2019

Pathway activity analysis of bulk and single-cell RNA-Seq data

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

PATHWAY ACTIVITY ANALYSIS OF
BULK AND SINGLE-CELL RNA-SEQ DATA

by

DAVID FOWLER JENKINS III

Sc.B., Brown University, 2011
M.S., Boston University, 2017

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2019

Approved by

First Reader

W. Evan Johnson, Ph.D.
Associate Professor of Medicine and Biostatistics

Second Reader

Joshua D. Campbell, Ph.D.
Assistant Professor of Medicine

DEDICATION

To my grandparents

Paul, Loretta, William, and Antoinette.

ACKNOWLEDGMENTS

My thesis work would not have been possible without a team of people helping me along the way, so I'd like to say thank you. First and foremost, to my advisor Evan, whose infectious enthusiasm I felt from the first day we met in 2014. I left every meeting with you inspired to work hard, and I won't ever forget that, thank you. To my collaborators, particularly Andrea Bild, Moom Rahman, and Shelley MacNeil, I'm proud of the work we did together, thank you. To my fellow lab mates, especially Tyler, Supriya, Mani, Yuqing, and Jason, thanks for your positive feedback and support. To my thesis committee, Paola, Josh, Jen, and Masanao, thanks for your suggestions and flexibility, it was a pleasure to work with you. To the BU Bioinformatics program, particularly Caroline Lyman, Johanna Vasquez, and David King, you have been more than generous with your time and resources; we all really appreciate the work you do to make our graduate work go as smoothly as possible. To my friends, both old and new, for always being there to make me laugh or listen to me when I needed you, thank you. Finally, to my family, Dave, Caroline, Emma, Sam, Gram L, and Paula, I love you.

PATHWAY ACTIVITY ANALYSIS OF BULK AND SINGLE-CELL RNA-SEQ DATA

DAVID FOWLER JENKINS III

Boston University Graduate School of Arts and Sciences

and College of Engineering, 2019

Major Professor: W. Evan Johnson, Associate Professor of Medicine and Biostatistics

ABSTRACT

Gene expression profiling can produce effective biomarkers that can provide additional information beyond other approaches for characterizing disease. While these approaches are typically performed on standard bulk RNA sequencing data, new methods for RNA sequencing of individual cells have allowed these approaches to be applied at the resolution of a single cell. As these methods enter the mainstream, there is an increased need for user-friendly software that allows researchers without experience in bioinformatics to apply these techniques. In this thesis, I have developed new, user-friendly data resources and software tools to allow researchers to use gene expression signatures in their own datasets. Specifically, I created the Single Cell Toolkit, a user-friendly and interactive toolkit for analyzing single-cell RNA sequencing data and used this toolkit to analyze the pathway activity levels in breast cancer cells before and after cancer therapy. Next, I created and validated a set of activated oncogenic growth factor receptor signatures in breast cancer, which revealed additional heterogeneity within public breast cancer cell line and patient sample RNA sequencing datasets. Finally, I created an R package for rapidly profiling TB samples using a set of 30 existing tuberculosis gene signatures. I applied this tool to look at pathway differences in a dataset of tuberculosis treatment failure samples. Taken together, the results of these studies serve as a set of user-friendly software tools and data sets that allow researchers to rapidly and consistently apply pathway activity methods across RNA sequencing samples.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANOVA	Analysis of variance
AR	Androgen receptor
ASSIGN	Adaptive Signature Selection and InteGratioN
ATCC	American Type Culture Collection
AUC	Area under the curve
BAD	Bcl-2-associated death promoter
BC	Breast cancer
BCA	Bicinchoninic acid
BCL-2	B-cell lymphoma 2 gene
BRCA	Breast cancer
cDNA	Complementary DNA
CDR	Cellular detection rate
CLARA	Clustering for Large Applications
CPM	Counts per million
DE	Differential expression
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DOTS	Directly observed treatment short course strategy
EC50	Concentration of a drug that gives half-maximal response
EGFR	Epidermal growth factor receptor
EBI	European Bioinformatics Institute
EDTA	Ethylenediaminetetraacetic acid

EMT	Epithelial–mesenchymal transition
ER	Estrogen receptor
FBS	Fetal bovine serum
FDR	False discovery rate
FISH	Fluorescence in situ hybridization
FPKM	Fragments per kilobase million
GEO	Gene Expression Omnibus
GFP	Green fluorescent protein
GFRN	Growth factor receptor network
GSEA	Gene Set Enrichment Analysis
GSVA	Gene Set Variation Analysis
GUI	Graphical user interface
HER2	Human epidermal growth factor receptor 2
HMECs	Human primary mammary epithelial cells
ICBP	Integrative Cancer Biology Program
IGF1R	Insulin-like growth factor 1 receptor
IHC	Immunohistochemistry
KEGG	Kyoto Encyclopedia of Genes and Genomes
LOOCV	Leave-one-out cross-validation
LTBI	Latent tuberculosis infection
MAIT	Mucosal associated invariant T cell
MAPK	Mitogen-activated protein kinase
MAST	Model-based analysis of single-cell transcriptomics

MCL-1	Induced myeloid leukemia cell differentiation protein Mcl-1 gene
MEBM	Mammary epithelial basal medium
mRNA	Messenger RNA
MSigDB	The Molecular Signatures Database
mTOR	Mammalian target of rapamycin
NA	Not available
NCBI	National Center for Biotechnology Information
PAM50	Point accepted mutation 50 matrix
PBS	Phosphate-buffered saline
PC	Principal component
PCA	Principal component analysis
PI3K	Phosphoinositide 3-kinase
PR	Progesterone receptor
PVDF	Polyvinylidene difluoride
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
ROC	Receiver operating characteristic
RPPA	Reverse phase protein array
RTK	Receptor tyrosine kinase
scRNA-Seq	Single-cell RNA sequencing
ssGSEA	Single sample gene set enrichment analysis
SCTK	Single Cell Toolkit
SDS	Sodium dodecyl sulfate

t-SNE	T-distributed stochastic neighbor embedding
TB	Tuberculosis
TCGA	The Cancer Genome Atlas
TNBC	Triple negative breast cancer
TPM	Transcripts per kilobase million
UCSC	University of California, Santa Cruz

Chapter 1. Introduction

Transcriptional Regulation and Disease

It is well established that cell signaling plays an important role in disease (Nahta, Hortobágyi & Esteva, 2003). The cell is a highly regulated and carefully controlled machine, with many signaling pathways working together to address the needs of each cell (Schlessinger, 2000). When these pathways are disrupted, there can be devastating effects. When attempting to quantify these changes in pathway activity there are several levels to probe. Alterations in DNA (DNA mutations) can cause changes in pathway activity, deactivating or activating a specific transcription factor, but redundancy in transcription factor pathways means that not every mutational event will have a strong, if any, effect (Spitz, Furlong, 2012). Further, many DNA mutations occur in positions in the genome that have no effect, so sifting through many mutations to identify the causal change is often difficult. An alternative approach is to quantify RNA expression. Messenger RNA (mRNA) is the transcriptional language that converts the instructions coded in DNA into protein products. By quantifying the level of RNA expression for each gene in a specific sample, a picture of the cellular activity of these genes can be identified. Differences in the levels of gene expression across samples can point to the causes of disease phenotypes or be used as a biomarker of a specific disease state. Often these biomarkers are not a single value, but a set of coordinately expressed genes that form a signature that can represent the activity of a cellular component, such as a pathway. By identifying biomarkers of disease, we can stratify patients into groups with similar cellular activity, which can often respond to disease treatments in similar ways (Groenendijk, Bernards, 2014, McCubrey et al., 2012).

Transcriptional Pathways in Cancer

In cancer, activating or inactivating mutations can disrupt the cell signaling cascades that control how cells grow, divide, and undergo apoptosis (McCubrey et al., 2012). These changes in cell signaling networks have become an important part of the set of 'hallmarks of cancer,' causing cells to grow uncontrollably, form tumors, and metastasize (Hanahan, Weinberg, 2011). Cancer is one of the leading causes of death globally with an estimated 18.1 million new cases in 2018 (The International Agency for Research on Cancer, 2018). Fortunately, targeted cancer drugs to address specific abnormalities in cancer signaling pathways have been developed (Gustafson et al., 2010). These targeted drugs can inhibit certain signaling pathways driven by key oncogenic growth factor receptors, such as EGFR and HER2 (Nahta, Hortobágyi & Esteva, 2003). Importantly, by identifying the signaling cascade that is driving a specific tumor, drugs that target components of that pathway can be administered. The ultimate goal of pathway analysis in cancer is to identify a set of biomarkers that can precisely characterize the drivers of a specific tumor and identify the drugs that would best target the exact combination of aberrations in a specific patient's tumor, an approach commonly known as personalized medicine.

Tuberculosis

Tuberculosis (TB) infection is a leading cause of death worldwide (World Health Organization, 2016). The majority of patients infected with TB will not progress to active TB disease (World Health Organization, 2016). Of those that do get infected, some will fail their treatment. In TB, many gene signatures have been produced that can accurately predict the likelihood of TB progression or predict active TB disease (Zak et al., 2016, Sambarey et al., 2017, Leong et al., 2018). These pathways typically contain genes involved in the immune and inflammatory responses (Scriba et al., 2017). Similar to their use in cancer, gene signatures of TB can help stratify patients into groups that are likely to progress to disease and those that are unlikely to get disease, monitor patient adherence to drug regimens, and ensure that infections are being successfully treated. This could be particularly important in situations where TB drugs are scarce, or resources for TB treatment are reduced, and help improve outcomes for TB treatment.

RNA Sequencing

RNA sequencing leverages high throughput sequencing technologies to quantify the gene expression levels in a sample. Standard sequencing pipelines involve aligning reads to a reference genome and counting the number of genes that overlap with each gene or transcript annotation. These raw counts can then be normalized to correct for differences in sequencing depth between samples or corrected for unwanted experimental variations called batch effects (Johnson, Li & Rabinovic, 2007) . The normalized counts can then be merged into a matrix of counts per sample for downstream analysis, which often involves identifying significantly differentially expressed pathways or gene signatures. Within the R programming language, several software tools have been created to make the storage of gene expression data easier. The SummarizedExperiment object allows for storage of multiple matrices which can be used to store sample and gene annotation data along with raw and normalized count data (Huber et al., 2015) . The object can be subset, automatically subsetting the annotation data along with the count data to make sure everything remains in sync.

Single-Cell RNA Sequencing

Typical bulk RNA sequencing combines the expression of genes from all cells in a sample. Recently, new techniques for performing single-cell RNA-Seq have been developed. These techniques involve either physically separating cells into individual wells in a plate or performing highly multiplexed bead-based library preparation for higher throughput results (Picelli et al., 2013, Macosko et al., 2015). The end result is expression data for an individual cell, allowing researchers to probe differences in gene expression across cell types or tumor subgroups. Due to the low amount of starting material for each individual cell, scRNA-Seq shows lower gene expression levels than typical bulk RNA-Seq datasets and some genes display a bimodal pattern of expression. To address these concerns, additional filtering and normalization steps are needed before standard RNA-Seq analysis techniques can be performed on scRNA-Seq datasets (Brennecke et al., 2013). Further, novel analysis methodologies that take into account the missingness that typically arises in scRNA-Seq data have been developed (Finak et al., 2015, Trapnell et al., 2014, Satija et al., 2015). Choosing which analysis methods to use can be dataset specific, and often involves iterating through several analysis techniques before settling on the best approach for a given dataset.

Dissertation Aims

The aims in this dissertation seek to develop novel software frameworks and pathway signatures to aid in the analysis of bulk and single-cell RNA-Seq datasets in the context of disease, specifically breast cancer and tuberculosis. Together, these aims will show that by creating interactive and intuitive tools for data processing, users can perform sophisticated analysis on RNA-Seq datasets without needing to write code or have a deep understanding of how to run the standard underlying algorithms for RNA-Seq analysis.

Aim 1: Create a user-friendly interface and a full-featured analysis toolkit for single-cell RNA-Seq datasets

Many tools for performing single-cell RNA-Seq (scRNA-Seq) analysis exist, but these tools are often only available on the command line and require significant bioinformatics expertise to use. While other software tools for analysis and visualization exist, there has yet to be a full scRNA-Seq analysis tool to help users take raw data through a standard pipeline to produce downstream analysis including quality control and filtering, visualization with dimensionality reduction methods, differential expression analysis, and pathway activity and gene enrichment approaches. In this aim, I present the Single Cell Toolkit (SCTK), the first fully interactive scRNA-Seq analysis tool written in R and Shiny. This tool allows users to perform a full scRNA-Seq analysis pipeline through an intuitive point-and-click interface, allowing improved access to scRNA-Seq analysis tools.

Aim 2: Create and apply oncogenic growth factor receptor network signatures across breast cancer cell lines and breast cancer patient tumor samples

Cell line derived gene expression signatures have been used to identify signatures of pathway activity in cancer samples (Bild et al., 2006). These signatures can then be used to stratify samples by cellular activity and predict the effectiveness of drugs that target activated oncogenic pathways. In this aim, I describe a new set of pathway activity signatures of breast cancer oncogenes in growth factor receptor networks. Pathway activity predictions were performed using Adaptive Signature Selection and Integration (ASSIGN) and can be run automatically through extensions of the ASSIGN R package (Shen et al., 2015). This set of signatures was applied to cancer cell line panels and patient breast cancer tumor samples, revealing additional heterogeneity within the cohorts and significant correlations to differences in drug response.

Aim 3: Collect available biomarkers of tuberculosis disease and progression, create an analysis framework to apply these signatures, and profile the pathway activity in a cohort of tuberculosis treatment failure samples

Several signatures of TB have been previously published and can accurately predict several aspects of TB progression into active disease or predict the effectiveness of TB treatment. Since numerous unique signatures have been developed, it is worthwhile to explore differences in pathway activity across several signatures rather than looking at them individually. In this aim, a set of 30 previously published gene signatures of TB were collected. To rapidly profile this set of 30 gene signatures we created the TB Signature Profiler, a software framework to easily profile a set of samples with a set of user defined signatures using common pathway activity prediction algorithms. With this tool, users can profile and visualize the pathway activity predictions easily, leveraging the SummarizedExperiment object within R to store raw data and pathway activity scores together. We used the TB Signature Profiler on a set of TB samples from treatment failure patients and identified heterogeneity that showed the published signatures can accurately show TB treatment response and highlight issues with adherence to drug treatment.

