (Fig. 5A, Table 1) including four exemplars for the kidney carcinoma

(KIRC) cohort, two for leukemia (LAML), four for liver (LIHC), six for the pheo- and paraglioma neuroendocrine tumors (PCPG), and 21 for prostate (PRAD). For example, four exemplars (X88, X197, X30, and X18) were found for LIHC that may reflect differentiation differences between the tumors. All four were associated with high hazard ratios, indicating poorer outcomes when the signature was detected. Moreover, the ratios were relatively unchanged in the multivariate models, indicating the exemplar-induced dichotomies of the patients are independent of the published subtypes (i.e. represent a different way of grouping the patients).

We plotted the Benjamini-Hotchberg-adjusted significance of the uncorrected and subtype-corrected SOS analysis. Most of the signatures discovered across these five tumor types were associated with poorer outcomes (red entries in Fig. 5A) and no exemplars in which the outcome separation was found to be significant only after accounting for published subtypes. We note that there are three borderline significant exemplars in PRAD (blue x's without boxes) that may represent cases the subtype correction does help reveal the survival separation. Other than these three exceptions in PRAD, we found that the outcome separation either remained significant (Fig. 5A, crossed circles) or was no longer significant in the case that an exemplar recapitulated a separation already accounted for by the published subtypes (Fig. 5A, open circles). Several cancer types (e.g. PRAD) had a linear trend near Y=X, indicating the published subtypes had little to no influence on most of the patients groupings based on signature scores. On the other hand, several tumor types (e.g. KIRC, LAML and PCPG), had linear trends off of Y=X revealing that subtype correction lessened the survival separation significance, suggesting many of the signature groupings are similar to the previously determined subtypes.

We found both exemplars that separate survival in a tumor-type specific manner as

tumor-type	signature	subtype-cor. SO-HR	naive SO-HR	subtype OS
KIRC	X125: cortical excitatory neuron	3.64 (s)	3.70 (n)	(G)
	from organoids			(-)
	X184: fetal fibroblast from placenta	2.23 (S)	2.88 (N)	
	X54: B cells from liver	1.84 (s)	1.68 (n)	
	X145: pancreatic stellate cell	0.58 (s)	0.50 (N)	
LAML	X112: stromal cell and metanephric cap	3.65 (s)	2.26 (n)	(G)
	from multiple tissues			(-)
	X92: astrocyte from brain	3.87 (s)	3.60 (N)	
LIHC	X88: oligodendrocyte precursor cell	5.32 (s)	5.40 (N)	(-)
	X197: iPSC normal culture	3.18 (s)	3.12 (n)	
	to maintain pluripotency			(-)
	X30: Spermatid and germ cells from testis	3.11 (s)	3.00 (n)	
	X18: Plasma cells from bone marrow	2.91 (s)	2.50 (n)	
PCPG	X68: Mammary epithelial cells from primary	5.93 (s)	4.59 (n)	
	breast cancer cells and lymph node			(-)
	X205: Muller cell and retinal rod cell	4.55 (s)	5.20 (n)	
	from retinal neural layer			(-)
	X132: Endothelial cells from embryonic heart	4.43 (s)	4.51 (n)	
	X167: Epithelial and basal cells	4.40 (s)	3.97 (-)	
	from lung carcinomas			(-)
	X54: B cells from liver	3.98 (s)	4.62 (n)	
	X39: type I pneumocyte	0.06 (s)	0.11 (-)	
PRAD	X38: lung ciliated cell	6.41 (s)	4.69 (n)	
	X150: acinar cell from pancreas	5.22 (S)	4.16 (n)	
	X2: Epithelial cells from lung	5.13 (s)	5.36 (n)	
	bronchioalveolar carcinoma			(-)
	X164: immune from lung carcinomas	5.07 (s)	3.38 (n)	
	X211: fetal hepatocytes	4.72 (s)	3.27 (-)	
	X134: Neurons from heart	4.02 (s)	3.52 (n)	
	X122: erythroid lineage cell	3.90 (s)	4.48 (n)	
	from multiple tissues			(-)
	X58: innate lymphoid cell from tonsil	3.49 (s)	3.30 (n)	

tumor-type	signature	subtype-cor. SO-HR	naive SO-HR	subtype OS
PRAD	X44: mast cell from lung	3.47 (s)	2.90 (-)	
	X145: pancreatic stellate cell	3.01 (s)	2.49 (-)	
	X205: Muller cell and retinal rod cell	2.93 (s)	3.03 (n)	
	from retinal neural layer			(-)
	X166: Epithelial cells from lung carcinomas	2.60 (s)	2.33 (n)	
	X196: induced Neural Plate Border	2.55 (s)	2.82 (n)	
	Stem Cells from fibroblast			(-)
	X80: acinar cell	2.54 (s)	2.93 (n)	
	X132: Endothelial cells from embryonic heart	2.42 (s)	2.70 (n)	
	X25: EC-Blood from testis	2.36 (s)	2.12 (-)	
	X151: Alpha cells from pancreas	0.36 (s)	0.39 (n)	
	X103: embryonic stem celll from H9 cell line	0.31 (s)	0.29 (n)	
	X39: type I pneumocyte	0.29 (s)	0.28 (n)	
	X32: Peritubular myoid cells from testis	0.28 (s)	0.40 (-)	
	X87: Macrophages from brain	0.19 (s)	0.22 (n)	
	X9: luminal epithelial cell of mammary gland	0.16 (s)	0.22 (n)	

Table 3.1: **High-confidence signature outcome separation (SOS) results.** All results with a subtype-corrected signature outcome separation (multivariate CoxPH model) fdr-adjusted p-value $p \leq 0.05$. SO-HR = Signature Outcome Hazard Ratio, OS = outcome separation. The subtype-corrected SO-HR is marked with "s" if the outcome separation has an False Discovery Rate adjusted p-value $p \leq 0.05$, and with "S" for $p < 0.001$. Similarly, the naive SO-HR is marked "n" for $p < 0.05$ and "N" for $p \leq 0.001$, and a tumor type for which the subtype groups show significant ($p \leq 0.001$) outcome separation is marked with "G".

well as those that separate patients by outcomes in two or more tumor types. For example, signature X132 shows a SOS in four tumor types, PRAD, KIRC, PCPG and LGG, whereas in all those four tumor types the detection of the signature X132 correlates with worse outcome (high SHS). The cells in signature X132 created from four centroids from the same human dataset, of which a majority of the cells (4568 cells out of 5782 total, 79%) are annotated as "endothelial cells from embryonic heart." Gene set enrichment analysis of X132 identifies "GO_MUSCLE_ORGAN_MORPHOGENESIS" as the most enriched pathway from Gene Ontology. Studies have shown endothelial cells play a role in tumor microenvironment in regulating tumor initiation, progression, and metastasis24. For example, endothelial cells promote prostate cancer metastasis25. Endothelial cell proliferation is known to be associated with tumor angiogenesis in gliomas, which contributes to malignant gliomas26. An emerging theme in metastasis is the involvement of endothelial dedifferentiation as a mechanism tumors use to transform and gain an immune privilege shared by developmental cell lineages (see Huijbers et al. 2022 for a review27). The SCEA contained 7 different prenatal and pediatric datasets (E-GEOD-114530, E-GEOD-124472, E-HCAD-10, E-HCAD-13, E-HCAD-7, E-MTAB-7381, E-MTAB-7407) from which 32 exemplar signature were derived by the scBEACON pipeline that include cell types originating from the liver, heart, kidney, umbilical cord blood, bone marrow, and tonsils. These signatures may implicate additional developmental associations in tumor subsets and are tabulated in the supplemental material (Supplemental Table S7). Signature X112 was derived from stromal cells and metanephric cap cells of the kidney. Studies have shown that Bone marrow stromal cells (BMSCs) promote chemoresistance in acute myeloid leukemia (AML) cells28 and potentially negatively influence patient survival rates. Thus, even

though exemplar X112 was derived from kidney, its stromal signature was robust enough to detect stromal presence in another tissue.

We further investigated specific exemplar-cohort pairs to illustrate microenvironment components relevant to patient outcomes. The associations for all exemplars and tumor types are provided in the supplemental material (see Supplementary Table S3) documenting numerous possible correlations worthy of exploration. For reasons that are not clear to us, many more signatures (n=22) were found to separate the patient samples of the PRAD cohort compared to other cohorts. Among these for example is exemplar X164 derived from lung carcinomas (dataset E-MTAB-6653), which was found to have a bimodal distribution for the PRAD samples (Fig. 5A). The presence of the X164 signature is associated with poorer outcomes for PRAD patients both with and without subtype correction (Fig. 5C). Our annotation pipeline associates the signature with NK cells and T-cells of the immune system (based on PanglaoDB). Because signature X164 is derived from another cancer cohort (lung carcinoma), it is possible this immune-related signature represents a cancer-permissive state (e.g. exhausted or inhibited T-cell populations). Consistent with this finding, some types of T cells, such as TH17 and/or Treg CD4+ T cells, have been shown to be involved in the development or progression of prostate cancer29.

As another example, exemplar X58 scores did not exhibit a bimodal distribution on PRAD samples but splitting the samples by the median signature score (Fig. 5B) produced a grouping of the patients into different outcome classes (Fig. 5D). X58 was derived from an "innate lymphoid cell" scRNA-seq dataset (E-GEOD-70580). Studies have shown that type 2 innate lymphoid cell are enriched in prostate cancer30, which produce interleukin (IL)-4 and -

13, which is known to regulate tumor microenvironment and promote cancer proliferation31,32.

3.3.4 New pan-cancer clustering is revealed on a Tumor Cell-Type (TCT) map using all cell-type exemplar signatures

In order to take a more comprehensive look at the cell type signature estimates in TCGA tumors, we projected the TCGA samples onto a two-dimensional landscape, using the estimates of all 217 cell types as input to Tumor Map20. The interactive Tumor Cell-Type (TCT) map is available online at bit.ly/TCTmap_217exemplars. We clustered the samples using hdbscan (see method), a spatial hierarchical clustering method33 to identify 50 TCT clusters. If a tumor type had at least five samples in multiple clusters, an outcome analysis described in section Methods was performed between the main cluster of that tumor type and the smaller minor cluster(s). Out of the 50 clusters, 35 were "pan-cancer", consisting of at least two or more tumor types. All the clustering and survival analysis results based on TCT map can be found in the supplement (Supplementary Table S4).

Most samples cluster by their tumor type (Fig. 6A). This is expected because the cell-of-origin signal in cancer molecular data is strong and the deconvolution estimates are based on mRNA-Seq data21. Even so, some exceptions were observed in which TCT clusters revealed unanticipated divisions with respect to previous publications of these tumors. Interestingly, the TCT clusters related to PanCan groupings to a higher degree compared to published subtypes in many cases. We compared the TCT clusters with previous subtypes quantitatively for each tumor type (Fig. 6B). For example, STAD had low similarity to both PanCan clustering solution and published subtypes. We investigated the TCT implications for STAD as it provided an

alternate perspective on how the tumors related to each other, The STAD samples were oriented into three main TCT clusters –, c36, c39 and c40 (Fig. 6D). Each of the three clusters contained a mixture of the published STAD subtypes. Thus, STAD as well as 6 other tumor types (i.e. CESC, LUSC, PRAD, LUAD, UCEC, UCS) may be interesting cases for further follow-up analyses to appreciate the TCT factors underlying why these tumors represent exceptions to the general rule in which tumors cluster with others primarily of their tissue-of-origin.

In several cases, the TCT map provided a new grouping for samples that had previously been grouped based on bulk molecular profiles. We found cases in which the TCT map split a published subtype into multiple new clusters and cases where the map merged samples of previously separated published subtypes into a single new cluster. A splitting pattern was found for the STAD cohort. A single copy-number instable STAD cluster, STAD-GI.CIN, was split into two different TCT map clusters (c39 and c40, Fig. 6D). C39 has higher signal from Exemplar X151 (alpha cells from the pancreas) and c40 has a higher signal from exemplar X101 (neural progenitor cells), compared to c39. An alpha cell is one type of endocrine cell that is responsible for secreting the peptide hormone glucagon. Endocrine cells are found throughout the GI tract 34, and enteroendocrine cells are dispersed in gut and stomach epithelium, comprising the endocrine elements of the GI tract 35. We suspect that the alpha cell exemplar reflects the enteroendocrine signal in STAD samples due to the current lack of a stomach enteroendocrine exemplar in the scBeacon collection. The association of X101 with c40 suggests the enteric nervous system (ENS) marks distinct tumor microenvironments, which is supported by work showing the ENS plays an essential role in regulating both the stem cell niche and the tumor microenvironment in many organs 36. Of potential interest to further characterize STAD tumor

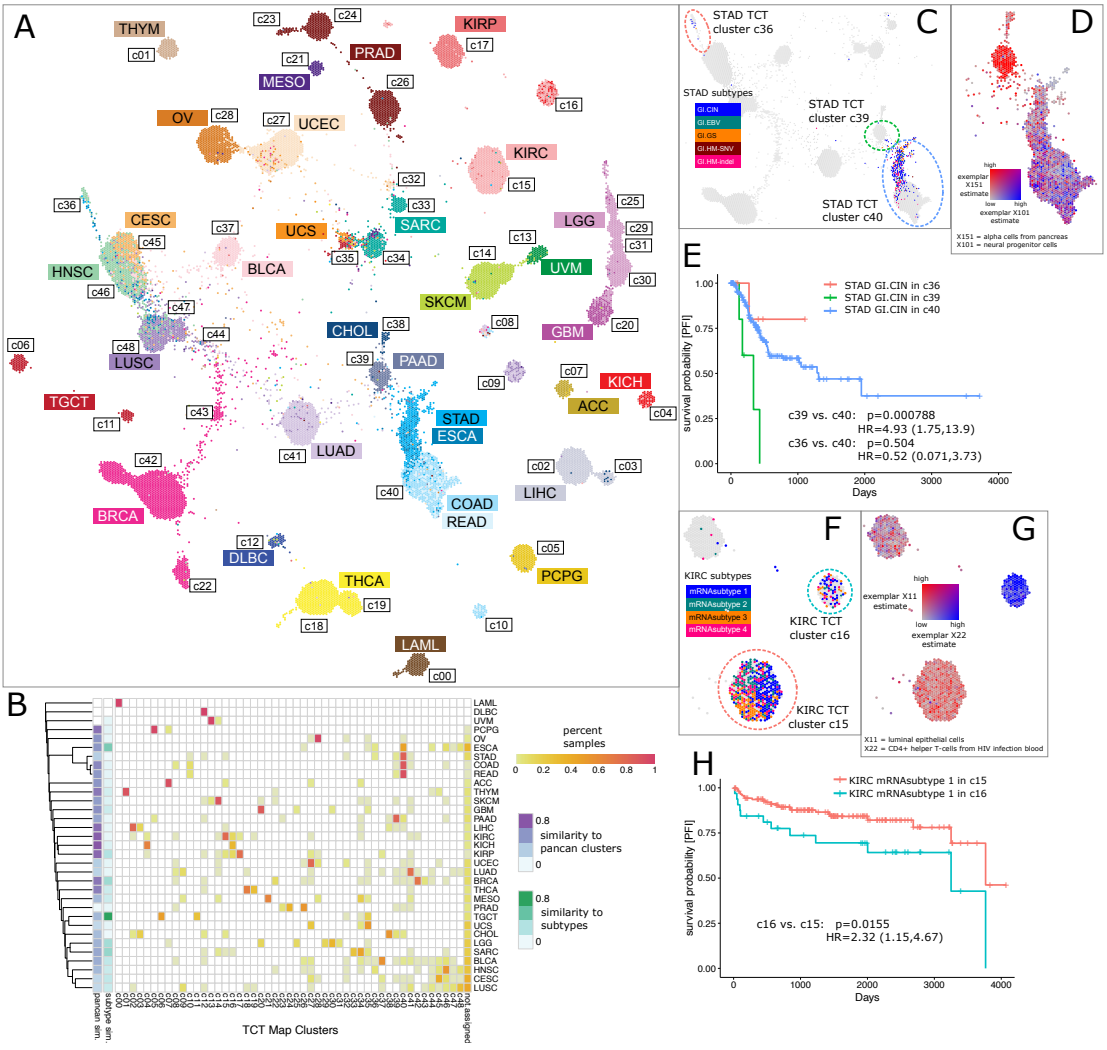

Figure 3.7: **Tumor Cell-Type (TCT) Map (A)** TCT map colored by TCGA tumor type. **(B)** Clustering of tumor types into clusters (clustering solution provided in Supplementary Fig. 20) and the similarity (adjusted rand index) of the clusters to the clustering solution derived in TCGA PancanAtlas21 as well as the grouping of samples into tumor subtypes. **(C)** STAD samples, colored by STAD subtypes. **(D)** Colors show the most differential signatures in STAD GI.CIN samples between cluster c39 and c40: Exemplar X151 (alpha cells from pancreas) and exemplar X101 (neural progenitor cells). **(E)** Survival of STAD subtype GI.CIN samples based on their clusters. **F** KIRC samples, colored by KIRC subtypes. **(G)** Colors show the most differential signatures in KIRC mRNA subtype 1 samples between cluster c15 and c16: Exemplar X11 (luminal epithelial cells of mammary gland) and exemplar X22 (CD4-positive helper T cells from HIV infection blood). **(H)** Survival of KIRC mRNA subtype 1 samples in clusters.

samples, we observed that the samples annotated originally by TCGA as copy-number instable (i.e. the STAD-GI.CIN subtype) were clustered into three TCT clusters that reveal a survival difference among the patients. E.g. cluster c39 patients have significantly lower PFI survival rates compared to c40 patients, and both c40 and c39 patients have lower PFI survival probability compared to c36 patients. Thus, the TCT map was found to reveal differences among tumors previously categorized as the same subtype of disease.

We found examples in which the TCT map clustered together samples belonging to different previously published subtypes. For example, TCT clusters c15 and c16 contain a mix of published KIRC subtypes (Fig. 6F-H). TCT cluster c15 shows higher signal from exemplar X11(luminal epithelial cells) while cluster c16 has higher X22 signal (CD4+ helper T-cells from HIV infection blood). For KIRC mRNA subtype 1 samples between cluster c15 and c16, samples in c15 have better prognosis survival rate compared to c16. This trend could be explained by the role that tumor epithelia plays in regulating immunotherapy outcomes and molecular components in tumor microenvironment37. In Zhang et al, the authors found KIRC-TCGA samples with high estimated fraction of CD8+ T cells have lower survival probability than samples with high estimated endothelial cells. Even though endothelial and epithelial cells have different cellular functions and structure, they are both derived from epithelium and have similar molecular characteristics. This could explain the reason that we observed dissimilar survival trend in TCT cluster c15 vs. c16.

Visual inspection of the TCT map revealed additional examples of groupings that go against the expected trends (cancer types or their subtypes clustering) that may suggest microenvironment factors associated with tumor state. For example, the TCT divides some of

the lung cancers into two distinct clusters (c09 versus c41) with both clusters having equal representation from the major subtypes (LUSC and LUAD). The division appears to separate potentially different lineages with c09 showing higher levels of X38 (lung ciliated cell) and X41 (transformed epithelial cell from lung) compared to those with higher levels in c41 such as X167 (Epithelial and basal cells from lung carcinomas) and X46 (type II pneumocytes). As another example, uterine carcinomas (UCEC) show an interesting pattern in the TCT. Among the copy number high UCEC subtype samples, several cluster with the serous ovarian tumors in c28 (n=30) while others cluster into the c27 group (n=114). The microenvironment factors that underlie the UCEC copy number high distinctions are complicated to interpret as both the high signatures in c28, such as X59 (neurons in the neocortex), and the high signatures in c27 such as X43 (B-cell from lung), are annotated with lower confidence. Other UCEC samples cluster with sarcomas into TCT cluster c32, distinguished by high levels of signature X108 (*Keratinocytes/Suprabasal cells of esophagus and low levels of the well-annotated signature X168 (Basal cells from lung carcinomas). More precise cell type signatures may be needed to understand the major determinants of the UCEC divisions by the TCT. On the other hand, some of the uterine sarcomas (UCS) cluster with the UCEC samples into c27 (n=13) instead of the main cluster (n=14) with immune-related signatures correlated with the division; e.g. with X22 (CD4-positive helper T cell from HIV infection blood) higher in the UCEC cluster compared to epithelial cells (X132 and X117) higher in the other cluster. Finally, the TCT map divided some of the prostate (PRAD) samples into two clusters that were not subtype related with some PRAD samples clustering into c24 (n=199) and others clustering with samples in c26 (n=296) with higher levels in c24 associated with exemplars X152 (acinar cell from pan-

creas) and X156 (CD8-Positive T-Lymphocytes from influenza patients), reflecting a lineage difference (e.g. involving the secretory glands) and/or a variation in the immune components underlying the disease. Thus, the TCT map reveals commonalities among tumors previously considered to have distinct molecular profiles.

3.4 Discussion

There is ever-growing evidence that the cell types present in a tumor's microenvironment influences the outcome of a cancer patient1. In recent years since single-cell sequencing became available, the characterization of various cell types in the human body has improved immensely38–40. A growing number of public single-cell sequencing datasets provides a more accurate and comprehensive definition of the human cell type repertoire. However, there are still challenges to efficiently integrate and analyze those datasets together. First, due to the high level of technical noise and systematic differences between sequencing platforms, simple concatenation could result in batch effects that become the dominant variance rather than biology. Batch effects have been shown to cause an increased number of false positives in downstream analyses41. To reduce the chance of false discoveries, integration of multiple datasets must eliminate batch effects42. Whole reference atlas initiatives such as the Human Cell Atlas (HCA) started collaborative projects to integrate as many datasets as possible to create a whole human cell type map, the data integration process for this task should not only be able to handle batch effects well, but also be computationally efficient and fast while ingesting and integrating datasets.

We used the 217 signatures from Chapter 2 for the deconvolution of 33 different

cancer types from TCGA. Many of the cell-type signatures are found to be correlated to patient outcomes in single tumor types, some also over multiple tumor types.

The interpretation of the deconvolution results has challenges. When a cell type is detected in a cancer sample, it may be due to the cell type being present in the tumor microenvironment. However, another possibility is that the tumor cells themselves have acquired certain characteristics of other cell types, which is ascribed to a particular cell type by the deconvolution method. Yet another possibility is that the usage of an incomplete reference might influence the deconvolution estimate to detect the most closely related cell type when the actual cell type is not included in the signature matrix. In addition, the annotation of the established collection of cell-type signatures is challenging since only a subset of the clusters of a dataset may have reliable annotations either assigned by the authors or inferred by computational methods like those presented in this study. Finally, the granularity of our cell type signatures may have an effect on the downstream analysis. Some datasets in our database are represented completely by just one cell type signature. This happens because all cells in the dataset are from a specific cell type and are very similar to each other compared to other datasets. Nevertheless, a more fine-grained cell type definition might be desirable in some cases, and a hierarchical definition of cell types and cell-type signatures might be a solution to this issue.

TCGA does not contain an exhaustive representation of all tissues and cell types in the body. Indeed, it has a limited set of cancer types. Thus, we expect many cell types to be absent from the TCGA collection. The fact that some signatures are not found when deconvolving may either the exclusion of certain types of cells in cancer tissues in the biased TCGA set or "odd" cell types found in scRNAseq data that are not present in bulks samples (although the latter is

hard to rule out as we did not analyze a comprehensive set of bulk tissue data). As the collection of signatures grows, there will concomitantly be increases in the number of signatures that fail to be detected in any analyzed set of tissues. However, at this stage, such extra cell types have not proven detrimental to the deconvolution or downstream analyses in any tangible way.