Chapter 2. An analysis toolkit for single-cell RNA-Seq data

Adapted from the following manuscript:

David F. Jenkins, Tyler Faits, Mohammed Muzamil Khan, Emma Briars, Sebastian Carrasco Pro, Steve Cunningham, Joshua D. Campbell, Masanao Yajima, and W. Evan Johnson. *(Manuscript submitted)*

Introduction

Single-cell RNA sequencing (scRNA-Seq) techniques allow researchers to explore the transcriptional landscape of a sample at the resolution of the individual cell. In the context of cancer, scRNA-Seq can identify the subclonality of a tumor sample to improve our ability to identify the cell-specific mechanisms that drive tumor growth and can characterize different cellular populations within the tumor microenvironment (Tirosh et al., 2016, Brady et al., 2017). However, different optimizations of parameters and algorithms are required for filtration, normalization, clustering, and differential expression of scRNA-Seq data compared to bulk RNA-Seq due to the low amount of starting material and technical bias introduced in the common scRNA-Seq library preparation techniques (Brennecke et al., 2013). Tools for normalization and analysis of scRNA-Seq data exist to overcome these technical biases, but these tools are not integrated and require command line processing of samples and knowledge of the many options available for each tool, which makes them difficult to use, especially for scientists without training in bioinformatics (McCarthy et al., 2017, Nakamura et al., 2015, Satija et al., 2015, Kharchenko, Silberstein & Scadden, 2014, Fan et al., 2016, Trapnell et al., 2014). Even for more advanced users, there is still a need to interactively explore scRNA-Seq results during processing to help make dataset specific decisions that can affect downstream analysis.

Shiny is an R package and toolkit developed by RStudio (https://www.rstudio.com) that allows for the creation of web based graphical user interfaces (GUIs) over R packages, allowing for interactive data exploration and analysis through familiar drop down menus and buttons (Chang et al., 2017). Users can load a Shiny app locally on their computer or the Shiny app can be hosted in the cloud and can be accessed through a web browser.

Package	**SCATER**	**SC3**	**SEURAT**	**SCDE**	**PAGODA**	**MONOCLE**	**SCTK**
Filtering and Data Summary	✓		✓				✓
Dimensionality Reduction	✓		✓			✓	✓
Clustering		✓	✓		✓	✓	✓
Batch Correction			✓	✓		✓	✓
Differential Expression		✓	✓	✓		✓	✓
Pathway Enrichment			✓		✓		✓
Experimental Design							✓
GUI	✓	✓			✓		✓
SingleCellExperiment Support	✓	✓					✓

Table 2.1. Comparison of SCTK and other popular scRNA-Seq analysis tools. While SCATER (McCarthy et al., 2017), SC3 (Nakamura et al., 2015), SEURAT (Satija et al., 2015), SCDE (Kharchenko, Silberstein & Scadden, 2014) , PAGODA (Fan et al., 2016), and MONOCLE (Trapnell et al., 2014) accomplish some steps in the scRNA-Seq analysis pipeline, the SCTK supports a full interactive scRNA-Seq analysis workflow and supports the SingleCellExperiment object for data storage.

Here, we present the Single Cell Toolkit (SCTK), an R/Shiny based package for both command line and interactive scRNA-Seq processing. While other tools can perform specific scRNA-Seq analysis steps, the SCTK is the first fully interactive scRNA-Seq analysis workflow available within the R language (Table 2.1). We applied the SCTK and our workflow on multiple data examples, including stimulated and unstimulated mucosal-associated invariant T cells, induced pluripotent stem cells from Yoruba male reference samples to identify batch effects, and tumor cells from breast cancer patients to identify pathway activity in response to treatment (Tung et al., 2017, Finak et al., 2015, Brady et al., 2017).

Methods

The SCTK is organized into several analysis modules. All modules can be run interactively through the Shiny web interface or through the R console. Below we describe the datasets available in the SCTK, the underlying architecture of the SCTK, and the analysis modules available through the interactive SCTK package and GUI.

Mucosal-associated Invariant T (MAIT) Cells

To demonstrate how interactive analysis can be performed in the SCTK, an example dataset of mucosal-associated invariant T (MAIT) cells was used (Finak et al., 2015). A set of 96 CD8+ MAIT cells were sorted, 47 cells were stimulated with cytokines, and the cells were processed and sequenced using the Fluidigm C1 system. The data was aligned to the human genome, quantified, and included with the MAST package. Cytokine stimulation of MAIT cells results in increased cytokine gene expression and pathway activity changes that can be identified with differential expression analysis and pathway activity analysis, which can serve as an effective control for our toolkit methods if cytokine genes and cytokine containing pathways are identified through analysis.

Pluripotent Stem Cells

A dataset demonstrating batch effects in single-cell data was created by Tung, et. al (Tung et al., 2017). Three induced pluripotent stem cell lines were sequenced in triplicate on the Fluidigm C1 platform using a total of 9 plates. The resulting data has a clear plate effect that represents an experimental batch effect that could affect downstream analysis if it is not corrected. After removing the batch effect, the experimental replicates should not separate during analysis, allowing the data to be used to identify biological differences between the individuals.

Data Structure

Steps in the analysis pipeline are performed on a SCTKExperiment object, an extension of the SingleCellExperiment and RangedSummarizedExperiment objects developed by the Bioconductor project (Huber et al., 2015). This object is organized into identically sized matrices designed to store counts, normalized counts, or batch corrected data; a data frame for sample annotation information; and a data frame for feature annotation information. These objects allow users to keep their scRNA-Seq data organized in a single object that automatically resizes all matrices and annotation information if the data is modified, ensuring annotation information and count data is always in sync. Additionally, data from dimensionality reduction approaches such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) can be stored in the object's reducedDims slot. The SingleCellExperiment object has been optimized to store large datasets by using sparse matrices and an efficient API to support data that would otherwise not fit into memory using a standard matrix (Lun, Pagès & Smith, 2018). Depending on the size of the data stored in the object, the matrices used in the object are automatically stored as either standard R matrices, sparse matrices, or on disk as a HDF5 file backed matrix. This allows users to take advantage of these memory saving strategies automatically without needing to specify which type of matrix that should be used on their dataset. The SingleCellExperiment can also store information about spike-in transcripts and sample specific size factors for normalization. By utilizing an object that can efficiently store both raw data and downstream analysis results, analysis can be performed within the SCTK, saved into the object, and loaded into R for additional analysis on the command line.

The SCTKExperiment is implemented as an S4 object, an object-oriented system available within R. This allows users to inherit methods and structure that exists in other S4 objects and add additional functionality while still being backwards compatible. The SCTKExperiment object also stores the percent variation explained by each principal component. This is accomplished by adding an additional slot named pcaVariances to the SCTKExperiment object that stores the percent variation explained by each principal component (PC) as a DataFrame. The `getPCA()` function available in the SCTK saves the PCs into the reducedDims slot and additionally stores the percent variation explained in the pcaVariances slot. This data can be accessed using the `pcaVariances()` function. The SCTKExperiment object can be further extended and will continue to be expanded in future versions of the SCTK to store additional single-cell data, annotations, and results.

Data Upload

After installing the SCTK, users can start the Shiny app by running the `singleCellToolkit()` function with a SCTKExperiment object as an input to automatically load the data into the app. Alternatively, a user can choose to upload a data matrix of raw count or normalized data directly through the Shiny app by uploading a text file, along with optional sample and feature annotation files. The SCTK will create a

SCTKExperiment object to store the toolkit analysis results. This object can be exported after analysis has been completed.

Data Summary and Filtering

After scRNA-Seq data has been loaded into the SCTK, a table of data summary metrics is presented. Because scRNA-Seq data is very sparse, dataset specific filtration and normalization can affect downstream analysis (Stegle, Teichmann & Marioni, 2015). In the data summary and filtering tab, the SCTK provides users with several summary statistics and options for manipulating their data and annotation information. First, within the Data Summary subtab, the SCTK displays a table of summary metrics including the number of samples, number or genes, average number of reads per cell, average number of genes per cell, and the number of genes with few or no counts across all samples. Additionally, histograms of the number of counts per sample and the number of expressed genes per sample are displayed. If the dataset is small, containing less than 50 cells, the entire data matrix is also displayed. Using this information, the user can make decisions about how best to filter their data for downstream analysis. Users can filter genes and samples with low or no expression, delete outlier samples, filter the dataset based on annotation information, and modify the annotation information by uploading a replacement annotation matrix. For larger datasets, users can also randomly subset their data on this tab, allowing the user to perform exploratory analysis on a reduced dataset within the SCTK. The filtering applied while using this tab modifies the underlying data that is used throughout the app. A snapshot of the original uploaded data is preserved so a user can always return to the original uploaded data to restart the analysis or try a different filtration protocol.

The SCTK doesn’t require users to import their data with a specific set of gene annotation information, but some tools within the SCTK are only available if the data uses specific gene annotations. Depending on the reference genome that is used during sequence alignment and quantification, users may have data that describes their genes using gene symbols (e.g. BRCA1), Entrez gene numeric IDs from the NCBI database (e.g. 672 for BRCA1), Ensembl gene IDs from the EBI (e.g. ENSG00000012048 for BRCA1), or from another source. The SCTK has the ability to convert gene ids to various formats using the org.*.eg.db Bioconductor annotation packages. These packages are not installed by default, so these must be manually installed before this function will work. After these packages have been installed, users can convert between the available gene annotations on the Data Summary and Filtering tab.

Additional modifications to the underlying data object are available in the Assay Details subtab. Lists of the available data matrices and reduced dimension data are displayed in the Assay Details tab. Users can add additional data matrices to their data object. Any existing matrix can be log-transformed, or if raw count data is available, a counts per million (CPM) normalization can be applied. Unwanted data matrices or reduced dimensionality can be deleted in this tab.

Finally, in the Visualize subtab, users can visualize gene expression data versus annotation data for genes of interest using a boxplot, scatterplot, barplot, or heatmap depending on the type of annotation data information that is available. This can be helpful for visualizing housekeeping genes for sample quality control, quantifying artificial spike-in controls for sequencing quality control, or visualizing individual genes of interest.

Dimensionality Reduction and Clustering

Visualization of scRNA-Seq data is crucial to identifying subclusters of cells present in the data. Dimensionality reduction techniques allow a user to visualize scRNA-Seq data by summarizing the observed variation into lower dimension space. PCA transforms the matrix into components that describe the variation observed in the data. An alternative to PCA, t-distributed stochastic neighbor embedding (t-SNE), is also frequently used when analyzing scRNA-Seq data because it is able to embed a large amount of variation into a small number of dimensions (Van, Hinton, 2008). When users open the dimensionality reduction and clustering tab in the SCTK, a list of available reduced dimension datasets and algorithms is provided. Because these algorithms can take a long time to compute on large datasets, users can precompute the reduced dimension data and store it in a SCTKExperiment object before uploading the data into the SCTK. For smaller datasets, users can perform PCA and t-SNE directly through the SCTK app. The resulting reduced matrices will be stored in the underlying object that can be downloaded when analysis is complete. The resulting data can be displayed in the dimensionality reduction and clustering tab. Annotation information can be added to the plot by selecting annotations with which to color or shape the points in the scatterplot.

After visualization of the data, users may want to stratify the scRNA-Seq data into clusters that appear during dimensionality reduction. Users can choose to cluster their data using k-means clustering, hierarchical clustering, or CLARA (Clustering for Large Applications). Clustering is typically performed on the PCA data, because t-SNE data does not retain the distance between clusters in its results. After the clustering algorithm is complete, the plot is automatically updated to display the resulting clusters. If the user wants to save the cluster results, the cluster assignments can be stored in the annotation data frame of the SCTKExperiment object and visualized on other reduced dimension data. Additionally, other clustering algorithms can be run on the command line, saved as annotation information in the SCTKExperiment object, and visualized in this tab.

ComBat Batch Correction

Because of the complexities of the library preparation and the low starting material in scRNA-Seq experiments, non-biological variation (batch effects) are present and can be a major source of variation present in single-cell experiments (Hicks et al., 2017). ComBat is a widely used method for adjusting for batch effects in microarray and RNA-Seq data (Johnson, Li & Rabinovic, 2007). If users identify variation associated with a technical effect, ComBat can be run within the SCTK to remove this variation before further downstream analysis. Users can choose an annotation present in the annotation data frame and add additional covariates to the ComBat model before performing batch correction. After batch correction, the ComBat results are stored as an additional assay in the SCTKExperiment object, which can then be used in the other analysis tabs within the SCTK.

Differential Expression and Biomarker Creation

Differential expression analysis can identify genes that are significantly up or down regulated between conditions. While many differential expression algorithms exist, their performance may vary on scRNA-Seq datasets. Users can apply common differential expression algorithms limma (Ritchie et al., 2015), DESeq2 (Love, Huber & Anders, 2014), or perform an ANOVA to identify differentially expressed genes by selecting one or multiple condition variables present in the annotation information. Users can customize the differential expression results by changing the number of genes to return, the p-value significance cutoff, and the p-value correction method applied to the results. The resulting gene list is displayed as a table and also in a heatmap which can also be customized. Users can download the gene list directly or create a biomarker list for a specific cell type or cell cluster, which can be stored in the gene annotation information in the SCTKExperiment object.

Single-cell RNA-Seq specific tools for differential expression have been developed that can adjust for some of the characteristics of scRNA-Seq data. MAST, Model-based Analysis of Single-cell Transcriptomics, has been developed to address these issues by using a hurdle model (Finak et al., 2015). A hurdle model allows for separate accounting of the processes that produce zero count values, and the ones that produce the positive count values. MAST allows users to identify this cutoff by using an adaptive threshold model that bins genes based on gene expression and identifies a cutoff for zero expression. This allows the dropout rate typical of scRNA-Seq data to be modelled. Additionally, MAST models the cellular detection rate (CDR), a measure of the percent of genes that are expressed in a given sample. Adding the CDR to the model can correct for biological and technical covariates when identifying differences in the condition of interest. MAST has been implemented within the SCTK. Users can choose whether to use MAST's adaptive thresholding model, choose fold change and expression thresholds, and identify significant genes based on conditions present in the annotation information provided. The results are presented in a table, violin plots, or visualized in a heatmap and can be saved as a biomarker in the SCTKExperiment object or downloaded directly.

Subsampling and Differential Power Analysis

The relative complexity of scRNA-Seq experimental designs makes it difficult for investigators to ensure that an experiment will have sufficient power while operating on a finite budget. Whereas there are tools for optimizing bulk RNA-Seq designs (Busby et al., 2013, Guo et al., 2014), these fail to account for the tiered nature of scRNA-Seq experiments, where each biological replicate may contribute any number of cells to be sequenced, each of which may belong to one of many cell types or subpopulations. Users of the SCTK can project estimated power metrics based on their dataset with variable simulated parameters including sequencing depth, number of sequenced cells, and number of biological replicates. To produce results within a reasonable timespan, the Shiny interface only allows users to vary one parameter at a time while keeping the others fixed. The command line allows users to probe all parameters at once, producing multidimensional power estimates which will help investigators optimize their scRNA-Seq experimental designs.

Pathway Activity Analysis

Gene expression measurements can be summarized into a signature or set of genes to create a score that represents the activity of that set of genes in a sample. By summarizing genes in known signaling pathways, cells with active signaling pathways or specific cellular functions can be identified. Gene Set Variation Analysis (GSVA) uses gene sets to create these signatures (Hänzelmann, Castelo & Guinney, 2013). The molecular signature database (MSigDb) is a database of molecular signatures that can be used in GSVA (Liberzon et al., 2011). GSVA has been implemented in the SCTK. Users can select their input data, gene set(s), and GSVA parameters interactively through the app. GSVA can be run across all MSigDB signatures, a user selected subset of MSigDB signatures, or a set of custom gene signatures saved as annotation columns in the rowData slot of the SCTKExperiment object. After GSVA is complete, scores will be displayed in either violin plots or a heatmap on the Pathway Activity tab of the SCTK. Users can save the pathway activity scores into the annotation data columns of the SCTKExperiment object or download the scores directly.

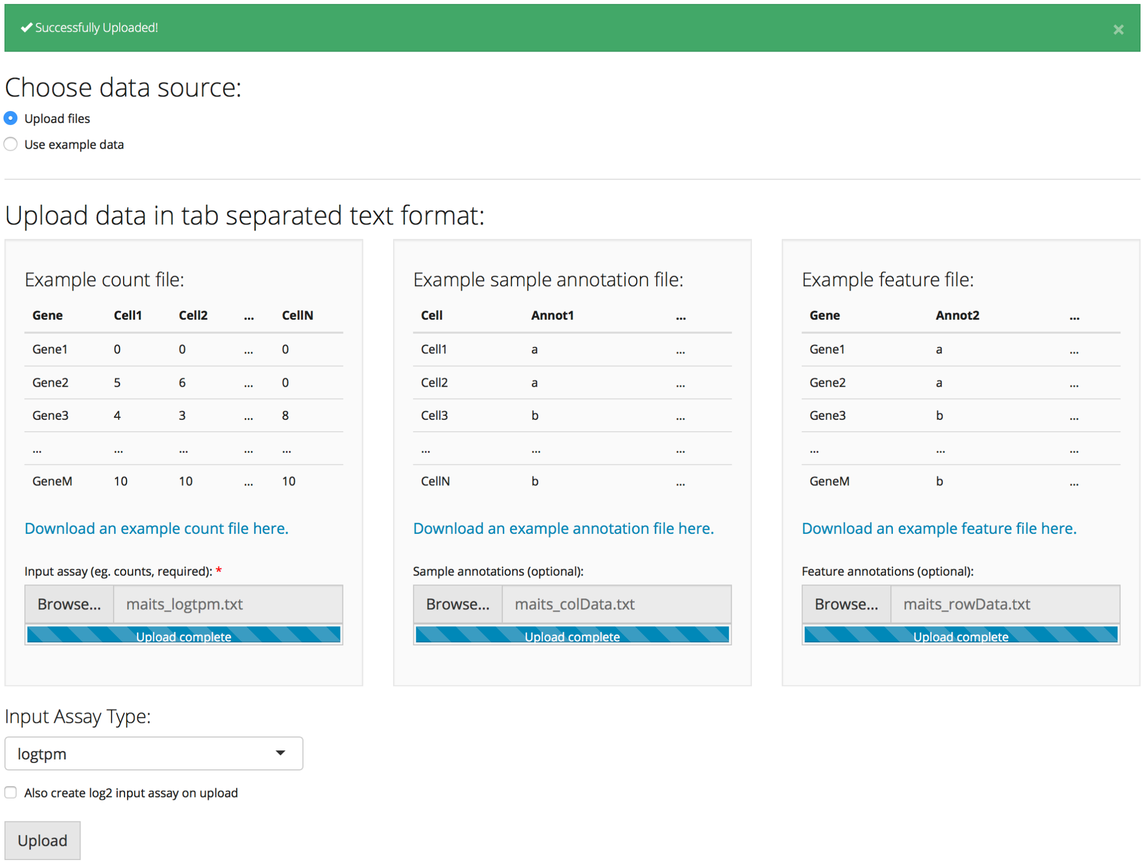
Results

The SCTK allows users to analyze data interactively through the Shiny web interface, or perform command line analysis and visualize the results when the analysis is complete. Interactive analysis works best for smaller studies of several hundred cells, which typically come from plate-based technologies such as SMART-Seq or CEL-Seq where cells are physically sorted into 96-well plates (Picelli et al., 2013, Hashimshony et al., 2016). For larger datasets, such as those created through commercially available tools such as the 10x Chromium Single Cell Solution and other droplet-based high throughput methods, analysis modules in the SCTK can be run on the command line, saved in the SCTKExperiment object, and loaded into the toolkit for efficient visualization (Macosko et al., 2015). To demonstrate a standard analysis workflow in the SCTK, two example datasets will be used. Equivalent analysis will be shown through the interactive modules and through the functions available on the R console.

Data Upload

To demonstrate how interactive analysis can be performed in the SCTK, we will begin using the MAIT cell example. The MAIT cell example should separate by experimental condition (cytokine stimulated vs unstimulated) and genes identified through differential expression and pathways identified through pathway activity analysis should be associated with cytokine stimulation.

To upload data into the toolkit for interactive analysis, data was extracted from the MAST package and the TPM matrix, sample annotations, and gene annotations were saved as tab separated text files. After starting the SCTK, the data can be uploaded on the "Upload" tab by selecting the text files and clicking upload (Figure 2.1). Optionally, the user can select "Create log(counts) assay" to store both the originally uploaded counts and a log transformed matrix.



✔ Successfully Uploaded! ×

Choose data source:

● Upload files

○ Use example data

Upload data in tab separated text format:

Example count file:

Gene	Cell1	Cell2	...	CellN
Gene1	0	0	...	0
Gene2	5	6	...	0
Gene3	4	3	...	8
...	...	...	...	...
GeneM	10	10	...	10

Download an example count file here.

Input assay (eg. counts, required): *

Browse... maits_logtpm.txt

Upload complete

Example sample annotation file:

Cell	Annot1	...
Cell1	a	...
Cell2	a	...
Cell3	b	...
...	...	...
CellN	b	...

Download an example annotation file here.

Sample annotations (optional):

Browse... maits_colData.txt

Upload complete

Example feature file:

Gene	Annot2	...
Gene1	a	...
Gene2	a	...
Gene3	b	...
...	...	...
GeneM	b	...

Download an example feature file here.

Feature annotations (optional):

Browse... maits_rowData.txt

Upload complete

Input Assay Type:

logtpm ▾

□ Also create log2 input assay on upload

Upload

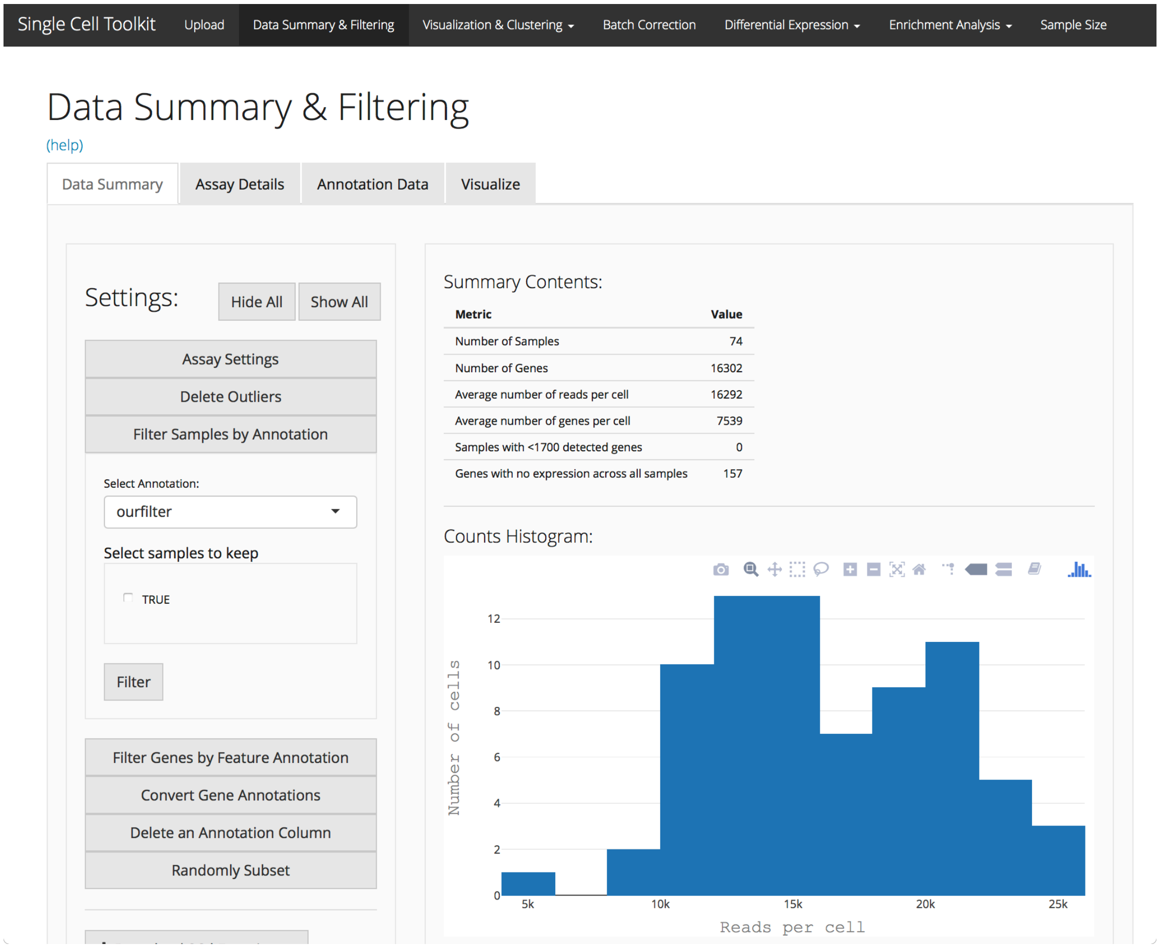
Figure 2.1. Single Cell Toolkit upload tab. Users can choose between uploading data through file upload boxes or preloaded example datasets. When the user clicks the 'Upload' button, the app creates a SCTKExperiment object to store raw data and analysis.

To perform analysis using the R functions available in the SCTK, the MAST data first must be loaded into a SCTKExperiment object. This can be accomplished with the `createSCE()` function.

```
R> library(MAST)

R> library(singleCellTK)

R> library(xtable)

R> data(maits, package="MAST")

R> maits_sce <- createSCE(assayFile = t(maits$expressionmat),

 +                        annotFile = maits$cdat,
 +                        featureFile = maits$fdat,
 +                        assayName = "logtpm",
 +                        inputDataFrames = TRUE,
 +                        createLogCounts = FALSE)
```

Data Summary and Filtering

On the second tab in the interactive toolkit, a table of summary metrics is rendered. Additionally, the user is provided with several options for filtering data and modifying the underlying SCTKExperiment object. The MAIT dataset contains an annotation column called “ourfilter.” The “Filter samples by annotation” filter was used to subset the original dataset of 96 cells to remove all cells that do not pass the filter, leaving 74 cells (Figure 2.2). This filter subsets all data assays, cell annotation data, and gene annotation data present in the SCTKExperiment object. The singleCellTK has the ability to convert gene ids to various formats in the “Convert Gene Annotation” section of the data summary and filtering page by selecting the organism, the source annotation type, and the annotation type to convert the gene annotations to.



Select Annotation: ourfilter

Select samples to keep

□ TRUE

Metric	Value
Number of Samples	74
Number of Genes	16302
Average number of reads per cell	16292
Average number of genes per cell	7539
Samples with <1700 detected genes	0
Genes with no expression across all samples	157

Figure 2.2. Single Cell Toolkit Data Summary and Filtering tab. In the right panel, a table of data summary metrics and a heatmap of counts per sample is displayed. The original 96 cells in the MAIT data are filtered to remove all samples that do not pass the “ourfilter” annotation column in the dataset using the “Filter samples by annotation” filter. 74 pass filter cells remain. Additional tools for data filtering are available in the left column.

In the R console, the `summarizeTable()` function produces summary metrics from a SCTKExperiment object. The user selects the assay to summarize and the table of summary metrics is produced (Table 2.2). Typically, these summary statistics would be run on a "counts" matrix, but the MAIT SCTKExperiment object only contains log(tpm) values so the average number of reads per cell is calculated from the normalized values instead of raw counts.

```
R> summarizeTable(maits_sce, useAssay = "logtpm")
```

Metric	Value

Number of Samples	96
Number of Genes	16302
Average number of reads per cell	17867
Average number of genes per cell	6833
Samples with <1700 detected genes	5
Genes with no expression across all samples	0

Table 2.2. Table of summary metrics produced by the `summarizeTable()` function. Five of the 96 cells in the MAIT dataset have fewer than 1,700 detected genes, indicating that these cells may have failed sequencing and should be removed for downstream analysis.

Sample annotation information is available in the colData data frame in the SCTKExperiment object. The "ourfilter" annotation can be used to subset the data within the SCTKExperiment object.

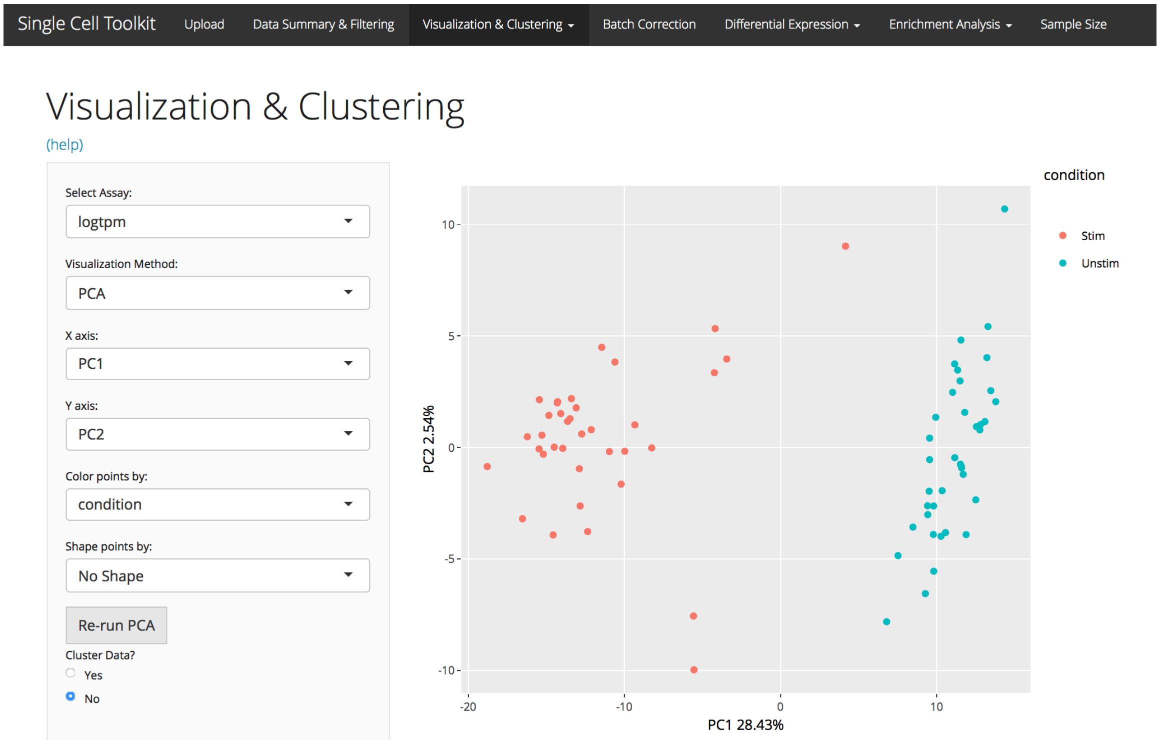
```
R> summarizeTable(maits_sce, useAssay = "logtpm")

R> colnames(colData(maits_sce))

 [1] "wellKey"          "condition"        "nGeneOn"

 [4] "libSize"          "PercentToHuman"   "MedianCVCoverage"

 [7] "PCRDuplicate"     "exonRate"         "pastFastqc"

[10] "ncells"           "ngeneson"         "cngeneson"

[13] "TRAV1"            "TRBV6"            "TRBV4"

[16] "TRBV20"           "alpha"            "beta"

[19] "ac"               "bc"               "ourfilter"

R> table(colData(maits_sce)$ourfilter)

FALSE  TRUE

   22    74

R> maits_subset <- maits_sce[, colData(maits_sce)$ourfilter]
```

To convert gene annotations in the R console, the `convertGeneIDs()` function can be used. Annotations can be converted between various formats available within the org.*.eg.db Bioconductor annotation packages which must be installed separately.

```
R> library(org.Hs.eg.db)
R> maits_entrez <- maits_subset
R> maits_subset <- convertGeneIDs(maits_subset, inSymbol = "ENTREZID",
 +                                 outSymbol = "SYMBOL",
 +                                 database = "org.Hs.eg.db")
```

Dimensionality Reduction

Next, the data is visualized in the Dimensionality Reduction and filtering tab. First, the 'logcounts' assay was selected. Since the PCA was not precalculated for this assay, PCA is performed, stored in the SCTKExperiment object, and then used for visualization in the scatter plot. The 'condition' variable in the colData annotation assay describes whether or not the cell was stimulated. There is a clear separation in the first principal component between stimulated and unstimulated cells, indicating a biological difference between the two cell conditions (Figure 2.3).