Validation of the scBEACON approach using the GTEx consortium data further affirmed its robustness and accuracy in identifying cell type signatures and their application in deconvolving bulk RNA-seq datasets. By leveraging the orthogonal dataset published by the GTEx consortium containing single nucleus RNA sequencing (snRNA-seq) derived signatures and subsequent deconvolution of GTEx bulk RNA-seq samples, we demonstrated that deconvolution with scBEACON-derived signatures for GTEx effectively grouped similar tissues, thereby underscoring its utility across diverse biological datasets. Notably, tissues from related organ systems such as the brain, gastrointestinal tract, and vascular structures exhibited coherent clustering, which is indicative of the tool's precision in capturing organ-specific cellular compositions. Moreover, the comparison of the scBEACON-derived signatures from the Single Cell Expression Atlas (SCEA) with the 35 obtained from GTEx revealed significant overlaps, with more than half of the GTEx signatures showing a high correlation with the SCEA-derived set. The majority of SCEA signatures were unique compared to those in GTEx, indicating a broader scope of cellular diversity captured by the SCEA dataset, reinforcing the capability of scBEACON to provide a detailed and expansive view of cellular landscapes across different conditions and tissues, which is crucial for understanding complex biological systems and their underlying mechanisms in health and disease.

In summary, we provide a comprehensive collection of cell-type signatures based on

the preprocessing of a large amount of scRNAseq data, strategies for identifying and merging signatures across datasets even from different platforms using rank-based centroids, a graph-based meta-clustering approach, and a novel enrichment-based cluster comparison metric. We provide annotations for all of the discovered 217 signatures and document survival associations for 33 exemplar signatures in 5 tumor types. We have made available a new interactive map of all TCGA tumors based on their TCT content. We found evidence for both merging pre-established subtypes into common TCT clusters as well as splitting samples of one subtype into multiple new TCT clusters. We found several examples in which regrouping samples, either using individual signatures on a single tumor cohort or using all signatures in a new pan-cancer TCT clustering, revealed new outcome implications.

Chapter 4

Characterizing cancer subtypes and cell-type components of Testicular Germ Cell Tumors

The project described in this chapter was part of the GDAN testicular germ cell cancer (TGCT) AWG working group at the National Cancer Institute (NCI). My main contributions are characterizing TGCT cancer subtypes in mixture samples, and deconvoluting and analyzing bulk tumor samples using cell-type signatures derived from scRNA-seq, especially cancer-specific tissue infiltrating macrophages. The manuscript is under preparation for submission.

4.1 Introduction

About 95% of all testicular cancers are represented by testicular germ cell tumors (TGCTs). TGCT is the most common solid tumor among males 15–34 years of age, with an estimated 8,850 new cases and 410 deaths during 2017 in the United States[16].

Testicular germ cell tumors (TGCTs) are the most common tumors in men aged between 15 and 44 years[17]. Germ cell tumors account for most of all testicular cancers. TGCT can be divided roughly into 2 groups: seminoma and non-seminoma, whereas Non-Seminoma can be further divided into spermatocytic seminoma, embryonal carcinoma, yolk sac tumor, choriocarcinoma, and teratoma, including many possible combinations of those[29].

The increasing incidence of TGCTs among males provides strong motivation to understand its histology, its genetic basis, and its gene and transcript regulatory properties[5]. TGCT affects several regulatory mechanisms on various molecular levels. Transcription factors and microRNAs can be considered key regulators of the transcriptome that frequently show aberrant activity in cancers. MicroRNAs (miRNAs) are short non-coding RNAs ($\tilde{2}2$ bases)[27, 22, 6] involved in many biological processes and human diseases[25]. miRNAs regulate mRNAs by either degrading them or preventing their translation[50]. The possible interplay between mRNAs, miRNAs, circular RNAs (circRNAs), long non-coding RNAs (lncRNAs), and other types of RNA that have miRNA binding sites gives rise to a large regulatory network between transcripts that carry miRNA binding sites. Network transcripts are referred to as competing endogenous RNAs (ceRNAs), as they compete for binding a limited pool of miRNAs[41]. In particular, circRNAs are potentially very potent ceRNAs as their circular form

is stable and protects these RNAs from degradation while at least some show a higher number of binding sites compared to other RNAs[23]. One hypothesis is that circRNAs act as miRNA buffers[54]. Several examples of interactions between miRNAs and circRNAs are known (e.g., CDR1as/CiRS-7, SRY[20], and circNCX1[26]). Even though individual studies confirmed the existence of ceRNAs in cancer, we lack knowledge of these phenomena in TGCT.

4.2 Subtype deconvolution in TGCT

Understanding subtypes in testicular germ cell cancer is crucial. Identifying and classifying these subtypes enables personalized treatment strategies tailored to individual patients, optimizing therapeutic efficacy while minimizing unnecessary side effects. Additionally, advances in molecular profiling have revealed heterogeneity within and across subtypes, highlighting the complexity of disease progression and the need for targeted interventions. Expert Pathologist Committee (EPC) review reported that many tumor samples contained mixtures of histological subtypes, which makes it harder to assign subtypes for the mixture samples. To extend a simplistic approach that assigned one subtype per sample, we deconvoluted subtypes for each sample, using data from DNA methylation, gene expression, and microRNA expression.

In the sample cohort, we have 4 pure histological types: seminoma(170), embryonal carcinoma(47), yolk sac(18), and mature teratoma(13). And 2 dominant histological types, choriocarcinoma(1) and immature teratoma(3), are the samples that have more than 70% of assigned histology but also consist of other histology types. The numbers in the parathesis are the number of the samples.

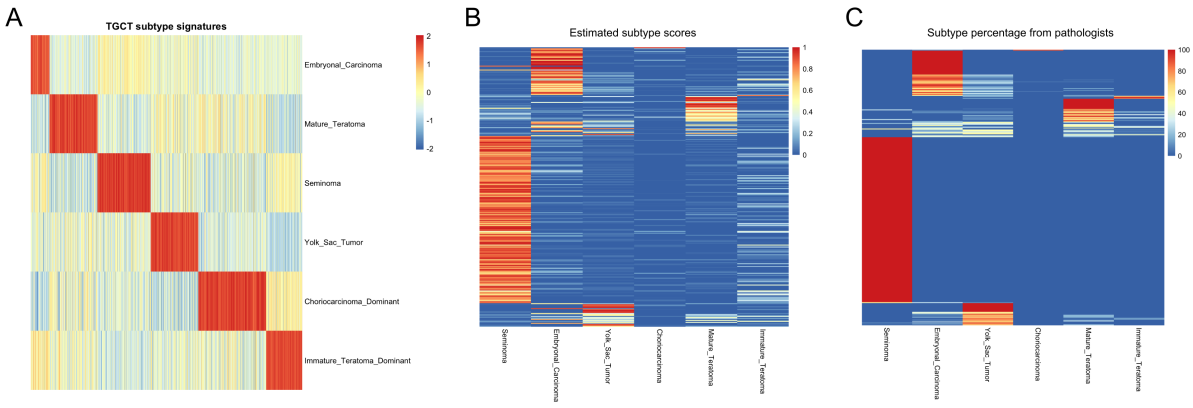

Figure 4.1: **Subtype Deconvolution in TGCT mixture Samples.** **(A)** Signature genes for the 6 TGCT subtypes, columns are the gene expression values for marker genes, rows are the subtypes. **(B)** Deconvolution results for TGCT samples. The results are 0-1 normalized **(C)** Pathologist review of the subtype percentage for TGCT samples.

First of all, I ranked normalized (ranking/number of genes) gene expression values in each sample and computed the average rank-normalized gene expression for each histological type to generate ranked centroids for each subtype. For choriocarcinoma and immature teratoma, I took the "purest" sample for each subtype and computed the ranked sample to represent ranked centroids for the two subtypes. Next, for the 6 ranked centroids, I subtracted the highest and second-highest expressed histological types for each gene and got a vector of differential expression values, then I ranked the differential expression vector, from highest to lowest. Next, I took the top 20% of genes as the differential expressed genes. Then I computed the signature matrix for CIBERSORT by subsetting differential expressed genes from the average rank normalized matrix generated in the previous step.

Then I calculated the accuracy of our deconvolution-based subtype estimation. We define the Estimate accuracy = # samples match with pathologists' call \ total # of pure or

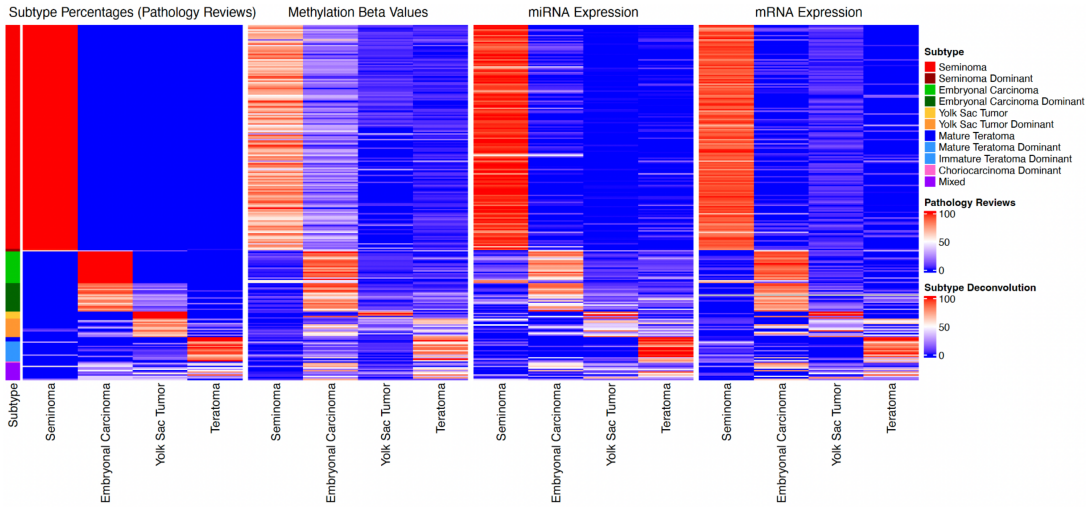

Figure 4.2: **Subtype Deconvolution in TGCT mixture Samples** Left to right: Subtype percentages from pathology (EPC) review and from subtype deconvolution using DNA methylation data, miRNA expression data, and mRNA expression data. Immature teratoma and mature teratoma were merged into a 'teratoma' subtype; the one choriocarcinoma was excluded from the analysis.

dominant samples. For pure samples, the accuracy goes up to 97.7%, for dominant and mixture samples, the accuracy is 78.6%.

Next, we applied this approach for DNA methylation data and microRNA data, to compare the results using these three omics. For this analysis, we merged the mature teratoma and immature teratoma into a single subtype type, and we removed the choriocarcinoma subtype since there were no 100% immature teratoma or choriocarcinoma samples. The predicted, deconvolved subtype proportions were concordant with subtype proportions from the pathology review. DNA methylation data estimated larger fractions of embryonal carcinoma components in seminoma samples, while results were similar to gene expression and miR expression data.