Select Assay: logtpm

Visualization Method: PCA

X axis: PC1

Y axis: PC2

Color points by: condition

Shape points by: No Shape

Cluster Data?
○ Yes
● No

Figure 2.3. Single Cell Toolkit Dimensionality Reduction and Clustering tab. Since no PCA values were present in the object, they are calculated, stored in the reducedDim slot in the object, and the first two principal components are displayed in the scatterplot. By selecting ‘condition,’ the points are colored by the condition column of the colData annotation assay in the SCTKExperiment object.

Dimensionality reduced data is stored in the reducedDims slot of the SCTKExperiment object, which can be accessed with the `reducedDims()` function. PCA and t-SNE data can be added to the object with the `getPCA()` and `getTSNE()` functions. In addition to storing the principal components in the reducedDims slot, the `getPCA()` function stores the percent variation explained by each principal component in the pcaVariances slot.

```
R> maits_subset <- getPCA(maits_subset, useAssay = "logtpm",
 +                            reducedDimName = "PCA_logtpm")
R> maits_subset <- getTSNE(maits_subset, useAssay = "logtpm",
 +                            reducedDimName = "TSNE_logtpm")
R> reducedDims(maits_subset)
```

```
 List of length 2
 names(2): PCA_logtpm TSNE_logtpm
```

PCA and t-SNE data can be visualized with the `plotPCA()` and `plotTSNE()` functions, respectively.

```
R> plotPCA(maits_subset, reducedDimName = "PCA_logtpm",
 +         colorBy = "condition")
R> plotTSNE(maits_subset, reducedDimName = "TSNE_logtpm",
 +          colorBy = "condition")
```

Similar to the PCA visualization, there is a clear separation between the stimulated and control cells in the t-SNE visualization (Figure 2.4).

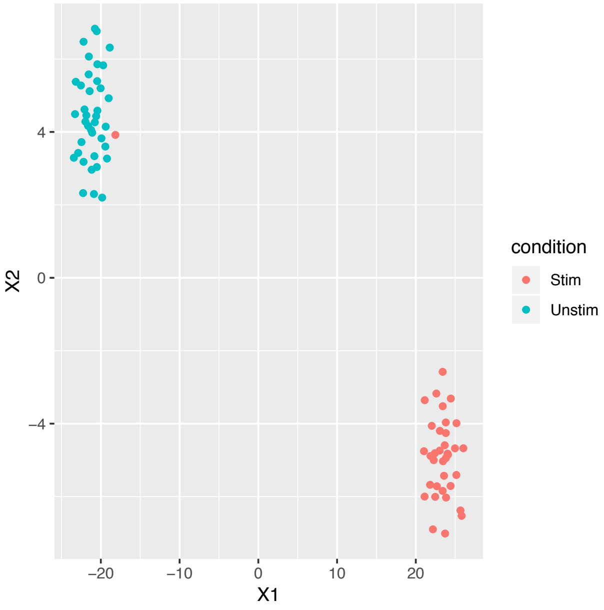
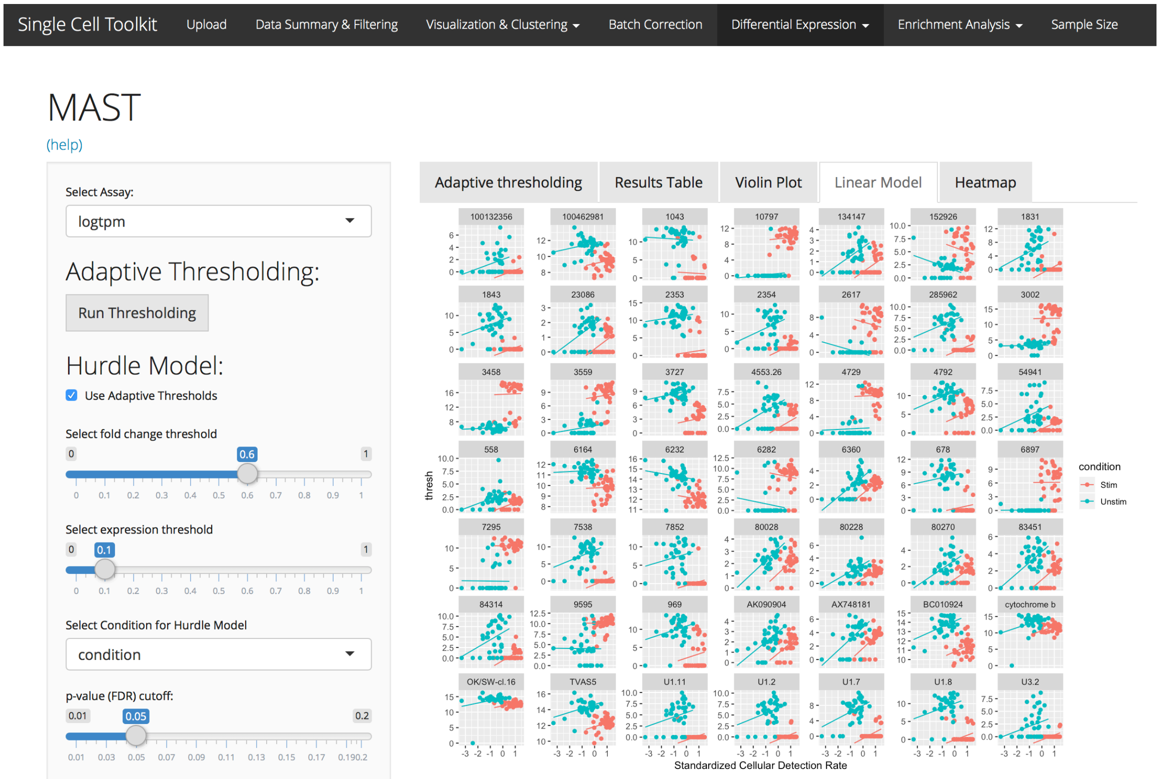


Figure 2.4. Result of `plotTSNE()` on the MAIT dataset. There is a clear separation between the stimulated and unstimulated MAIT cells. One sample marked as stimulated clusters with the other unstimulated cells. This could indicate a mislabeled sample.

Differential Expression with MAST

Differential expression analysis can identify the genes associated with the biological difference induced by cytokine stimulation that was identified during visualization with PCA. The MAST differential expression tab was used on the logcounts assay. The default options (Use adaptive thresholding, minimum fold change of 0.6, 0.1 expression threshold of 0.1, and an FDR cutoff of 0.05) were used in accordance with the MAST package example vignette (Finak et al., 2015).



Select Assay: logtpm

☑ Use Adaptive Thresholds

Select fold change threshold: 0.6

Select expression threshold: 0.1

Select Condition for Hurdle Model: condition

p-value (FDR) cutoff: 0.05

Figure 2.5. Single Cell Toolkit MAST tab. One available visualization of the MAST differential expression results is a plot of expression values vs standardized cellular detection rate for the top significantly expressed genes.

MAST analysis can be run by selecting the analysis options on the MAST page and clicking the “Run DE Using Hurdle” button. After MAST analysis completed, the 953 significant genes could be visualized as a result gene table, a set of violin plots, a heatmap, or a set of linear models of logtpm values vs cellular detection rate (Figure 2.5). Among the top differentially expressed genes was interferon gamma, a cytokine that is known to be produced in response to stimulation. The resulting significant gene list can be downloaded at the bottom of the tab.

To run MAST analysis through the R console on a SCTKExperiment object, first run adaptive thresholding on the object. After adaptive thresholding is complete, the `MAST()` function in the SCTK can be used. After MAST analysis is complete, the `MASTviolin()`, `MASTregression()`, and `plotDiffEx()` functions can be used to visualize the results (Figure 2.6).

```
R> thresholds <- thresholdGenes(maits_subset, useAssay = "logtpm")
R> mast_results <- MAST(maits_subset, condition = "condition",
 +                      useThresh = TRUE, useAssay = "logtpm")
R> MASTviolin(maits_subset, useAssay = "logtpm",
 +            fcHurdleSig = mast_results, threshP = TRUE,
 +            condition = "condition", samplesize = 16)
R> MASTregression(maits_subset, useAssay = "logtpm",
 +                fcHurdleSig = mast_results, threshP = TRUE,
 +                condition = "condition", samplesize = 16)
R> plotDiffEx(maits_subset, useAssay = "logtpm",
 +            condition = "condition",
 +            geneList = mast_results$Gene[1:100],
 +            annotationColors = "auto",
 +            displayRowLabels = FALSE, displayColumnLabels = FALSE)
```

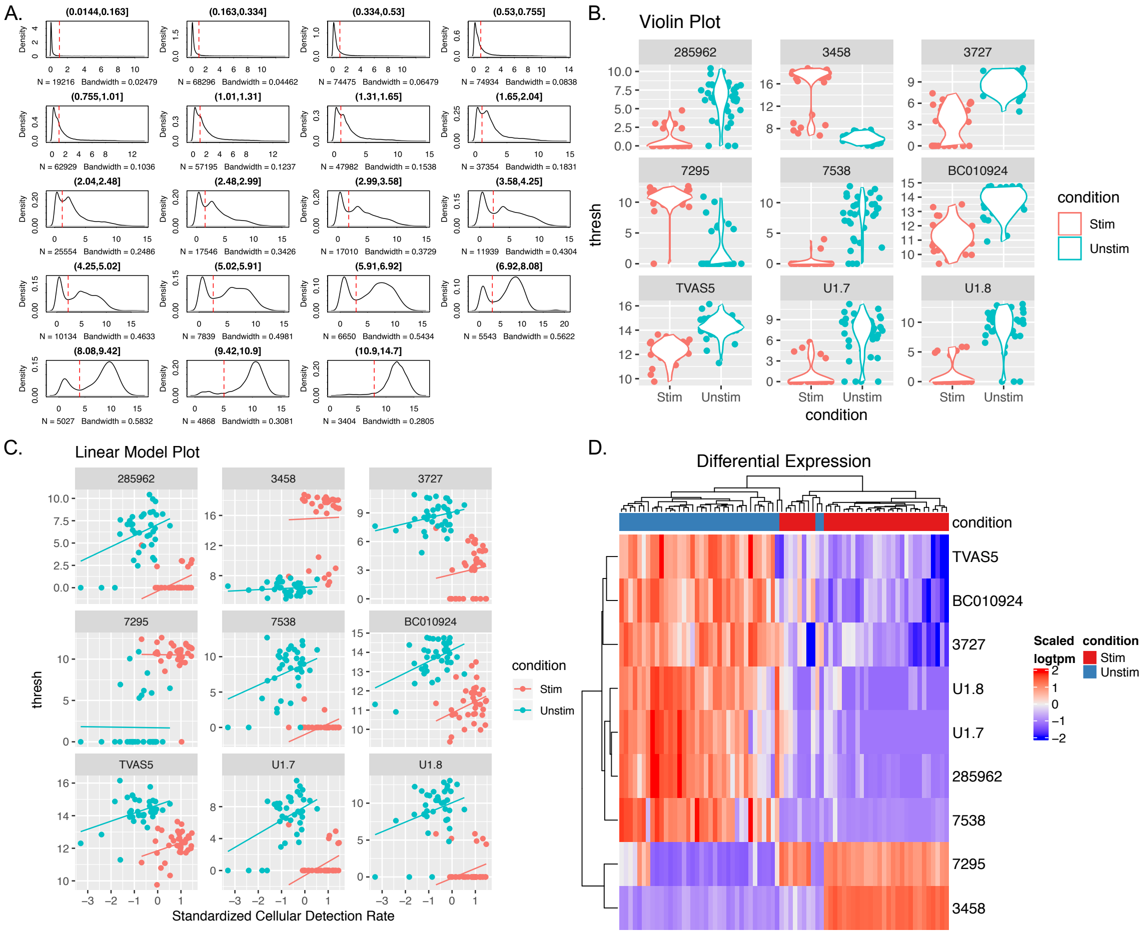
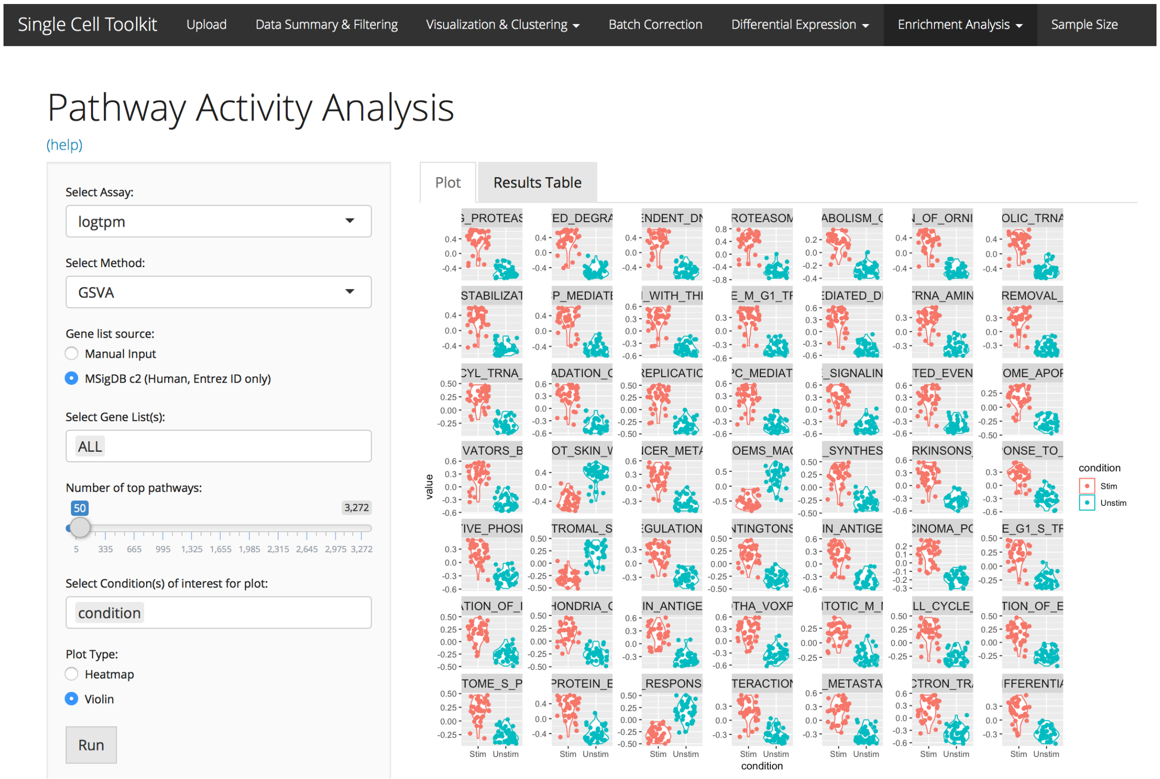


Figure 2.6. MAST result visualizations available in the SCTK. a. The `thresholdGenes()` function bins genes based on expression profile and displays a density plot for each bin. The red line indicates the cutoff for zero expression. **b.** The `MASTviolin()` function displays the top differentially expressed genes using a violin plot. **c.** The `MASTregression()` function displays the top differentially expressed genes and the CDR used in the model **d.** The `plotDiffEx()` function can be used to display a heatmap of a set of differentially expressed genes.

Pathway Activity Analysis with GSVA

To identify gene lists that show differences in pathway activity level between unstimulated and stimulated cells, the GSVA tab was used. GSVA was used to calculate pathway activity levels for all pathways in MSigDB c2. The 50 top significantly different pathway gene lists when comparing stimulated vs unstimulated cells were displayed as violin plots (Figure 2.7). Among the top pathways that showed increased activity in the stimulated cells was KEGG_PROTEASOME, indicating proteasome related genes showed increased activity in the stimulated T cells. This pathway includes interferon gamma. The pathway results can be downloaded at the bottom of the pathway activity analysis tab.



Pathway Activity Analysis

(help)

Select Assay:

logtpm ▾

Select Method:

GSVA ▾

Gene list source:

○ Manual Input

● MSigDB c2 (Human, Entrez ID only)

Select Gene List(s):

ALL

Number of top pathways:

50 — 3,272

5 335 665 995 1,325 1,655 1,985 2,315 2,645 2,975 3,272

Select Condition(s) of interest for plot:

condition

Plot Type:

○ Heatmap

● Violin

Run

Figure 2.7. Single Cell Toolkit Pathway Activity Analysis tab. Currently GSVA is supported. Users can choose to manually input a gene list or use a subset or all of the gene lists in MSigDB c2. After clicking 'Run' users can visualize a heatmap or violin plot of results if a condition of interest is given. Results can be downloaded or saved into the SCTKExperiment object.

The `gsvaSCE()` function can be used to run GSVA on an SCTKExperiment object using signatures from MSigDB. Currently, the SCTKExperiment object must use Entrez Gene IDs. Users can run GSVA using the full set of MSigDB signatures or a subset of signatures. The signatures run below are known to separate the stimulated and unstimulated cells:

```
R> gsvaRes <- gsvaSCE(maits_entrez, useAssay = "logtpm",
 +                "MSigDB c2 (Human, Entrez ID only)",
 +               c("KEGG_PROTEASOME",
 +               "REACTOME_VIF_MEDIATED_DEGRADATION_OF_APOBEC3G",
 +               "REACTOME_P53_INDEPENDENT_DNA_DAMAGE_RESPONSE",
 +               "BIOCARTA_PROTEASOME_PATHWAY",
 +               "REACTOME_METABOLISM_OF_AMINO_ACIDS",
 +               "REACTOME_REGULATION_OF_ORNITHINE_DECARBOXYLASE",
 +               "REACTOME_CYTOSOLIC_TRNA_AMINOACYLATION",
 +               "REACTOME_STABILIZATION_OF_P53",
 +               "REACTOME_SCF_BETA_TRCP_MEDIATED_DEGRADATION_OF_EMI1"),
 +               parallel.sz=1)
R> gsvaPlot(maits_subset, gsvaRes, "Violin", "condition", text_size=5)
R> gsvaPlot(maits_subset, gsvaRes, "Heatmap", "condition",
 +           show_column_names = FALSE, text_size = 5)
```

After performing GSVA, the `gsvaPlot()` function can be used to produce a set of violin plots or a heatmap of the GSVA results (Figure 2.8).

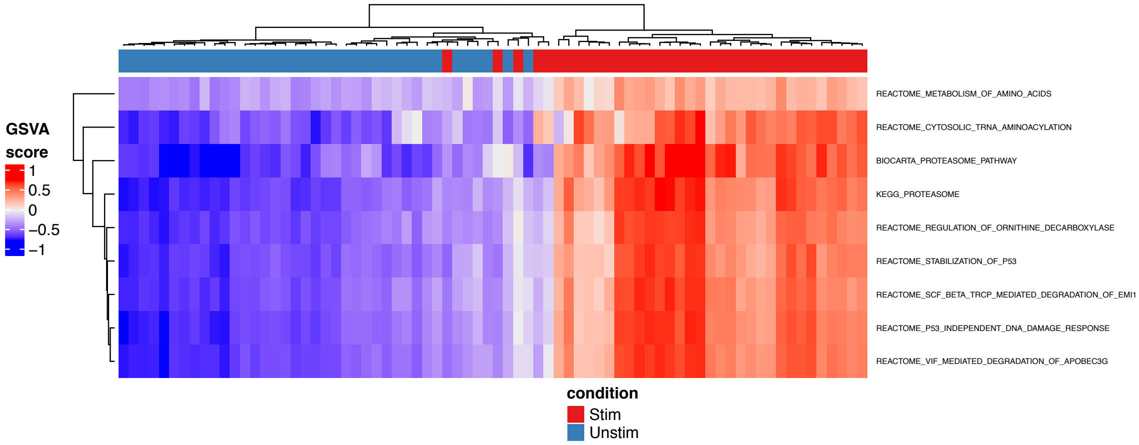


Figure 2.8. Pathway activity heatmap from the `gsvaPlot()` function. The results of GSVA pathway activity analysis can be visualized using a heatmap. If a condition of interest is chosen, a color bar is displayed on the top of the heatmap indicating the condition.

Single-Cell Batch Effects

We used the induced pluripotent stem cell line data to demonstrate the SCTK's ability to detect and correct for batch effects (Tung et al., 2017). In this dataset, three reference samples were prepared and sequenced in triplicate separately in order to introduce an experimental batch effect. Because these initial samples were identical, any difference between the replicates of the same sample represent an unwanted technical effect that could affect downstream analysis to identify biological differences between the samples.

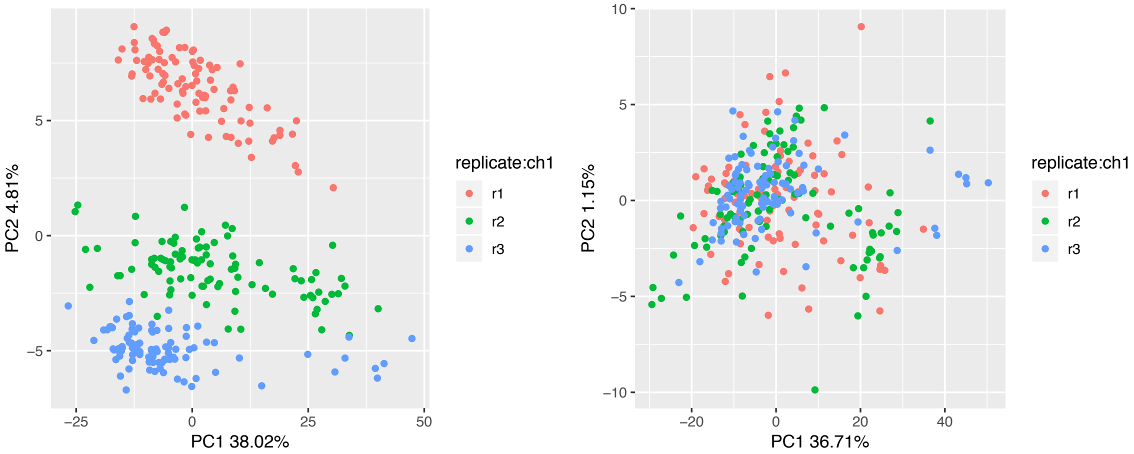


Figure 2.9. PCA before and after ComBat batch correction. The three replicates show a clear separation in the log(counts) data (left), which is corrected after running ComBat (right).

The dataset was downloaded and loaded into the SCTK. In order to reduce the effect of genes with low or no expression, cells with less than 1,700 detected genes and genes with average expression in the bottom 50 percent of the dataset were removed using the filtering tab. The three replicates from the NA19239 sample were used for downstream analysis. The resulting filtered data was visualized on the Dimensionality Reduction and Clustering tab. The batch effect resulting from the plate effect was clearly seen in the data from the log(molecules) assay (Figure 2.9, left). ComBat was run on the log(molecules) assay using default parameters (replicate as batch condition, no additional covariates, parametric combat) and saved in an assay named “combat”. The Dimensionality Reduction and clustering tab was then used to visualize a PCA of the combat assay (Figure 2.9, right). After ComBat, the plates display no signs of batch effects in the first two principal components, indicating that the technical plate artifact has been removed.

The `ComBatSCE()` function can be used to perform ComBat batch correction on a SCTKExperiment object. Batch effects can be visualized using reduced dimension data, using functions such as `plotPCA()` and `plotTSNE()`. To perform this analysis on the R console, first the data must be loaded and subset to contain the NA19239 samples only.

```
R> library(GEOquery)
R> #download data from GEO
R> GSE77288 <- getGEO('GSE77288', GSEMatrix=TRUE)
R> con <- gzcon(url(paste(
 +   "ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE77nnn",
 +   "GSE77288/suppl",
 +   "GSE77288_molecules-raw-single-per-sample.txt.gz", sep="/")))
R> txt <- readLines(con)
R> dat <- read.table(textConnection(txt), sep = "\t", header=T)
R> #extract annotation data from the GSE record
R> pdatasub <- pData(GSE77288$GSE77288_series_matrix.txt.gz)[
 +   pData(GSE77288$GSE77288_series_matrix.txt.gz)$title %in%
 +     paste(as.character(dat[,1]), as.character(dat[,2]),
 +           as.character(dat[,3]), sep="-"),]
R> rownames(pdatasub) <- pdatasub$title
R> #transform the count matrix
R> datsub <- t(dat[, 4:ncol(dat)])
R> colnames(datsub) <- paste(as.character(dat[,1]),
 +                           as.character(dat[,2]),
 +                           as.character(dat[,3]), sep="-")
R> #create SCtkExperiment object
R> GSE77288_sce <- createSCE(assayFile = datsub,
 +                           annotFile = pdatasub,
 +                           inputDataFrames = T)
R> #subset data to NA19239 only
R> GSE77288_sce <- GSE77288_sce[ ,
 +   colData(GSE77288_sce)[,"individual:ch1"] == "NA19239"]
R> #remove genes with no expression across all cellss
R> GSE77288_sce <- GSE77288_sce[
 +   rowSums(assay(GSE77288_sce, "counts")) != 0, ]
R> #log transform the count matrix
R> assay(GSE77288_sce, "logcounts") <- log2(assay(GSE77288_sce) + 1)
R> #plot before combat
R> plotPCA(GSE77288_sce, useAssay = "logcounts", runPCA = T,
 +         colorBy = 'replicate:ch1')
R> #run combat
R> assay(GSE77288_sce, "combat") <- ComBatSCE(GSE77288_sce,
 +                                                batch = "replicate:ch1")
R> #plot after combat
R> plotPCA(GSE77288_sce, useAssay = "combat", runPCA = T,
 +         colorBy = 'replicate:ch1')
```

Discussion

We have developed the Single Cell Toolkit (SCTK), a framework for analyzing and visualizing scRNA-Seq data interactively in R. With this toolkit users can process data, visualize the results, and save the data into a convenient object for further downstream analysis. Because the SCTK uses the SCTKExperiment object, the resulting data object is compatible with other tools that accept SummarizedExperiment or SingleCellExperiment objects. The toolkit supports various use cases from a user who just wants to visualize preprocessed analysis stored in a data object to a user who wants to perform a full scRNA-Seq pipeline from filtering to pathway activity analysis. The SCTK is the first fully interactive toolkit that allows a user to perform a standard scRNA-Seq workflow from uploading a count matrix to differential expression and pathway activity analysis without writing any code.

Additionally, we have used the analysis workflow available in the SCTK to identify pathway activity differences between breast cancer cells before and after drug treatment. By performing pathway activity analysis on this dataset, we found significant increases in receptor tyrosine kinase (RTK) and epithelial to mesenchymal transition (EMT) pathways, indicating the SCTK can be used to identify biologically meaningful results. (Brady et al., 2017).

We demonstrated the SCTK is an effective analysis tool by presenting a differential expression analysis and pathway activity prediction workflow for MAIT cells, and batch correction with ComBat on a set of three technical replicates of a Yoruba male reference sample. These workflows show the flexibility and interactive capability of the SCTK, which is not possible in any other currently available R package.

The SCTK is flexible and additional analysis modules will be added over time. Future improvements include additional ways to normalize input data including using ERCC spike in information, scRNA-Seq specific methods for pathway activity analysis and ComBat batch correction, additional analysis modules, and additional updates to the SCTKExperiment object to store additional results and downstream analysis and support nested study designs where cells in an experiment come from various donors or tissues.

Software Availability

The SCTK is freely available through Bioconductor (https://bioconductor.org/packages/devel/bioc/html/singleCellTK.html) and GitHub (https://github.com/compbiomed/singleCellTK). Detailed installation instructions are available on the SCTK help website (https://compbiomed.github.io/sctk_docs/).

Acknowledgments

We thank the Boston University MSSP consulting services, Zijian Han, Ziyan Li, Zichun Liu, Shiyi Yang, and Ada McFarlane for contributions to the MAST wrapper code. This research was supported by U01CA220413 from the NIH.

Chapter 3. Oncogenic growth factor receptor network signatures in TCGA and metastatic breast cancer

Adapted from the following manuscript:

Mumtahena Rahman[†], Shelley M. MacNeil[†], David F. Jenkins[†], Gajendra Shrestha, Sydney R. Wyatt, Jasmine A. McQuerry, Stephen R. Piccolo, Laura M. Heiser, Joe W. Gray, W. Evan Johnson and Andrea H. Bild. 2017. "Activity of distinct growth factor receptor network components in breast tumors uncovers two biologically relevant subtypes." *Genome Medicine* 9 (1): 40.