4.3 Cell type deconvolution in TGCT

The tumor microenvironment (TME) consists of various non-cancerous cell types, including immune cells, fibroblasts, endothelial cells, and extracellular matrix components, which interact with tumor cells and influence tumor behavior. Tumor purity refers to the proportion of tumor cells in a tumor sample relative to non-tumor cells. The components of the TME can significantly impact tumor purity through multiple mechanisms. For instance, infiltrating immune cells such as tumor-infiltrating lymphocytes (TILs) can target and eliminate tumor cells, leading to decreased tumor purity. In TGCT samples, seminoma and non-seminoma samples showed distinct differences in their gene expression profile(see Figure 4.3(A)) as well as tumor purity level, as shown in Figure 4.3 (B). Non-seminoma samples have higher purity levels compared to seminoma samples. In Figure 4.3(C), LM22 is the default immune signature consisting of 22 major immune cell populations, the high negative correlation indicates the presence of immune cells in the tumor microenvironment contributed to tumor impurity in TGCT samples. Therefore, here we are using deconvolution to understand and characterize the cell type ratios for tissue components and microenvironment in TGCT samples.

The CIBERSORT analysis was performed using cell-type signatures that were derived from single-cell RNAseq studies. The analysis results in a score that serves as an estimation of cell type abundance within each tumor sample. Note that there is no requirement for a sample's cell-type scores to sum to one. That is, if no cell-type signature matches well with the sample data, the sample does not get forced into the closest cell-type. We begin with a signature set that was derived from a scRNA-seq study of healthy fetal and postnatal human

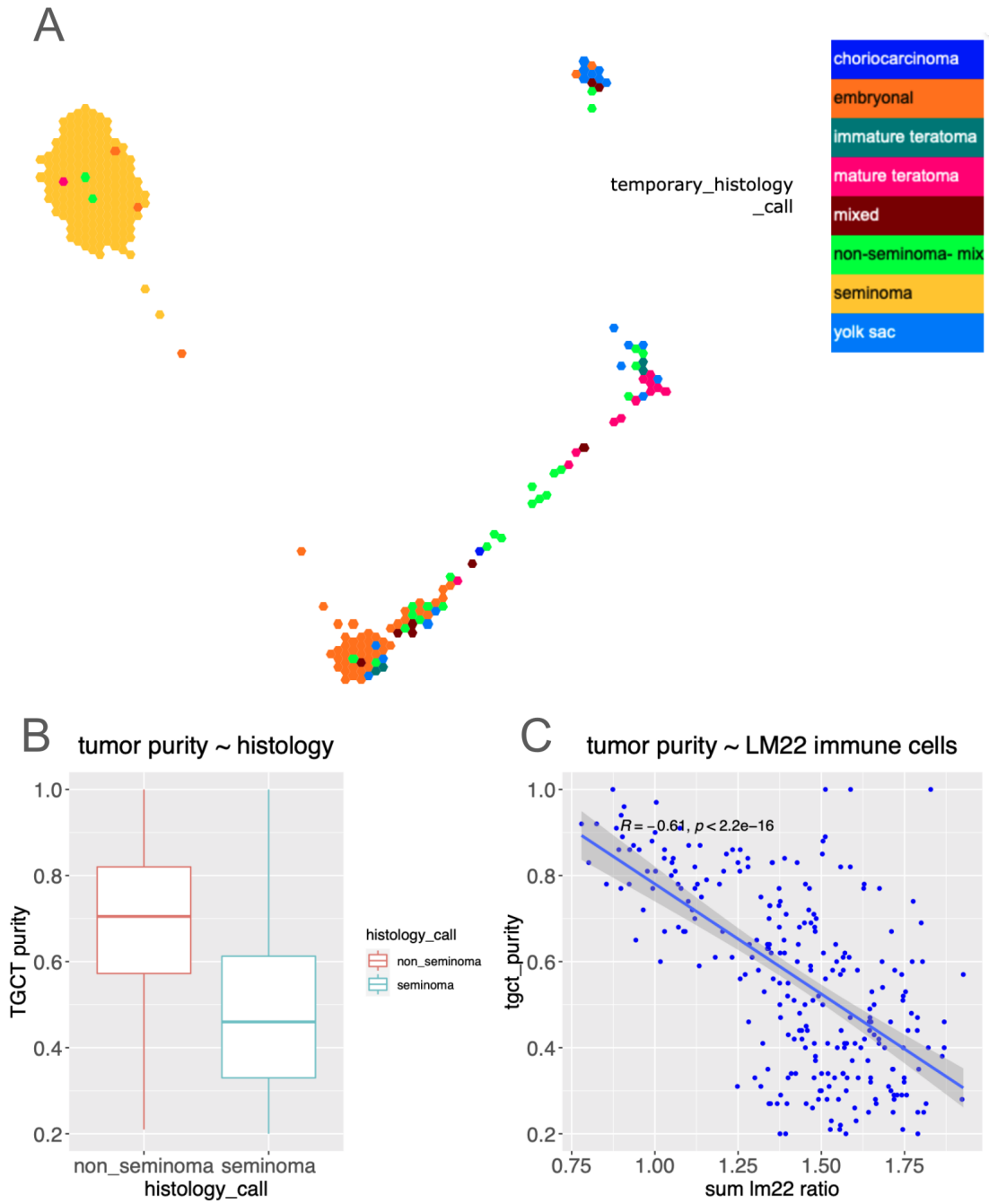

Figure 4.3: **TGCT tumormap and purity.** **(A)** Tumormap using RNA-seq gene expression, with histology calls in colors. **(B)** Tumor purity in seminoma and non-seminoma samples. **(C)** purity score in correlation with the sum of the LM22 score in the scatter plot.

testis[18]. The most striking difference between the histologic subtypes is in the germ cell signature score. Overall, the seminoma samples have the highest germ cell signature scores. Of the non-seminomas, the embryonal (and embryonal dominant) samples had the highest germ cell signature scores. The remaining non-seminomas scored low in the germ cell signature scores. One thing to note is that there seems to be a small subset of KIT-mutated seminomas that have low scores in the germ cell signature compared to the other seminomas. There is a much smaller set of KIT-wt seminomas that has a similar pattern but to a lesser extent. These germ cell-low, KIT-mutated samples had signature scores that were distinct in other CIBERSORT analyses, as well.

Preliminary analysis also reveals some evidence of differences between histology subtypes in macrophage signature scores derived from the study conducted by Guo et al[18]. Seminomas and embryonals score slightly higher in that signature. We performed additional analysis to determine whether the signature is detecting tissue-resident macrophage or infiltrating macrophage. For this, we further investigated cell type using signature sets derived from immune studies of various organ systems[11]. We found that, in all four organ signature sets and the average across all organs, the macrophage scores were the highest compared to other immune cell types, which is in concordance with the previous human testis studies, which conclude testicular macrophages are the largest immune cell population in the human testis.[52, 55]

To better understand the macrophage populations in TGCT samples, we generated three macrophage signatures from the scRNAseq study of human gonadal development[15] for deconvolution: SINGLEC15_ftM fetal testicular macrophages (ftMs), which had an osteoclast-like signature; TREM2_tfM, which had a microglialike signature; and tissue_repair.macrophage.

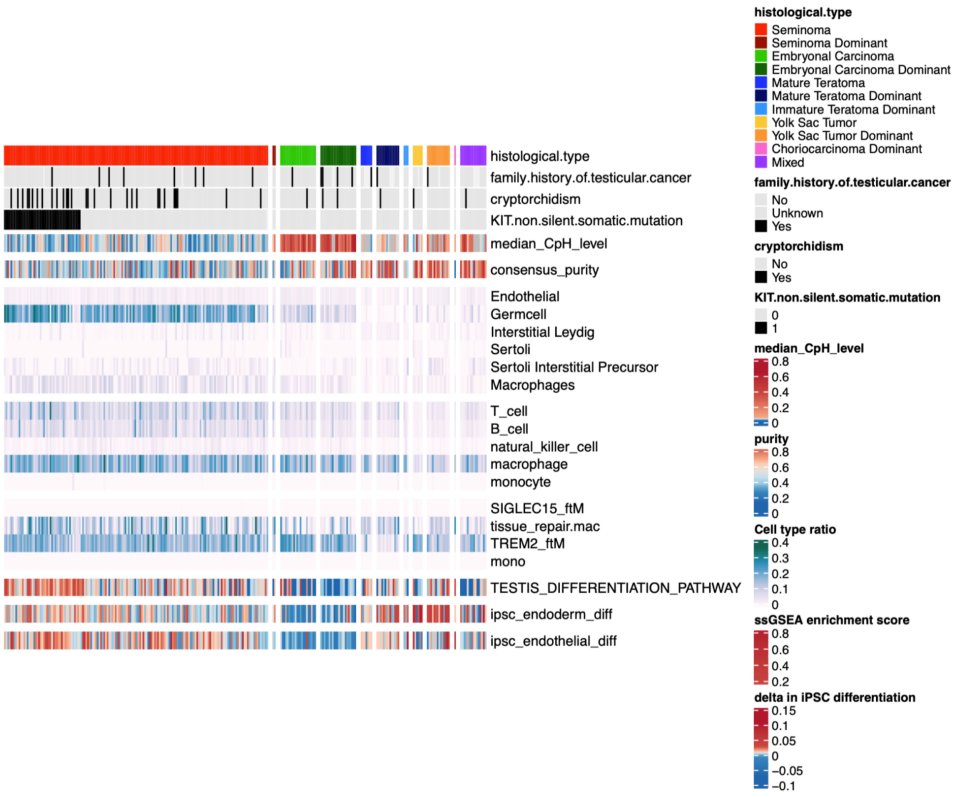

Figure 4.4: **TGCT cell type deconvolution.** Cibersort deconvolution for TGCT samples using cell type signatures derived from multiple single-cell RNA-seq datasets

We also included a monocyte signature in the same study to represent the infiltrating macrophage population. Monocytes originate in the bone marrow and circulate in the blood. SINGLEC15_ftM and TREM2_tfM are two types of testis-specific resident macrophages. tissue_repair. macrophages, which are present in all developing tissues, make up the majority of the macrophage population in the Garcia Alonzo study. In TGCT samples, TREM2_tfM and tissue_repair.macrophage are the two major macrophage populations, with little SINGLEC15_ftM and monocyte signals showing up in CIBERSORT deconvolution results.

In this study, we then did correlation analysis for macrophage scores from embryonic testis study with different populations of macrophage. We computed the correlation of fetal or postnatal macrophage scores[18] with scores from a human gonadal development study[15]. We observed that the correlation was highest with tissue-repair macrophage signature (R= 0.52). The TREM2_tfM signature had a lower correlation (R= 0.18). Again, TGCT samples had very low SINGLEC15_ftM and monocyte signature scores, so no correlation was observed for those cell types. This correlation analysis suggests that the fetal or postnatal macrophage scores that we observe from the Guo study can be attributed mostly to the tissue-repair macrophage cell type.