[†]Contributed equally

Introduction

Breast cancer remains one of the leading causes of cancer-related death in women (DeSantis et al., 2014). It is well established that growth factor receptors and their downstream signaling pathways, contribute to breast cancer proliferation, survival, and metastasis (Lemmon, Schlessinger, 2010, Mosesson, Yarden, 2004). Molecular aberrations can occur in various growth factor receptor network (GFRN) members and have been described in breast cancer (Nahta, Hortobágyi & Esteva, 2003, Hynes, 2000, Masuda et al., 2012). These findings have paved the way for GFRN-targeted treatments which are currently approved for use and being evaluated in various stages of clinical development and in clinical trials (De Abreu et al., 2014, Davis et al., 2014). Although these treatments do hold promise, relatively few data are available on the cooperativity and diversity of complicated GFRN signaling in actual breast tumors. Additionally, assessing GFRN activity in patient tumors is extremely challenging due to the lack of methods capable of measuring signaling events in tumors. Drug selection is often guided by expression of protein biomarkers, and drug resistance often develops due to compensation by interacting pathways within the GFRN (Groenendijk, Bernards, 2014, McCubrey et al., 2012). Therefore, there is a strong need to develop better methods for measuring and understanding GFRN signaling events in breast tumors in order to deliver the most effective treatment regimens and combat drug resistance (Lemmon, Schlessinger, 2010, Groenendijk, Bernards, 2014, Perona, 2006).

Growth factor receptors, such as epidermal growth factor receptor 1 (EGFR), human epidermal growth factor receptor 2 (HER2), and insulin-like growth factor 1 receptor (IGF1R), are key regulatory nodes of the GFRN and are often aberrantly activated across breast cancer subtypes (Masuda et al., 2012, Iqbal, Iqbal, 2014, Farabaugh, Boone & Lee, 2015). Approximately 15–30% of breast cancer patients are diagnosed with HER2-positive breast cancer, which is characterized by amplification of HER2 (Iqbal, Iqbal, 2014). EGFR amplifications occur in 25% of all triple-negative breast cancer (TNBC) patients and are often associated with poor outcomes (Masuda et al., 2012, Davis et al., 2014, Perou, Charles M., 2010). High IGF1R activity occurs in up to 50% of breast tumors and is seen across all breast cancer subtypes (Farabaugh, Boone & Lee, 2015). These receptors can activate downstream oncogenic growth cascades such as the phosphoinositide 3-kinase (PI3K) and mitogen-activated protein kinase (MAPK) pathways, forming a complex, interconnected, and dynamic signaling network (Lemmon, Schlessinger, 2010, Davis et al., 2014). Activation of PI3K by growth factor receptors triggers the PI3K/AKT/mammalian target of rapamycin (mTOR) pathway, leading to cell proliferation, metabolic changes, and cell survival (Baselga, 2011, Paplomata, O'Regan, 2014, Saini et al., 2013). In the MAPK pathway, following growth factor receptor activation, RAS becomes activated followed by activation of RAF1, MEK, and ERK, leading to transcriptional changes that impact cellular proliferation, motility, and evasion of apoptosis (Masuda et al., 2012, Davis et al., 2014, Santen et al., 2002, Roberts, Der, 2007). Both the PI3K and MAPK pathways contribute to tumor progression by disrupting the balance of pro- and anti-apoptotic proteins of the BCL-2 protein family in the mitochondrial (also known as intrinsic) pathway of apoptosis (Czabotar et al., 2014, Vo, Letai, 2010). Particular GFRN members can upregulate anti-apoptotic proteins such as BCL-2, BCL-XL, and MCL-1 and downregulate pro-apoptotic proteins such as BAD, BAX, and BIM, all of which contribute to apoptosis evasion and resistance to cancer treatments in patients (Letai, 2008, Datta et al., 1997, Franke et al., 2003, Townsend et al., 1998, Carpenter, Lo, 2013, Weston et al., 2003, Ley et al., 2003, Deng et al., 2007). ERBB receptor tyrosine kinases, such as EGFR and HER2, have a great deal of overlap in the downstream pathways they activate; however, individual ERBB receptors have the capability to preferentially bind particular downstream signaling molecules (Arteaga, Engelman, 2014, Yarden, Sliwkowski, 2001). Furthermore, preclinical studies have shown that EGFR- and HER2-driven cancers show differential response to targeted therapies. EGFR mutant cancers are less responsive to single-agent PI3K/AKT inhibitors in comparison to HER2-amplified cancers and require the inhibition of both the PI3K and MEK pathways (Faber et al., 2009). These suggest that ERBB proteins can couple to distinct signaling pathways and invoke non-redundant physiological effects, which warrants for specificity for the different GFRN components. Therefore, an accurate assessment of global GFRN activity is pivotal for selecting targeted treatment strategies that consider the diversity of growth and cell survival mechanisms in breast cancer patients.

Despite advances in the cellular and molecular characterization of breast cancer, effective personalized breast cancer treatment remains elusive. Immunohistochemical and gene expression profiling-defined breast cancer molecular classification has advanced our understanding of breast cancer prognosis, treatment, and improved survival. Currently, breast cancers are stratified into different clinical subtypes in order to determine specific treatments, and several breast cancer subtyping approaches are currently available. For example, fluorescence in situ hybridization (FISH) or immunohistochemistry (IHC) techniques are often used to determine clinical subtypes based on common receptor protein alterations such as estrogen (ER), progesterone (PR), and HER2 receptor amplification (De Abreu et al., 2014, Weigel, Dowsett, 2010). Additionally, Ki-67 (proliferation marker), CK 5/6 (cytokeratin marker), EGFR, androgen receptor (AR), and p53 (apoptosis marker) are used as biomarkers to further classify breast cancer using IHC methods. Although helpful, IHC methods are often subjected to bias due to tissue handling, fixation, antibody sources, and need for physical evaluation by pathologists (Hammond et al., 2010, Wolff et al., 2013). More recently, Perou and Sørlie et al. proposed five “intrinsic subtypes” that have shown utility in guiding therapy by leveraging gene expression data, differences in clinical outcomes, and responses to neoadjuvant chemotherapy (Perou, 2010, Parker et al., 2009, Sørlie et al., 2001, De Abreu et al., 2014, Patani, Martin & Dowsett, 2013). Further, evaluation of gene expression has led to the proposition of several additional subtypes, including claudin-low, molecular apocrine, and a novel luminal-like subtype (Herschkowitz et al., 2007, Prat et al., 2010, Vera-Badillo et al., 2014, Farmer et al., 2005, Guedj et al., 2012, Dvorkin-Gheva, Hassell, 2014). While molecular subtypes continue to emerge, routine use of such subtypes in clinical settings is not sensitive and specific due to some critical limitations. For example, tumors of the same clinical or intrinsic subtype can show differences in growth, survival, and response to therapies, and clinical and intrinsic subtypes are sometimes discrepant (Marusyk, Polyak, 2010, Huang et al., 2012). Approximately one-third of HER2+ tumors are not classified as the HER2-enriched intrinsic subtype and up to 25% of clinically characterized ER+ tumors are not classified as the luminal intrinsic subtype (Parker et al., 2009). While IHC methods are single protein based, intrinsic subtypes are fundamentally empirical and do not focus on distinct biological properties. Thus, both IHC and intrinsic subtypes fail to recapitulate the biological heterogeneity within each subtype (Cheang et al., 2015). Recent studies highlight the discordance between the IHC and intrinsic subtypes, which calls for additional work (Cheang et al., 2015, Tang, Tse, 2016). To address these challenges, pathway-level subtyping may provide complementary information for determining therapeutic targets. For example, identification of specific aberrant pathways within the triple negative and basal-like subtypes may help to explain additional heterogeneity and better target these subtypes pharmacologically (Badve et al., 2011). Here, breast cancer inter-tumor heterogeneity was explored in terms of GFRN activity for its well-known role in growth, evasion of apoptosis, and drug response.

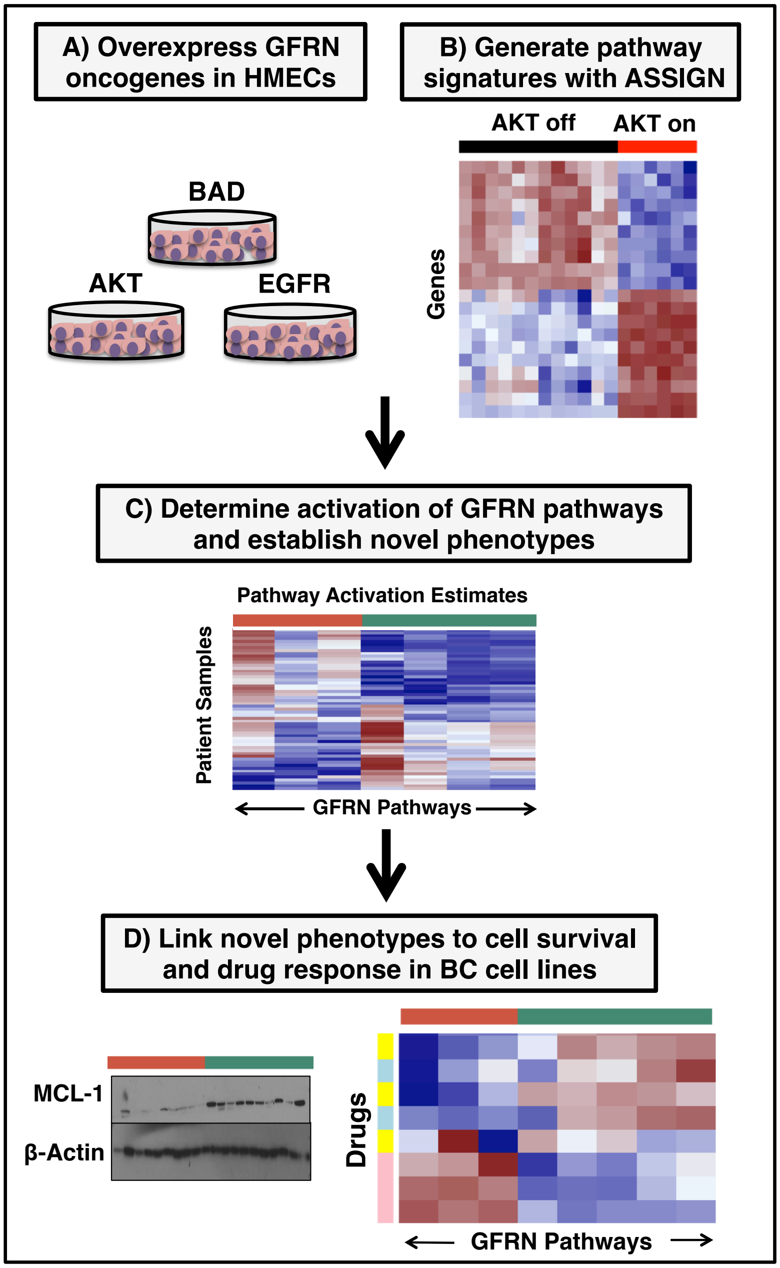

A) Overexpress GFRN oncogenes in HMECs
B) Generate pathway signatures with ASSIGN
BAD
AKT
EGFR
AKT off
AKT on
Genes
C) Determine activation of GFRN pathways and establish novel phenotypes
Pathway Activation Estimates
Patient Samples
GFRN Pathways
D) Link novel phenotypes to cell survival and drug response in BC cell lines
MCL-1
β-Actin
Drugs
GFRN Pathways

Figure 3.1. High-level overview for probing growth factor receptor networks in breast cancer. a. Overexpression of growth factor receptor network (GFRN) genes in human mammary epithelial cells (HMECs): AKT, BAD, EGFR, HER2, IGF1R, RAF1, and KRAS (G12V). **b.** Generation of RNA sequencing data from HMECs overexpressing GFRN genes and signature generation using ASSIGN. **c.** Determination of GFRN pathways activation across TCGA breast tumors and ICBP breast cancer cell lines and identification of novel phenotypes based on GFRN activity. **d.** Linking novel phenotypes to survival and drug response mechanisms in biochemical and drug response assay.

While biochemical measurement of pathway activity is challenging in human tumors due to limited tissue availability and instability of specific proteins, patterns of activity across multiple genes—or gene expression signatures—can be used as surrogates for pathway activation in tumors and to model biological phenotypes (Bild et al., 2006, Watters, Roberts, 2006, Cohen et al., 2011, Soldi et al., 2013, El-Chaar et al., 2014). Pathway activation has been used to predict drug response to targeted therapies in cell lines (Cohen et al., 2011, El-Chaar et al., 2014, Gustafson et al., 2010), but to the best of our knowledge, this is the first study which measures activity of seven GFRN members concurrently at the pathway level in patient samples. In this study, 1,119 breast tumors were profiled for GFRN activity across The Cancer Genome Atlas (TCGA) and across 55 breast cancer cell lines from the Integrative Cancer Biology Program (ICBP43) (Figure 3.1) (Cancer Genome, 2012, Daemen et al., 2013). Pathway activity was estimated in samples using novel GFRN gene expression signatures for the HER2, IGF1R, AKT, EGFR, KRAS (G12V mutation), RAF1, and BAD pathways. These GFRN signatures were generated by performing sequencing on RNA collected from primary human mammary epithelial cells (HMECs) overexpressing HER2, IGF1R, AKT1, EGFR, KRAS (G12V), RAF1, or BAD for 18–36 h. These signatures capture early transcriptional events, which occur shortly after oncogene activation, and represent the transcriptional profile of pathway activation, and not of a transformed cell.

Using the pathway analysis toolkit Adaptive Signature Selection and InteGratioN (ASSIGN), the signatures were projected onto each breast cancer data set and uncovered two discrete patterns of GFRN activity (Shen et al., 2015). One pattern was characterized by concurrent activation of the HER2, IGF1R, and AKT pathways, and another was characterized by concurrent activation of the EGFR, KRAS, RAF1, and BAD pathways. Typically, when one set of pathways was active, the other set was inactive, indicating that each sample tends to have a dominant GFRN phenotype. Pathways activation of HER2, IGF1R, and AKT was nicknamed the "survival phenotype" and activation of EGFR, KRAS, RAF1, and BAD as the "growth phenotype". These names were chosen for simplicity and based on the known role of AKT signaling in cancer cell survival and the known role of EGFR/RAS signaling in cellular growth (Zhang, Liu, 2002, McCubrey et al., 2007). Importantly, genomic pathway activity corresponded to apoptotic phenotypes. The growth phenotype showed upregulation of anti-apoptotic protein MCL-1 and downregulation of pro-apoptotic protein BIM as a mechanism of escaping apoptosis. Additional subgroups were also identified within each phenotype, including HER2 high and HER2 low activity groups within the survival phenotype and BAD high and BAD low activity groups within the growth phenotype. These discrete subgroups displayed differences in response to targeted therapies and chemotherapies. Therefore, these phenotypes can serve as surrogates for GFRN activity that capture significant variability in the gene expression data, differentiate survival mechanisms, and correlate to drug response significantly. A major component of the heterogeneity found across tumor expression data was contributed by GFRN signaling and was independent of ER, PR, and HER2 status compared to intrinsic subtypes. Additionally, a unique aspect is that GFRN activity explained the data in a biologically meaningful way. For example, while intrinsic subtyping approaches are based on empirical patterns of gene expression and do not necessarily represent a biological process, the subgrouping approach represents aberrant activity in specific GFRN pathway signaling. Therefore, pathway-based phenotypes and subgroups have the potential to complement existing methods and identify biologically and clinically relevant patterns in tumors. Taken together, pathway signatures not only aid in assessing general pathway activity patterns in a biologically relevant way, but also show promise to select better treatment targets for breast cancer patients.

Methods

Overexpression of genes of interest in human mammary epithelial cells

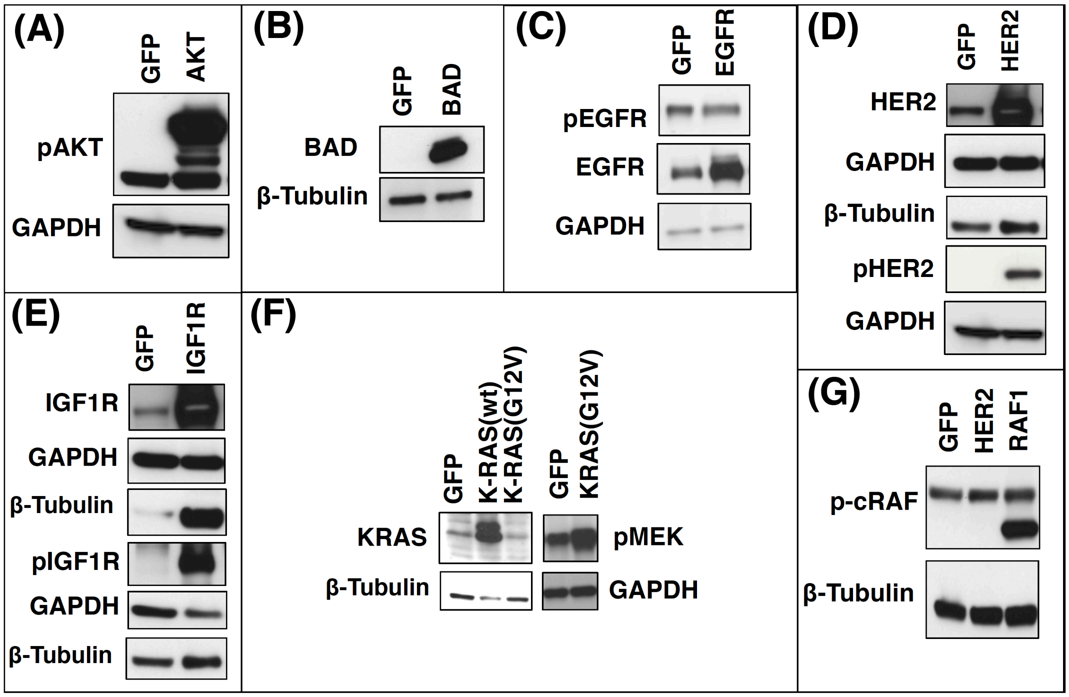


Figure 3.2. Validation of protein overexpression for each GFRN signature. Protein lysates from human primary mammary epithelial cells (HMECs) overexpressing GFRN genes were compared to GFP control protein lysates using Western blotting. **a.** HMECs overexpressing AKT1 compared to GFP (GAPDH loading control) **b.** HMECs overexpressing BAD, compared to GFP (βtubulin loading control) **c.** HMECs overexpressing EGFR and pEGFR compared to GFP (GAPDH loading control) **d.** HMECs overexpressing HER2 and pHER2 compared to GFP (GAPDH and β-tubulin loading controls) **e.** HMECs overexpressing IGF1R and pIGF1R (GAPDH and β-tubulin loading controls) **f.** HMECs overexpressing pMEK compared to GFP (β-tubulin and GAPDH loading controls) **g.** HMECs overexpressing RAF1 compared to GFP controls (β-tubulin loading controls).

In order to create gene expression signatures representative of pathway activation, GFRN oncogenes were overexpressed in HMECs. HMECs from a non-cancer-related breast reduction surgery performed at the University of Utah were isolated and cultured according to previously published protocols (Ian Freshney, Freshney, 2004). Cells were grown in serum-free mammary epithelial basal medium (MEBM) plus the addition of a “bullet kit” (Lonza) and supplemented with 5 mg/ml transferrin and 10-5 M isoproterenol at 5% CO2. Cells were brought to quiescence by growth in low serum conditions (0.25% MEBM + bullet kit, no EGF) for 36 hours. Cells were infected with recombinant adenovirus (at 500 MOI) expressing either human oncogenes AKT1, IGF1R, BAD, HER2, KRAS (G12V), and RAF1 or GFP control (Figure 3.2). Cells were incubated with virus for 18 hours except for KRAS (G12V), which was incubated for 36 hours. The adenoviral expression systems invoke transient gene expression changes, which allow us to capture the early transcriptional events of each oncogene, as opposed to the transcriptional profile of a transformed cell. Recombinant adenoviruses were amplified and concentrations were determined using previously published protocols (Luo et al., 2007). All viruses were obtained from Vector Biolabs, except RAF1 (Cell Biolabs) and EGFR (gift from Duke University).

Western blot analysis for expression of growth factor proteins in HMECs and apoptotic proteins in breast cancer cell lines

Proteins from HMECs and the following cell lines were extracted: HCC3153, HCC1395, ZR75B, HCC1569, HCC2218, SKBR3, LY2, SUM52PE, ZR7530, MDAMB361, AU565, BT474, BT483, CAMA1, HCC1419, HCC1428, MCF7, MDAMB175, T47D, ZR751, HCC1954, JIMT1, BT549, HCC1143, HCC1806, HCC1937, HCC38, HCC70, HS578T, and MDAMB213 (Appendix A). To collect protein, cells were washed with PBS, scraped on ice into PBS, pelleted by centrifugation, lysed in lysis buffer for 15 minutes (50 mM Tris (pH 8.0), 140 mM NaCl, 5 mM EDTA, 1% TritionX-100, 0.1% SDS, protease cocktail (Sigma), phosphatase inhibitors cocktails 2 and 3 (Sigma), and centrifuged at 13,000 × g for 15 minutes. Protein quantification of lysates was determined using a BCA assay (Pierce). Electrophoresis was performed on a 8–12% Tris-HCl polyacrylamide gel (BioRad) for HMEC Western blots and 18% Criterion TGX Tris/Glycine gels (BioRad) for apoptotic protein western blots. Proteins were then transferred to a PVDF membrane using the iBlot® 2 Dry Blotting System (Thermo Fisher Scientific). Membranes were blocked for 1 hour with SuperBlock™ (Thermo Fisher Scientific) and probed with the following primary antibodies: AKT (#9272), pAKT (#13038), BAD (#9292), EGFR (#4267), pEGFR (#2234), HER2 (#2165), pHER2 (#2244), IGF1R (#3027), pIGF1R (#3021), KRAS (sc-30), pMEK (#9154), p-cRAF (#9427), GAPDH (#5174), and β-tubulin (#2146). Of note, pAKT ran higher than expected due to AKT myristoylation. Breast cancer cell line lysates were probed with the following: MCL-1 (#5453), BIM (#2933), and B-actin (#3700). All antibodies were obtained from Cell Signaling Technology, besides KRAS, which was obtained from Santa Cruz.

Dose response assay

Cell lines were plated at 2000 cells per well in 384 well plates for 24 hours at 37°C. Detailed information on the cell lines and their growth conditions is provided in (Appendix A). All cell lines were obtained from American Type Culture Collection (ATCC). Drugs were diluted to six doses in media containing 5% FBS (Gibco/Life technologies) and 1% anti–anti (Gibco/Life technologies). Erlotinib, trametinib, UMI-77, obatoclax, doxorubicin, and neratinib were purchased from Selleckchem, and bafilomycin and AKT1/2 inhibitor were from Sigma-Aldrich. Drugs were dissolved in

100% DMSO and stored at -80°C. Detailed information on drug doses is provided in Table 3.1. Cell viability and growth was measured using CellTiter-Glo (Promega) 72 hours post-treatment. All treatment doses were performed in four replicates. The Drug Discovery Core Facility, a part of the Health Sciences Cores at the University of Utah, performed the dose response assay. EC50s (concentration of each drug that provides half of the maximum response) were determined and converted to drug sensitivity values defined as the negative log of the EC50s (-logEC50) (Table 3.2). EC50 values were calculated from dose response data by plotting in GraphPad Prism 4 and using the equation Y = 1/(1 + 10ˆ((logEC50 − X) × HillSlope)) with a variable slope (Y min = 0 and Y max = 1).

Drug	Company	Stock Conc.	Dose 1	Dose 2	Dose 3	Dose 4	Dose 5	Dose 6
Erlotinib	Selleck	30 mM	100 um	30 uM	10 uM	3 uM	1 uM	0.3 uM
Neratinib	Selleck	1 mM	1 uM	0.1 uM	0.05 uM	0.01 uM	0.005 uM	0.001 uM
UMI-77	Selleck	30 mM	30 uM	10 uM	3 uM	1 uM	0.3 uM	0.1 uM
Bafilomycin	Sigma-Aldrich	1 mM	5 uM	1 uM	0.5 uM	0.1 uM	0.05 uM	0.01 uM
Doxorubicin	Selleck	10 mM	3 uM	1 uM	0.3 uM	0.1 uM	0.03 uM	0.01 uM
SigmaAKT	Sigma-Alrich	10 mM	10 uM	3 uM	1 uM	0.3 uM	0.1 uM	0.03 uM
Tramatinib	Selleck	100 mM	30 uM	10 uM	3 uM	1 uM	0.3 uM	0.1 uM
Obatoclax	Selleck	5 mM	2.5 uM	1 uM	0.5 uM	0.2 uM	0.1 uM	0.05 uM

Table 3.1. Drug dose information for the drug response assay.

Cell Line	Bafilomycin	Doxorubicin	Erlotinib	Neratinib	UMI.77	Obatoclax	Sigma AKT Inhibitor	Tramatinib
HCC1143	6.940	5.993	6.113	4.461	5.390	5.951	4.365	4.181
HCC1806	8.534	6.457	6.212	5.769	4.412	6.408	4.508	4.522
HCC1937	5.042	5.751	4.553	3.901	4.866	5.910	4.654	3.102
BT549	7.487	6.460	NA	NA	5.220	5.796	4.413	NA
HCC38	7.896	6.213	3.692	NA	3.018	6.404	3.192	2.051
HS578T	6.275	6.026	3.853	1.410	5.298	6.389	4.579	NA
AU565	8.411	6.965	6.426	8.410	5.769	6.828	6.257	4.694
HCC1569	NA	5.287	4.408	NA	5.336	5.889	4.876	2.627
MDAMB361	8.508	6.234	4.862	5.816	5.345	6.527	7.133	5.098
SKBR3	7.796	5.716	4.538	6.897	NA	6.047	5.000	-10.117
BT483	7.774	6.137	3.242	NA	5.289	6.980	NA	4.200
MCF7	8.018	5.529	NA	NA	NA	6.271	4.976	NA
T47D	8.362	6.116	4.846	3.997	5.632	6.829	5.572	NA
ZR751	6.837	4.590	NA	5.677	2.412	6.093	4.687	6.770
BT474	8.425	5.741	NA	7.317	0.381	5.669	4.802	3.537
CAMA1	8.178	5.848	NA	2.451	4.740	6.507	4.985	-1.851
HCC1395	7.593	6.375	5.083	3.059	5.914	7.391	6.616	6.324

HCC1419	7.904	5.523	4.524	7.210	4.002	5.849	4.906	NA
ZR7530	7.504	5.443	5.953	7.943	5.830	6.306	4.790	4.185
HCC1954	7.527	7.308	6.422	7.594	6.127	7.041	6.313	6.838
HCC2218	10.000	NA	4.442	10.000	4.974	6.164	4.814	10.000
HCC70	7.529	6.440	5.503	4.709	5.188	6.229	6.257	10.000
JIMT1	8.569	6.261	5.206	4.822	4.981	6.145	4.201	4.852

Table 3.2. -log(EC50) drug sensitivity values from the dose response assay. All concentration are in M. NA indicates that an EC50 value could not be determined.

RNA preparation and RNA sequencing

After transfection with adenovirus and Western blot validation, cells were pelleted, washed in PBS, and stored in RNAlater (Ambion). Cells were then DNase treated, and RNA was extracted using the RNeasy kit (Qiagen). RNA replicates were generated for each overexpressed gene: six each for AKT, BAD, IGF1R, and RAF1; five for HER2; and 12 for GFP control (Gene Expression Omnibus (GEO) accession GSE83083). Additionally, 9 replicates of each of KRAS and GFP control were generated (GEO accession GSE83083). The EGFR signature and its corresponding GFP control were previously generated with six replicates of each (GEO accession GSE59765). RNA concentration was determined with a Nanodrop (ND-1000). cDNA libraries were prepared from extracted RNA using the Illumina Stranded TruSeq protocol (Illumina). cDNA libraries were sequenced at Oregon Health and Sciences University using the Illumina HiSeq 2000 sequencing platform with six samples per lane. Single-end reads of 101 base pairs were generated.

Gene expression data processing, normalization, and datasets

The Rsubread R package (version 1.14.2) was used to align and summarize RNA-Seq reads to the UCSC hg19 reference genome and annotations (Liao, Smyth & Shi, 2014, Liao, Smyth & Shi, 2013). All RNA-Seq data in this study, including HMEC overexpression data (GSE83083, GSE59765), TCGA breast cancer data (GSE62944), and ICBP breast cancer RNA-Seq dataset (GSE48213), were processed and normalized using a pipeline that can be found at https://github.com/srp33/TCGA_RNASeq_Clinical (McCubrey et al., 2007, Johnson, Li & Rabinovic, 2007).

Generation of gene expression signatures

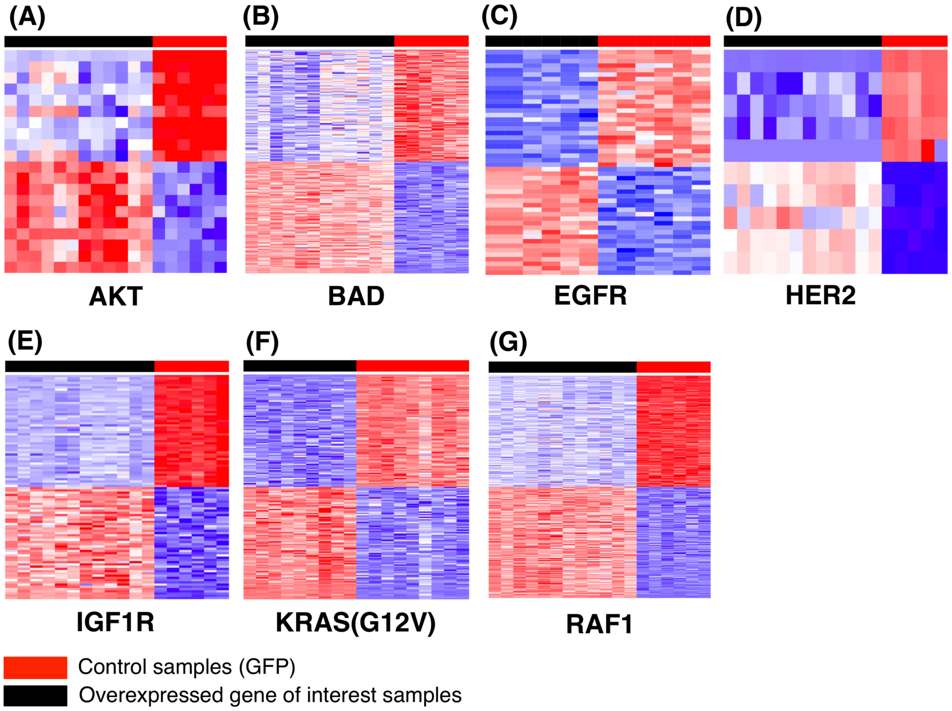


Figure 3.3. Gene expression signatures for key GFRN pathways generated by ASSIGN. a. AKT 20 gene signature, **b.** BAD 250 gene signature, **c.** EGFR 50 gene signature, **d.** HER2 10 gene signature, **e.** IGF1R 100 gene signature, **f.** KRAS (G12V) 200 gene signature, and **g.** RAF1 350 gene signature. The horizontal black bar indicates green fluorescent protein (GFP) overexpressing control samples, and the red bar indicates the overexpressed genes of interest (i.e., AKT1, BAD, EGFR, ERBB2 (HER2), IGF1R, KRAS (G12V), and RAF1, respectively) signature samples.