We found that the correlation of the guo_macrophage signature with either the tissue-repair.mac signature or the TREM2_tfM signature depended upon the subset of TGCT samples that we considered. We separated our TGCT cohort into seminoma and non-seminoma groups in this analysis. For the seminoma group, the guo_macrophage signature was more highly correlated with the tissuerepair.mac signature compared to the non-seminoma group (pearson_r_seminoma= 0.60, pearson_r_non_seminoma= 0.18). We saw a different pattern for

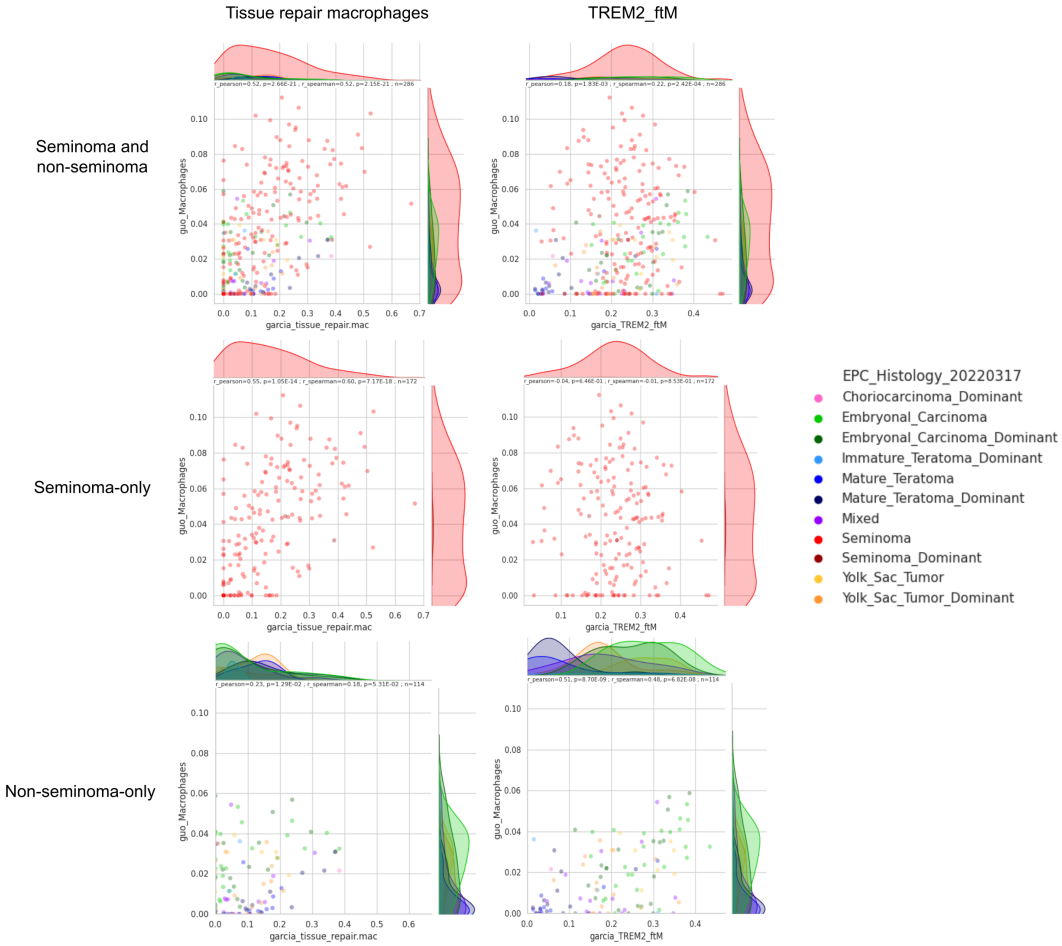

Figure 4.5: **TGCT subtypes consist of different macrophage populations.** Cibersort deconvolution for TGCT samples using cell type signatures derived from multiple single cell RNA-seq datasets

the non-seminomas. The guo_macrophage signature correlation with TREM2_tfM signature was higher for non-seminomas than for seminomas (pearson_r_non_seminoma= 0.48, pearson_r_seminoma= -0.01). We also observed that, of the non-seminoma histology types, the embryonal carcinoma samples tended to have the highest guo_macrophage and TREM2_tfM signature scores. The mature teratoma samples tended to have the lowest guo_macrophage and TREM2_tfM signature scores. [reference scatterplots] [Perhaps seminomas come about early in testis development, before testis immune privilege has been established, allowing access by tissue-repair macrophages. Non-seminomas develop at later stages when the immune privilege is in place, so we see the more dominant signal from tissue-resident TREM2_tfM.]

We used scRNAseq data from studies that investigated the developmental processes involving iPSC endoderm[62] and iPSC endothelium[63]. Cell type signatures at different time points were derived from these datasets, representing timepoint snapshots of tissues as they progress through various developmental time points. As before, we used these developmental cell-type signatures to score the TGCT samples using CIBERSORT. Using the TGCT sample scores from the signatures that represent the earliest and latest developmental time points, we computed differential iPSCendothelial and differential iPSCendodermal scores to get a readout of how the TGCT samples relate to each other along those developmental axes. We found that embryonal tumors have the lowest differential scores. This result suggests that the cells of embryonal tumors are the least differentiated of the histology subtypes in our dataset. The result is corroborated by GSEA scores in a testis differentiation pathway (ref name of the pathway as it appears in MSigDB). The GSEA result also includes the highest differentiation in KITmutated seminoma.

4.4 Discussion

In this study, we characterized the TGCT subtype for the mixture samples that consist of multiple subtypes and compared the results to the histology review. Using the deconvolution pipeline, we are able to achieve high accuracy in assigning subtypes for TGCT samples.

We also showed that different macrophage populations are associated with TGCT histology, which is also associated with mutation status in TGCT samples. We generated developmental scores derived from iPSC stem cells, and the differential score correlated well with the testis differential pathway enrichment score. This approach could potentially be used to characterize cell state transition and developmental stages in tumors.

Chapter 5

Identifying cell types and cell states in glioblastoma

The project described in this chapter is part of the Treehouse Childhood Cancer Initiative collaboration with University of Helsinki and University of Arkansas for Medical Sciences. The work is still subject to active research and will be published under the project lead of Analiz Rodriguez, Vadim Le Joncour and Olena Vaske.

My main contributions are extraction of cell state signatures in previous studies, and deconvolution for gliablastoma bulk RNA-seq samples with predefined marker genes and brain cell developmental trajectories.

5.1 Introduction

Glioblastoma multiforme (GBM) is the most common malignant primary brain tumor among adults and carries a grim prognosis with less than a 5% chance of 5-year survival following standard therapy[38]. However, the standard treatment such as radiation and chemotherapeutic has remained the same for the last decades. This one-size-fits-all treatment approach to heterogeneous tumors has contributed to the poor treatment outcome in patients. There has been tremendous work on identifying and characterizing the molecular subtypes in cancer samples including glioblastoma. Studies of inter-tumor heterogeneity based on bulk gene expression data in TCGA suggest that at least three subtypes of glioblastoma exist, namely proneural (TCGA-PN), classical (TCGA-CL) and mesenchymal (TCGA-MES)[46, 48]. Based on glioblastoma subtypes, researchers identified genetic heterogeneity between tumor samples.

Another layer of heterogeneity is the developmental states of glioblastoma cells in the tumor[30]. Glioblastoma contains subsets of glioblastoma stem cells (GSCs), which interfere with neuron development and contribute to treatment resistance and tumor metastasis[3, 9]. In addition, multiple subtypes or developmental states can co-exist in different regions of the same tumor[32]. It is important to understand the cell types and states in glioblastoma at the single-cell level, including tumor components and microenvironment, either with single-cell sequencing technology or bulk tumor deconvolution using cell type signatures derived from single-cell gene expression profiles.

There are a few studies for defining and characterizing the cell type and states in glioblastoma, in this project, I used two studies: First, 1) Liu, Ilon, et al. "The landscape of tu-

mor cell states and spatial organization in H3-K27M mutant diffuse midline glioma across age and location." Nature Genetics 54.12 (2022): 1881-1894. In this study, the authors dissected H3-K27M mutant diffuse midline glioma using single-cell transcriptomic, epigenomic, and spatial data. They identified 5 tumor metaprograms: "astrocyte-like" (AC-like), "oligodendrocyte-like" (OC-like), "mesenchymal-like" (MES-like), oligodendrocyte precursor cell (OPC-like) and cycling. OPC-like cells were further resolved into three subpopulations (OPC-like-1, OPC-like-2, and OPC-like-3). Oligodendrocyte precursor cells (OPC)-like glioma cells demonstrated a cancer stem cell-like state that is capable of self-renewal and tumor initiation. This indicates that OPC-like cells are at the core of K27M mutation-mediated tumorigenesis.

In 2) Neftel, Cyril, et al. "An integrative model of cellular states, plasticity, and genetics for glioblastoma." Cell 178.4 (2019): 835-849, the authors analyzed scRNA-seq data from glioblastoma patients and concluded that malignant cells in glioblastoma exist in 4 cellular states: (i) neural progenitor-like (NPC-like), (ii) oligodendrocyte-progenitor-like (OPC-like), (iii) astrocyte-like (AC-like) and (iv) mesenchymal-like (MES-like) states. Mesenchymal-like (MES-like) states can be further separated into two meta-modules: hypoxia-independent (MES1) and dependent (MES2) signatures. Neural progenitor-like (NPC-like) is also further subdivided into two subprograms: OPC-related genes in NPC1 vs. neuronal lineage genes in NPC2.

There are overlaps and discrepancies in the definition of cell states or metaprograms in glioblastoma in those two papers, in this project, I analyzed the two sets of cell states separately.

However, brain tissue has been known for its complexity and the amount of different cell types that exist in the brain. Many cell types that are derived from the same progenitor

cells often share similar marker gene profiles, which makes it harder to generate a unique set of marker genes for brain tumor cell types needed for deconvolution. Because of this, many tumor deconvolution studies or methods avoid brain tumor cell type deconvolution. Here I proposed to use a multi-group, hierarchical approach for constructing a signature matrix for glioblastoma. This approach minimized the problem of overlapping marker genes and resulted in more robust deconvolution results.

In this context, the definition of cell states and cell types might overlap to a degree, but there is an important distinction where cell type refers to the specific kind or category of cell based on its structural, functional, and molecular characteristics, and cell states address the current condition or status of a cell, which can be dynamic and can change in response to various internal and external factors. Cell state can include factors such as whether the cell is actively dividing (in the cell cycle), and whether it is undergoing differentiation (changing into a specialized cell type). Here we emphasize the approach to generate a ”cell signature”, regardless of the difference between cell types and cell states.

5.2 Deconvolute glioblastoma based on cell type developmental trajectories

There are two developmental trajectories in glioblastoma samples that we are interested in, see Figure 5.1. We are particularly interested in trajectory_1, because trajectory_1 summarized the glioblastoma-related neuron cell types development, and it contains the cell types we are interested in such as oligodendrocyte-progenitor-like (OPC), astrocyte, and oligo-

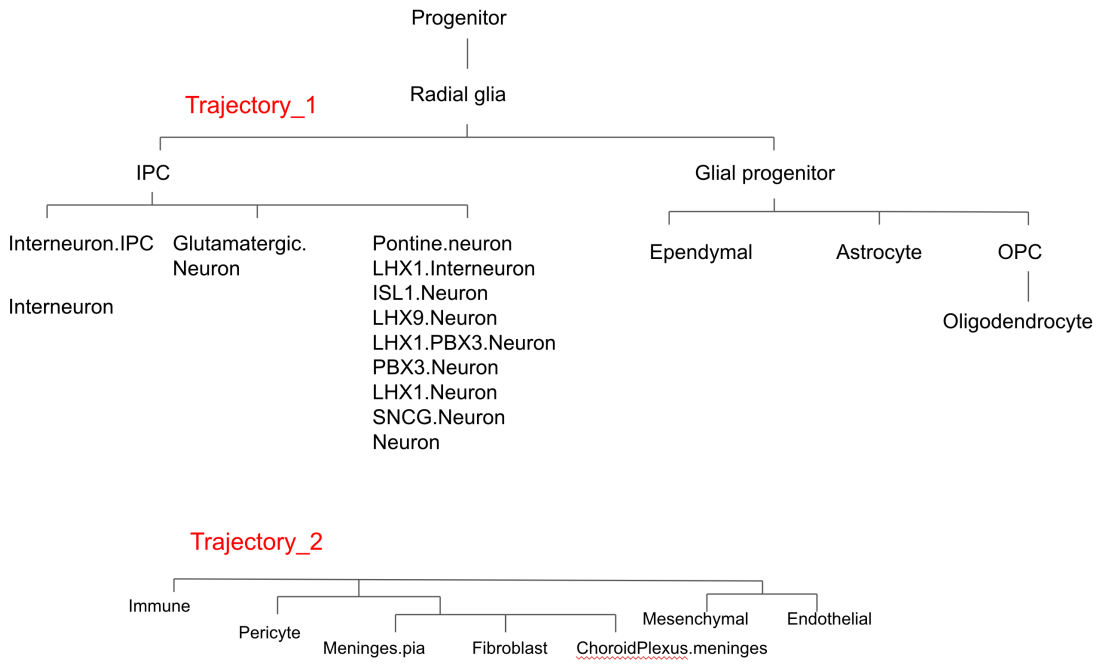

Figure 5.1: **Glioblastoma cell type developmental trajectories**

dendrocyte. In Figure 5.2, I grouped cell types into 8 groups for a lower-resolution first-round deconvolution. In this way, similar cell types, or cell types derived from the same trajectory branches are grouped together and minimize the impact of overlapping marker genes for individual similar cell types.

Figure 5.3 shows the first round deconvolution results for the 8 cell type groups, and figure 5.3(A) shows the signature matrix with a set of differentially expressed genes for each group. Figure 5.3(B) shows deconvolution results using synthetic bulk generated using the glioblastoma single-cell RNA-seq dataset from the Treehouse.

As shown in Figure 5.3(B), the correlation between the ground truth and deconvolu-

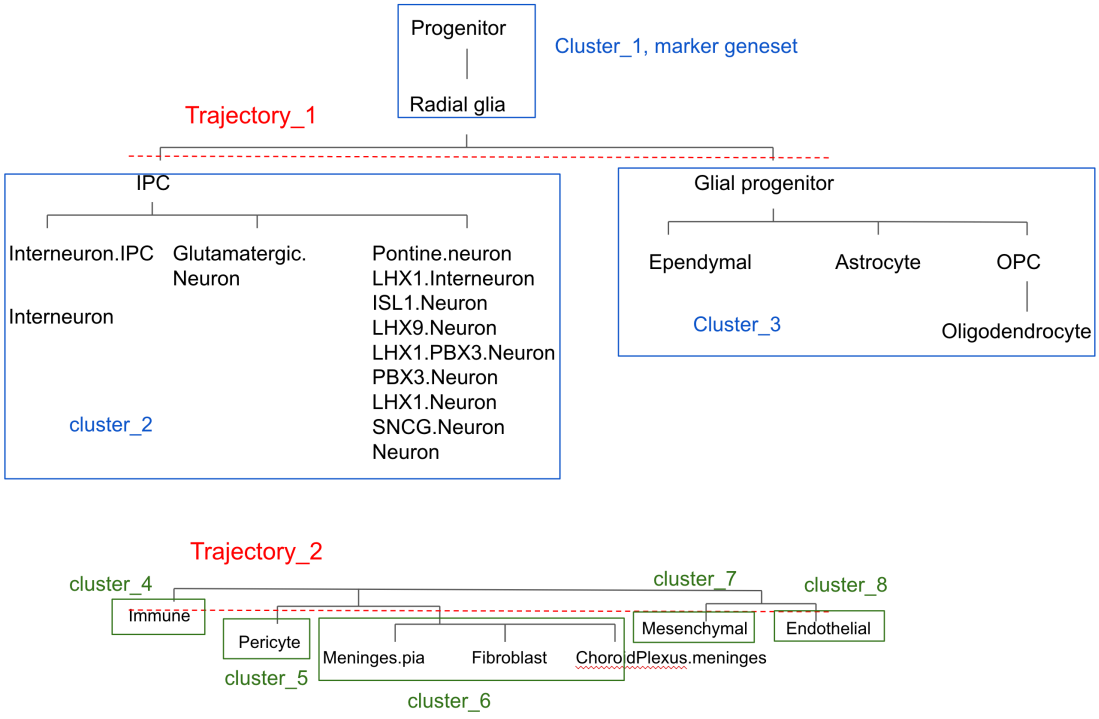

Figure 5.2: **Group glioblastoma cell types based on developmental trajectories**

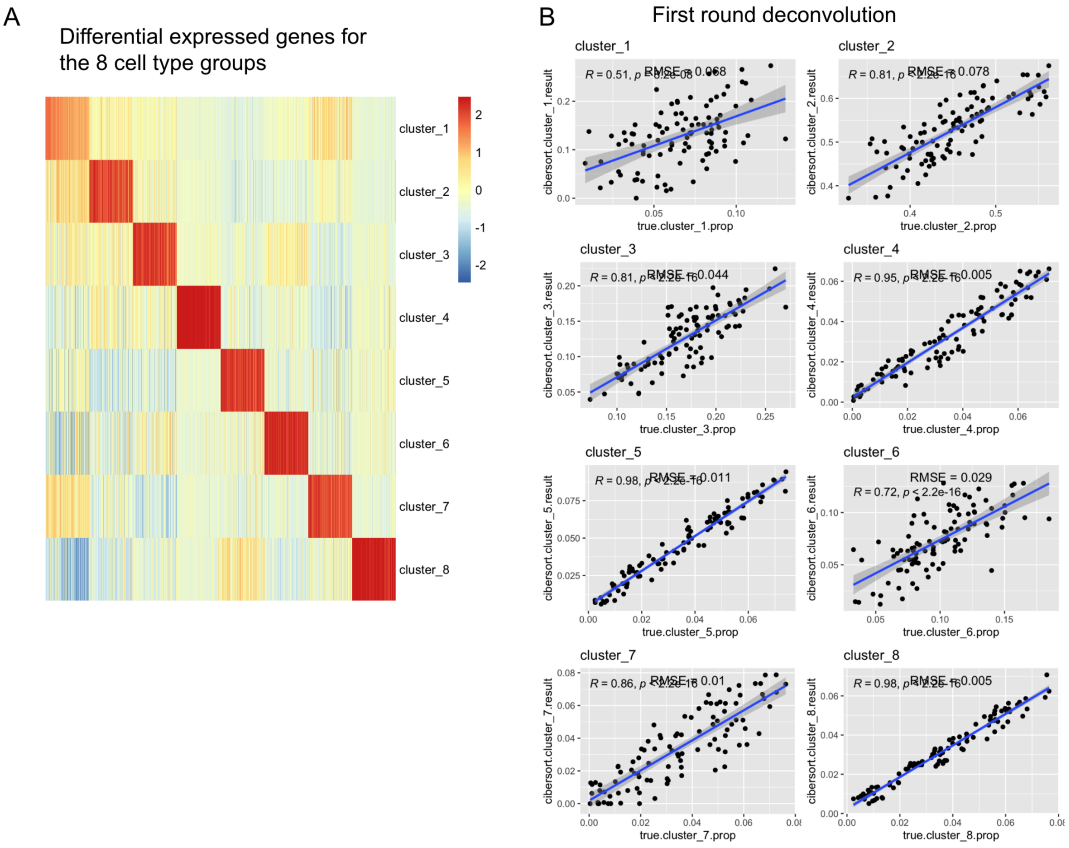

Figure 5.3: **Group glioblastoma cell types deconvolution, first round. (A)** Signature matrix with a set of differential expressed genes for each group. **(B)** First round of deconvolution results using synthetic bulk samples generated from scRNA-seq data

tion results is high, especially for group 3, which consists of ependymal, astrocyte, OPC, and oligodendrocyte, the cell types we are most interested in.

Next, since we are very interested in characterizing astrocytes, OPC, and oligodendrocytes in glioblastoma, I picked group 3 for further analysis. As shown in Figure 5.4, I further separated Group 3 into two clusters(cluster_9 and cluster_10). Differentially expressed genes in the second round of deconvolution are selected from the marker genes of group 3, and the estimated ratios for cluster_9 and cluster_10 are normalized by the estimated ratio of group 3.

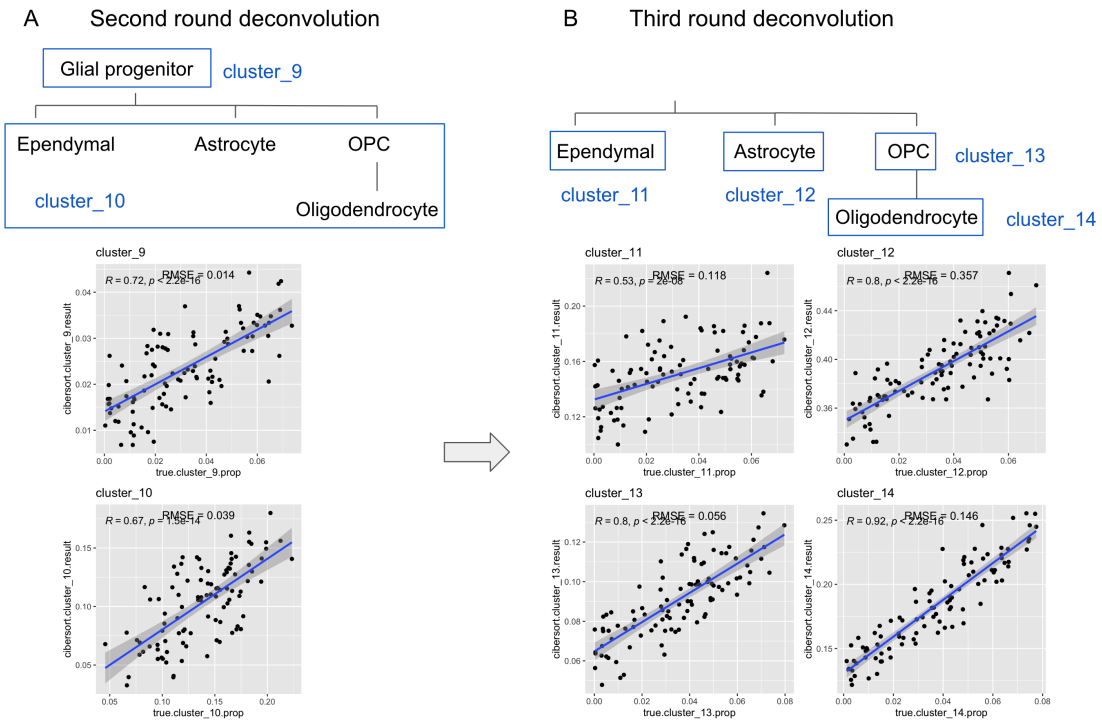

Figure 5.4: **Group glioblastoma cell types deconvolution, first round.** **(A)** Second round deconvolution on group 3. **(B)** Third round of deconvoluton results

Using the same strategy, I performed the third round of deconvolution shown in Figure 5.4(B), in which astrocyte, OPC, and oligodendrocyte ratios are estimated individually with high correlation with the ground truth ratios with low RMSE.

Using a hierarchical strategy for marker gene selection for deconvolution addressed the nature of cell type differentiation. This approach made it possible to have a finer resolution for deconvolution, especially when some rare cell types come from a subpopulation of a very similar cell type. Compared results to the traditional way of analyzing all cell types possible at once in one deconvolution run, separating the deconvolution into multiple steps based on cell type developmental trajectory showed decent results.

5.3 Estimate cell states in glioblastoma tumor with H3-K27M mutation

Histone 3 lysine27-to-methionine (H3-K27M) mutations most frequently occur in diffuse midline gliomas (DMGs) of the childhood pons. It is important to understand the heterogeneity of glioblastoma tumors with H3-K27M mutations. Here we collected two papers that defined cell states in glioblastoma tumors with H3-K27M mutation using bulk RNA-seq data, with marker genes associated with each cell state. For this project, we want to extract the cell states in these two publications and deconvolute bulk tumor samples in Treehouse, the Childhood Cancer Initiative in UC Santa Cruz.

The two papers are listed below: 1) Liu, Ilon, et al. "The landscape of tumor cell states and spatial organization in H3-K27M mutant diffuse midline glioma across age and location." Nature Genetics 54.12 (2022): 1881-1894. 2) Neftel, Cyril, et al. "An integrative model of cellular states, plasticity, and genetics for glioblastoma." Cell 178.4 (2019): 835-849.

In order to extract cell state signatures for deconvolution, I collected single-cell RNA-seq data GSE102130[14] to generate a signature matrix for cell states. I annotated the samples in the scRNA-seq dataset with cell state marker genes and performed deconvolution analysis using CIBEROSRT on Treehouse glioblastoma samples. See results in Figure 5.5 and Figure 5.6.

The two results correlated well in terms of the AC and MES like cell states, which made up for the majority of the cell states composition in glioblastoma samples. OC, OPC, and NPC population varies across samples and the two results, could caused by the different

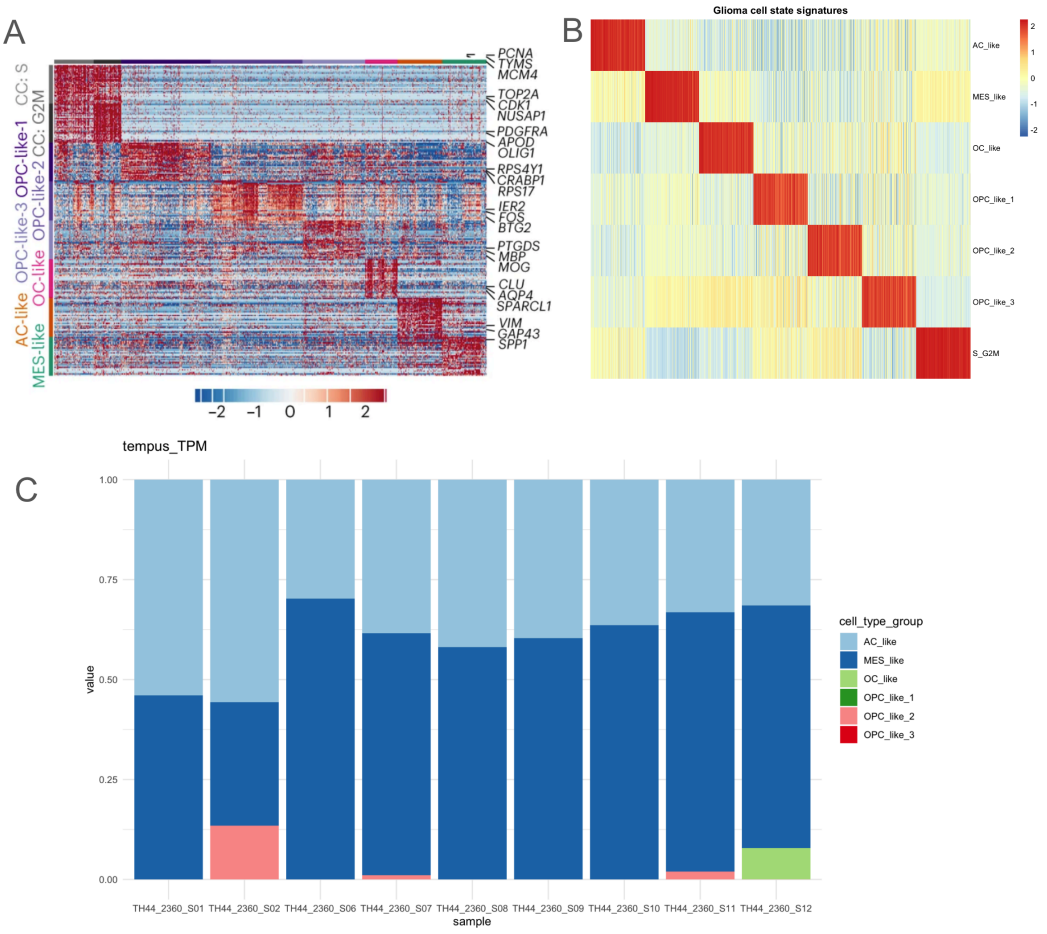

Figure 5.5: **Cell states estimation in patients based on Liu et al.** **(A)** Cell states heatmap with marker genes in Liu et al. **(B)** Signature matrix for cell states based on Liu et al, in GSE102130 glioblastoma scRNA-seq data. **(C)** Cell states deconvolution results in patient tumor samples.

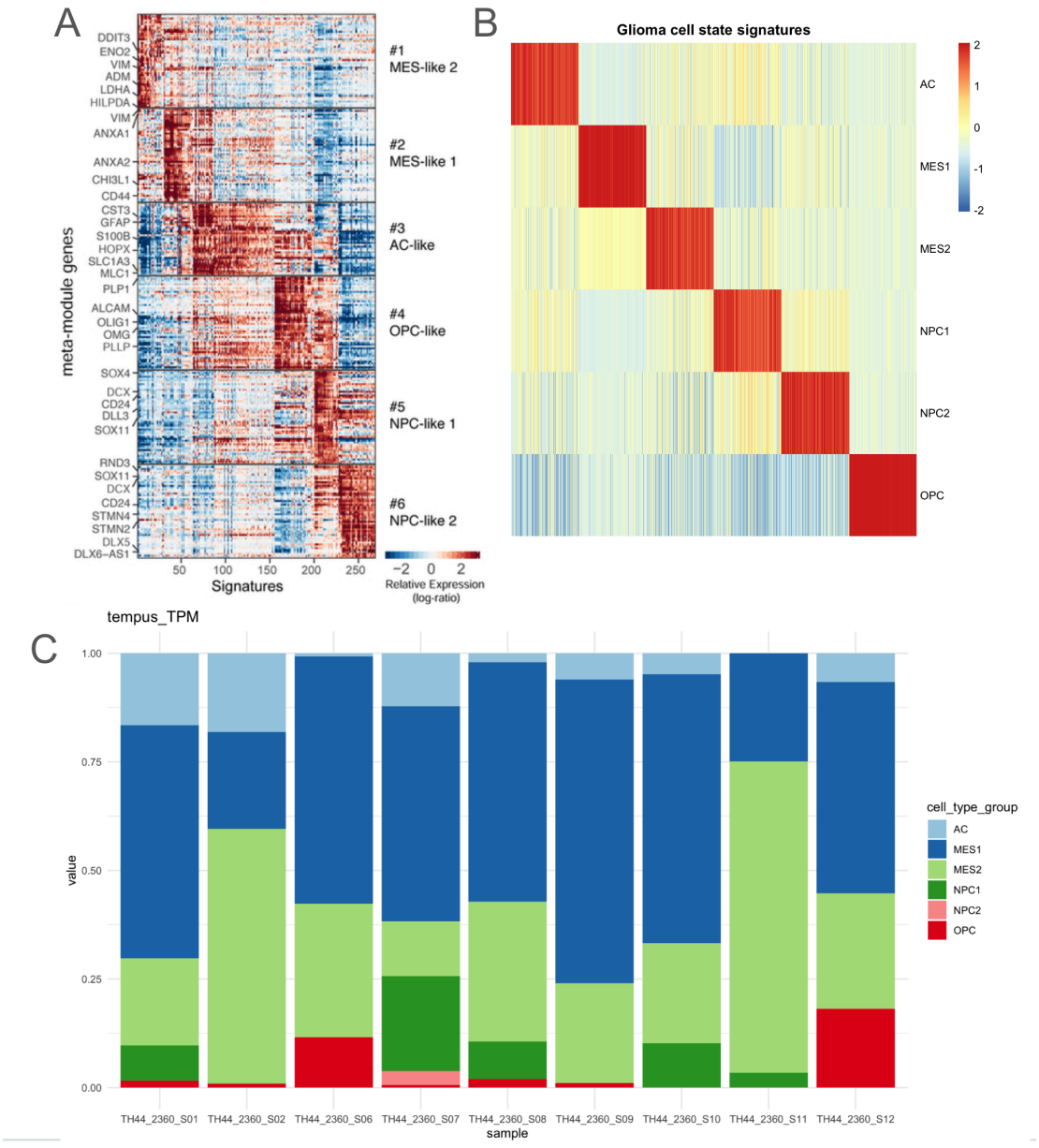

Figure 5.6: **Cell states estimation in patients based on Neftel et al.** **(A)** Cell states heatmap with marker genes in Neftel et al. **(B)** Signature matrix for cell states based on Neftel et al, in GSE102130 glioblastoma scRNA-seq data. **(C)** Cell states deconvolution results in patient tumor samples.

molecular definitions for OC, OPC, and NPC in the two papers.

5.4 Discussion

In this chapter, two independent methods were applied to deconvolute bulk RNA-seq glioblastoma tumor samples: the first using synthetic bulk RNA seq data generated from scRNA-seq glioblas- toma samples, and the second using bulk RNA seq data from tumor biopsies of individuals with glioblastoma. The two results demonstrated the complexity and heterogeneity of glioblastoma tumors. Using a hierarchical strategy for marker gene selection for deconvolution addressed the nature of cell type differentiation. This approach made it possible to have a finer resolution for deconvolution, especially when some rare cell types come from a subpopulation of a very similar cell type. Compared to the traditional way of analyzing all cell types possible at once in one deconvolution run, separating the deconvolution into multiple steps based on cell type developmental trajectory showed decent results.

Inferring cell states based on previous studies from Treehouse bulk RNA tumor samples also has clinical importance in understanding the heterogeneity of glioblastoma tumors. Using cell state signatures generated in an independent scRNA-seq dataset showed the robustness of the rank-based deconvolution pipeline to capture the biological signals.

Chapter 6

Conclusion

Even though cancer is a disease originating in the genome, the cell phenotype is often better classified with transcriptional signatures.

In chapters 2 and 3, first, I built scBeacon, a tool that derives cell type signatures by integrating and clustering multiple scRNA-seq datasets to extract signatures for deconvolving unrelated tumor datasets on bulk samples. Through the employment of scBeacon on the TCGA cohort, we find previously unrecognized cellular and molecular attributes within specific tumor categories, many with patient outcome relevance. We developed a tumor cell-type map (TCT) to visually depict the relationships among TCGA samples based on the cell-type inferences. In Chapter 4, I used a similar deconvolution approach to characterize and estimate the mixed histologies in testicular germ cell cancer and deconvoluted the bulk tumors using cell type signature derived from scRNA-seq data derived from germ cell testicular tissue. I also came up with developmental scores using iPSC data to characterize the differentiation stages in TGCT

samples.

In Chapter 5, I characterized cell type and cell states in glioblastoma using a hierarchical strategy for marker gene selection for deconvolution and addressed the nature of cell type differentiation. To estimate the cell state component in the Treehouse samples, we relied on previous studies on glioblastoma with predefined marker gene sets.

Altogether, I presented a variety of research projects that demonstrate the use of transcriptional signatures in the analysis of tumors and their microenvironment. I described the new insights gained from gene expression analysis about cancer subtypes, cell states, and ultimately patient risk and outcomes.

Bibliography

[1] Bárbara Andrade Barbosa, Saskia D van Asten, Ji Won Oh, Arantza Farina-Sarasqueta, Joanne Verheij, Frederike Dijk, Hanneke WM van Laarhoven, Bauke Ylstra, Juan J Garcia Vallejo, Mark A van de Wiel, et al. Bayesian log-normal deconvolution for enhanced in silico microdissection of bulk gene expression data. *Nature communications*, 12(1):6106, 2021.

[2] Frances R Balkwill, Melania Capasso, and Thorsten Hagemann. The tumor microenvironment at a glance. *Journal of cell science*, 125(23):5591–5596, 2012.

[3] Shideng Bao, Qiulian Wu, Roger E McLendon, Yueling Hao, Qing Shi, Anita B Hjelmeland, Mark W Dewhirst, Darell D Bigner, and Jeremy N Rich. Glioma stem cells promote radioresistance by preferential activation of the dna damage response. *nature*, 444(7120):756–760, 2006.

[4] Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. *Cell systems*, 3(4):346–360, 2016.

[5] Aiman Batool, Nasim Karimi, Xiao-Ning Wu, Su-Ren Chen, and Yi-Xun Liu. Testicular germ cell tumor: a comprehensive review. *Cell Mol Life Sci*, 76:1713–1727, 2019.

[6] Francesco Boniolo, Michael Hoffmann, Nora Roggendorf, Burcu Tercan, Jan Baumbach, Mauro AA Castro, et al. spongeffects: cerna modules offer patient-specific insights into the mirna regulatory landscape. *Bioinformatics*, 2023.

[7] Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019.

[8] Daniel Charytonowicz, Rachel Brody, and Robert Sebra. Interpretable and context-free deconvolution of multi-scale whole transcriptomic data with unicell deconvolve. *Nature communications*, 14(1):1350, 2023.

[9] Jian Chen, Yanjiao Li, Tzong-Shiue Yu, Renée M McKay, Dennis K Burns, Steven G Kernie, and Luis F Parada. A restricted cell population propagates glioblastoma growth after chemotherapy. *Nature*, 488(7412):522–526, 2012.

[10] Tinyi Chu, Zhong Wang, Dana Pe'er, and Charles G Danko. Cell type and gene expression deconvolution with bayesprism enables bayesian integrative analysis across bulk and single-cell rna sequencing in oncology. *Nature Cancer*, 3(4):505–517, 2022.

[11] Cecilia Domínguez Conde, Chao Xu, Louie B Jarvis, Daniel B Rainbow, Sara B Wells, Tamir Gomes, et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376:eabl5197, 2022.

[12] Meichen Dong, Aatish Thennavan, Eugene Urrutia, Yun Li, Charles M Perou, Fei Zou, and Yuchao Jiang. Scdc: bulk gene expression deconvolution by multiple single-cell rna sequencing references. *Briefings in bioinformatics*, 22(1):416–427, 2021.

[13] Rose Du, Vince Carey, and Scott T Weiss. deconvseq: deconvolution of cell mixture distribution in sequencing data. *Bioinformatics*, 35(24):5095–5102, 2019.

[14] Mariella G Filbin, Itay Tirosh, Volker Hovestadt, McKenzie L Shaw, Leah E Escalante, Nathan D Mathewson, Cyril Neftel, Nelli Frank, Kristine Pelton, Christine M Hebert, et al. Developmental and oncogenic programs in h3k27m gliomas dissected by single-cell rna-seq. *Science*, 360(6386):331–335, 2018.

[15] Laura Garcia-Alonso, Vania Lorenzi, Carla I Mazzeo, Joao Paulo Alves-Lopes, Keith Roberts, Maria Sancho-Serra, et al. Single-cell roadmap of human gonadal development. *Nature*, 607:540–547, 2022.

[16] Andrea Garolla, Ugo De Giorgi, and Domenico Milardi. Testicular cancer: New insights on the origin, genetics, treatment, fertility, general health, quality of life and sexual function. *Frontiers in Endocrinology*, 11:514990, 2020.

[17] Armen A Ghazarian, Scott P Kelly, Sean F Altekruse, Philip S Rosenberg, and Katherine A McGlynn. Future of testicular germ cell tumor incidence in the united states: Forecast through 2026. *Cancer*, 123:2320–2328, 2017.

[18] Jingtao Guo, Enrique Sosa, Tsotne Chitiashvili, Xichen Nie, Ernesto Javier Rojas, Elizabeth Oliver, Kathrin Plath, James M Hotaling, Jan-Bernd Stukenborg, Amander T Clark, et al. Single-cell analysis of the developing human testis reveals somatic niche cell specification and fetal germline stem cell establishment. *Cell Stem Cell*, 28(4):764–778, 2021.

[19] Sen Guo and Chu-Xia Deng. Effect of stromal cells in tumor microenvironment on metastasis initiation. *International journal of biological sciences*, 14(14):2083, 2018.

[20] Thomas B Hansen, Torben I Jensen, Bettina H Clausen, Jesper B Bramsen, Bente Finsen, Christian K Damgaard, et al. Natural rna circles function as efficient microrna sponges. *Nature*, 495:384–388, 2013.

[21] Eishu Hirata and Erik Sahai. Tumor microenvironment and differential responses to therapy. *Cold Spring Harbor perspectives in medicine*, 7(7):a026781, 2017.

[22] Michael Hoffmann, Elisabeth Pachl, Marc Hartung, and Verena Stiegler. Spongedb: a pan-cancer resource for competing endogenous rna interactions. *Narodonaselenie*, 2021.

[23] Michael Hoffmann, Lukas Schwartz, Ovidiu-Andrei Ciora, Nicole Trummer, Lena-Luise Willruth, Julius Jankowski, et al. circrna-sponging: a pipeline for extensive analysis of circrna expression and their role in mirna sponging. *Bioinform Adv*, 3:vbad093, 2023.

[24] Brandon Jew, Marcus Alvarez, Elior Rahmani, Zong Miao, Arthur Ko, Kristina M Garske, Jae Hoon Sul, Kirsi H Pietiläinen, Päivi Pajukanta, and Eran Halperin. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature communications*, 11(1):1971, 2020.

[25] Ramesh V Kartha and S Subramanian. Competing endogenous rnas (cernas): new entrants to the intricacies of gene regulation. *Front Genet*, 5:8, 2014.

[26] Meng Li, Wei Ding, Muhammad Akram Tariq, Wei Chang, Xiaowen Zhang, Wenbin Xu, et al. A circular transcript of ncx1 gene mediates ischemic myocardial injury by targeting mir-133a-3p. *Theranostics*, 8:5855–5869, 2018.

[27] Markus List, Azadeh Dehghani Amirabad, Dennis Kostka, and Marcel H Schulz. Large-scale inference of competing endogenous rna networks with sparse partial correlation. *Bioinformatics*, 35:i596–i604, 2019.

[28] Bogdan A Luca, Chloé B Steen, Magdalena Matusiak, Armon Azizi, Sushama Varma, Chunfang Zhu, Joanna Przybyl, Almudena Espín-Pérez, Maximilian Diehn, Ash A Alizadeh, et al. Atlas of clinically distinct cell states and ecosystems across human solid tumors. *Cell*, 184(21):5482–5496, 2021.

[29] Seiji Morinaga and Noboru Sasano. Histopathology of testicular germ cell tumors. *Gan To Kagaku Ryoho*, 11:2460–2467, 1984.

[30] Cyril Neftel, Julie Laffy, Mariella G Filbin, Toshiro Hara, Marni E Shore, Gilbert J Rahme, Alyssa R Richman, Dana Silverbush, McKenzie L Shaw, Christine M Hebert, et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell*, 178(4):835–849, 2019.

[31] Irene Papatheodorou, Pablo Moreno, Jonathan Manning, Alfonso Muñoz-Pomer Fuentes, Nancy George, Silvie Fexova, Nuno A Fonseca, Anja Füllgrabe, Matthew Green, Ni Huang, et al. Expression atlas update: from tissues to single cells. *Nucleic acids research*, 48(D1):D77–D83, 2020.

[32] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. Single-cell rna-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

[33] Jeff S Pawlikowski, Tony McBryan, John van Tuyn, Mark E Drotar, Rachael N Hewitt, Andrea B Maier, Ayala King, Karen Blyth, Hong Wu, and Peter D Adams. Wnt signaling potentiates nevogenesis. *Proceedings of the National Academy of Sciences*, 110(40):16009–16014, 2013.

[34] Gregory D Poore, Evguenia Kopylova, Qiyun Zhu, Carolina Carpenter, Serena Fraraccio, Stephen Wandro, Tomasz Kosciolek, Stefan Janssen, Jessica Metcalf, Se Jin Song, et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature*, 579(7800):567–574, 2020.

[35] Aleix Prat, Estela Pineda, Barbara Adamo, Patricia Galván, Aranzazu Fernández, Lydia Gaba, Marc Díez, Margarita Viladot, Ana Arance, and Montserrat Muñoz. Clinical implications of the intrinsic molecular subtypes of breast cancer. *The Breast*, 24:S26–S35, 2015.

[36] Sidharth V Puram, Itay Tirosh, Anuraag S Parikh, Anoop P Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, Christina L Luo, Edmund A Mroz, Kevin S Emerick, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, 171(7):1611–1624, 2017.

[37] Emad A Rakha, Jorge S Reis-Filho, and Ian O Ellis. Basal-like breast cancer: a critical review. *Journal of clinical oncology*, 26(15):2568–2581, 2008.

[38] Megan R Reed, A Geoffrey Lyle, Annick De Loose, Leena Maddukuri, Katrina Learned, Holly C Beale, Ellen T Kephart, Allison Cheney, Anouk van den Bout, Madison P Lee, et al. A functional precision medicine pipeline combines comparative transcriptomics and tumor organoid modeling to identify bespoke treatment strategies for glioblastoma. *Cells*, 10(12):3400, 2021.

[39] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. The human cell atlas. *elife*, 6:e27041, 2017.

[40] Andrew D Rouillard, Gregory W Gundersen, Nicolas F Fernandez, Zichen Wang, Caroline D Monteiro, Michael G McDermott, and Avi Ma'ayan. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, 2016, 2016.

[41] Leonardo Salmena, Laura Poliseno, Yvonne Tay, Lev Kats, and Pier Paolo Pandolfi. A cerna hypothesis: the rosetta stone of a hidden rna language? *Cell*, 146:353–358, 2011.

[42] Radoslav Savić, Jialiang Yang, Simon Koplev, Mahru C An, Priyanka L Patel, Robert N O'Brien, Brittany N Dubose, Tetyana Dodatko, Eduard Rogatsky, Katyayani Sukhavasi, et al. Integration of transcriptomes of senescent cell models with multi-tissue patient samples reveals reduced col6a3 as an inducer of senescence. *Cell Reports*, 42(11), 2023.

[43] Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, and Ahmedin Jemal. Cancer statistics, 2021. *CA: a Cancer Journal for Clinicians*, 71(1):7–33, 2021.

[44] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

[45] Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.

[46] Roel GW Verhaak, Katherine A Hoadley, Elizabeth Purdom, Victoria Wang, Yuan Qi, Matthew D Wilkerson, C Ryan Miller, Li Ding, Todd Golub, Jill P Mesirov, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in pdgfra, idh1, egfr, and nf1. *Cancer cell*, 17(1):98–110, 2010.

[47] Alexandra-Chloé Villani, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, Andrew Butler, Shiwei Zheng, Suzan Lazo, et al. Single-cell rna-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335):eaah4573, 2017.

[48] Qianghu Wang, Baoli Hu, Xin Hu, Hoon Kim, Massimo Squatrito, Lisa Scarpace, Ana C DeCarvalho, Sali Lyu, Pengping Li, Yan Li, et al. Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. *Cancer cell*, 32(1):42–56, 2017.

[49] Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, 10(1):380, 2019.

[50] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna RM Shaw, Bradley A Ozenberger, Kyle Ellrott, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet*, 45:1113–1120, 2013.

[51] Mengyu Xie, Kyubum Lee, John H Lockhart, Scott D Cukras, Rodrigo Carvajal, Amer A Beg, Elsa R Flores, Mingxiang Teng, Christine H Chung, and Aik Choon Tan. Timex: tumor-immune microenvironment deconvolution web-portal for bulk transcriptomics using pan-cancer scrna-seq signatures. *Bioinformatics*, 37(20):3681–3683, 2021.

[52] Dong Zhang, Ying Yu, Tianzhi Duan, and Qing Zhou. The role of macrophages in reproductive-related diseases. *Heliyon*, 8:e11686, 2022.

[53] Yuanyuan Zhang and Zemin Zhang. The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications. *Cellular & molecular immunology*, 17(8):807–821, 2020.

[54] Zhicheng Zhang, Teng Yang, and Jing Xiao. Circular rnas: Promising biomarkers for human diseases. *EBioMedicine*, 34:267–274, 2018.

[55] Shuzhen Zhao, Wei Zhu, Shanshan Xue, and Dong Han. Testicular defense systems: immune privilege and innate immunity. *Cell Mol Immunol*, 11:428–437, 2014.