Adaptive Signature Selection and InteGratioN (ASSIGN; version 1.9.1), a semi-supervised pathway profiling toolkit, was used to generate gene expression signatures. A formal definition of the ASSIGN model and software implementation was reported previously (Shen et al., 2015). RNA-Seq data from HMECs overexpressing GFP control were compared to HMECs overexpressing AKT1, IGF1R, BAD, HER2, KRAS (G12V), RAF1, and EGFR. ASSIGN uses a Bayesian variable approach to select genes with the highest weights and signal strengths, indicating differential expression. These genes represent oncogenic signatures (Figure 3.3).

Gene set enrichment analysis on RNA-Seq signatures

The R package Gene Set Variation Analysis for microarray and RNA-Seq data (GSVA; version 1.22.0), a non-parametric, unsupervised method for estimating variation of gene set enrichments in gene expression data, was used to perform this gene set enrichment analysis (Hänzelmann, Castelo & Guinney, 2013). GSVA was downloaded from Bioconductor (3.4) (Huber et al., 2015). RNA-Seq data from HMECs overexpressing GFP (control), AKT1, IGF1R, BAD, HER2, KRAS (G12V), RAF1, and EGFR was used as input for the GSVA algorithm. The following gene sets were used and downloaded from the Molecular Signatures Database (Liberzon et al., 2011). 1,320 gene sets from the C2: canonical pathways collection (*c2.cp.v5.2.symbols.gmt*) and 50 gene sets from the hallmarks collection (*h.all.v5.2.symbols.gmt*). The following GSVA parameters were used: minimum gene set size = 10, maximum gene set size = 500, verbose = TRUE, rnaseq = TRUE, and method = "ssgsea". GSVA returns a matrix containing enrichment scores for each sample and gene. The R package limma (version 3.30.2) was used to perform a differential expression analysis between each overexpressed gene sample and its respective GFP control sample (Ritchie et al., 2015). The full results from the gene set enrichment analysis can be found in (Rahman et al., 2017).

Batch adjustment and estimation of pathway activity in ICBP and TCGA BRCA patient samples

HMEC oncogenic signatures (training data) were applied to 55 ICBP breast cancer cells and 1,119 TCGA breast cancer patient gene expression datasets (test data) to estimate pathway activation status. To avoid confounding batch effects within and between the training and test data, the data were adjusted for batch effects. First, in order to visualize batch effects in the data a principal component analysis (PCA) was performed on the training (HMEC overexpression RNA-Seq) data. The training data were sequenced separately in three batches, and significant batch effects were observed. Batch effects were adjusted using the ComBat function from the R package sva (version 3.21.1) (Johnson, Li & Rabinovic, 2007, Leek et al., 2012). ComBat was run using the reference-batch option, which adjusts the data to match an indicated batch. The sequencing batch containing AKT1, IGF1R, BAD, HER2, and RAF1 was selected as the reference batch. A model-matrix indicating which pathway was associated with each training replicate was also included. After the first batch adjustment, PCA was performed on the adjusted training data and the test data (ICBP breast cancer cell lines or TCGA breast tumors). Significant batch effects were identified between the training and test data and performed a second round of ComBat adjustment, using the training data as the reference batch. After the second batch adjustment, PCA was performed to confirm the resolution of the batch effect. Additionally, background baseline gene expression differences were adjusted between oncogenic signatures and test samples (ICBP cell lines and TCGA patient data) using ASSIGN’s adaptive background parameter. The variation in magnitude and direction of signature-relevant gene expression between oncogenic signature training samples and test samples was adjusted using ASSIGN's adaptive signature parameter. The model specification options for all analyses are listed in Table 3.3. Default ASSIGN settings were used for all other parameters.

Parameter	**Value**
adaptive_B	TRUE
adaptive_S	TRUE
mixture_beta	FALSE
S_zeroPrior	FALSE
sigma_sZero	0.05
sigma_sNonZero	0.5
iter	100,000
burn_in	50,000

Table 3.3. ASSIGN parameters used for all analyses. The default values were used for all other parameters.

Optimization of single-pathway estimates in ICBP cell line and TCGA BRCA patient data

To determine the optimum number of genes for each oncogenic signature, signatures with gene list lengths from 25 to 500 genes, in 25 gene increments, were generated using ASSIGN's single pathway settings. By default, ASSIGN chooses gene lists that contain an equal number of genes that have increased or decreased expression with pathway activation. ASSIGN also allows a specific gene to be anchored in the signature, making sure that the gene is always included in the signature, even if it is not chosen during gene selection or if it is removed from the signature after Monte Carlo simulation. Anchor genes were chosen based on the oncogene overexpressed in each signature. Pathway predictions generated by ASSIGN are represented as values from zero to one. Values of zero represent no pathway activity and values of one represent high pathway activity. For all the signatures that passed internal leave-one-out cross-validation, pathway estimates were included for further validation in proteomics, mutation, and gene expression. To determine optimal signature gene list lengths and evaluate the robustness of the generated signatures, pathway activation estimates from ICBP and TCGA were correlated with proteins that reflect downstream pathway activation from corresponding ICBP and TCGA RPPA data as a measurement of protein quantity (Hennessy et al., 2010, Paweletz et al., 2001).

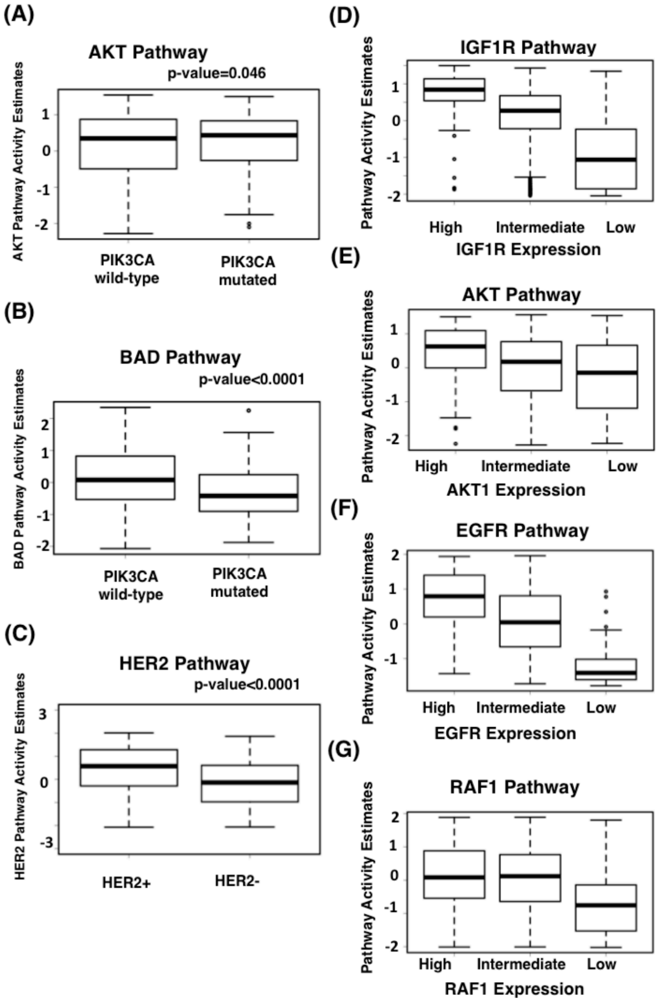


Figure 3.4. GFRN gene expression signature validations in TCGA breast cancer data. Pathway activity estimate boxplots between the **a.** AKT pathway and **b.** BAD pathway between PI3KCA mutated and PI3KCA wild-type TCGA breast cancer samples (n=787). Any mutation in PI3KCA was considered pathogenic in this mutation analysis. **c.** HER2 pathway activation estimates between HER+ and HER- patient TCGA samples (n=708). Pathway activation estimates for **d.** IGF1R, **e.** AKT, **f.** EGFR, and **g.** RAF1 between 'high', 'intermediate', and 'low' expressing samples in 1,119 BRCA TCGA samples. Samples with 90 percentile or higher expression were considered 'high', 10 percentile or lower were considered 'low', and 10 to 90 percentile were considered 'Intermediate' expressing samples for AKT1, EGFR and RAF1. For IGF1R validation, samples with 80 percentile or higher IGF1R expression were considered 'high', 20 percentile or lower was considered 'low', and 20 to 80 percentile expression were considered 'Intermediate' expressing samples.

Significant correlations were found between pathway activation estimates for all GFRN signatures and appropriate downstream pathway proteins (Farabaugh, Boone & Lee, 2015, Corbit et al., 2003, Kolch et al., 1993, Matallanas et al., 2011) (Table 3.4). Mutation-based analysis was performed using t-tests between patient groups based on mutation status in oncogenic proteins. For example, TCGA mutation data were analyzed and higher AKT activation and lower BAD activation estimates were found in patients with PI3KCA mutations (Figure 3.4a, b) and higher HER2 pathway activation estimates were found in HER2-positive tumors (Figure 3.4c). In gene expression data, higher pathway activity for AKT, EGFR, IGF1R, and RAF1 in TCGA samples classified as “high” expressing using percentiles from TCGA RNA-Seq dataset for their respective genes AKT1, EGFR, IGF1R, and RAF1 were found (Figure 3.4d–g). Samples with 90th percentile or higher expression were considered “high”, 10th percentile or lower “low”, and 10th to 90th percentile “intermediate” expressing samples for AKT1, EGFR, and RAF1. For IGF1R validation, samples with 80th percentile or higher IGF1R expression were considered “high”, 20th percentile or lower “low”, and 20th to 80th percentile “intermediate” expressing samples. Finally, pairwise Spearman correlation values and calculated p-values between pathway predictions and corresponding TCGA reverse phase protein array (RPPA) data were used to determine which gene numbers gave the best correlations. The HER2 and AKT signatures performed better with fewer genes. Therefore, 5, 10, 15, and 20 gene signatures for HER2 and AKT were generated. Significant correlations were seen between pathway estimates and RPPA protein scores. For example, AKT pathway activation estimates were significantly correlated with AKT, PDK1, and phosphorylated-PDK1 protein levels in both ICBP and TCGA (p-values <

0.0001) samples. Due to the lack of proteins available to validate the BAD signature, negative correlations between BAD pathway estimates and AKT protein based on the knowledge that activation of AKT leads to BAD inhibition were used (Datta et al., 1997). The optimized gene list was the list that gave the best average correlation in the expected direction for the RPPA data correlated with each pathway in TCGA data and was significant both in ICBP and TCGA data, with an ICBP correlation of at least 0.3 and a maximum gene list length of 300 genes. Appendix B includes a gene list of optimum gene numbers determined for each signature. Scaled ASSIGN pathway activity predictions for each of the seven optimized pathways in TCGA and ICBP are available in (Rahman et al., 2017).

Pathway	Number of Genes	Protein	ICBP		TCGA	
			Cor.	p-value	Cor.	p-value
ATK	20	Akt	0.576	2.03E-04	0.192	1.54E-07
		PDK1	0.574	2.14E-04	0.239	5.93E-11
		PDK1_pS241	0.535	6.50E-04	0.339	5.74E-21
BAD	250	Akt	-0.456	4.33E-03	-0.150	4.43E-05
		PDK1	-0.605	8.14E-05	-0.313	4.37E-18
		PDK1_pS241	-0.518	1.02E-03	-0.232	2.23E-10
EGFR	50	EGFR	0.470	0.050	0.357	2.09E-23
		EGFR_pY1068	0.397	0.028	0.129	4.50E-04
		EGFR_pY1173	NA	NA	0.155	2.44E-05
HER2	10	HER2	0.923	0.00E+00	0.376	1.61E-05
		HER2_pY1248	0.953	0.00E+00	0.356	1.37E-04
IGF1R	100	IRS1	NA	NA	0.324	2.37E-19
		IGF1R	0.086	0.608	NA	NA
		PDK1	0.569	2.45E-04	0.371	2.68E-25
		PDK1_pS241	0.509	1.26E-03	0.403	5.33E-30
KRAS (G12V)	200	EGFR	0.423	8.57E-03	0.493	4.05E-46
		EGFR_pY1068	0.296	7.17E-02	0.089	1.60E-02
		EGFR_pY1173	NA	NA	0.090	1.47E-02
		MEK1	NA	NA	0.116	1.69E-03
RAF	350	MEK1	0.285	0.084	0.245	1.72E-11
		PKC.alpha	0.467	3.46E-03	0.396	6.36E-29
		PKC.alpha_pS657	0.462	3.83E-03	0.415	0.00E+00

Table 3.4. Spearman correlations for protein correlations. Spearman correlations between pathway activation estimates and proteomics data for optimum selection in ICBP cell line and TCGA proteomics data. NA indicates that the value is not available.

Software implementation of pathway activity prediction with generated signatures

The signatures presented here have been included in the latest version of the ASSIGN package (version 1.11.3) so that pathway activity prediction can be easily performed on other datasets. Because the gene list length can affect the results of ASSIGN analysis, the signatures can be used in their original form, or the gene list lengths can be optimized based on maximizing correlations between ASSIGN activity predictions and a set of variables, such as RPPA data.

Determination of growth factor phenotypes in ICBP and TCGA

Cell lines from ICBP, patient tumors from TCGA, and breast cancer cell lines for in vitro experiments were classified as either the survival or growth phenotype by calculating the mean of scaled pathway activation values for HER, IGF1R, and AKT for the survival phenotype and the mean of scaled pathway activation values for BAD, EGFR, KRAS, and RAF1 for the growth phenotype. Each sample was classified as either survival or growth phenotype based on which phenotype had the highest mean.

Identification of additional drug response heterogeneity within growth factor phenotypes

To classify samples into subgroups within the growth factor phenotypes that corresponded to high and low HER2 activity within the survival phenotype and high and low BAD activity within the growth phenotype, the R function kmeans was used to perform k-means clustering on the scaled pathway activity data for AKT, HER2, BAD, and EGFR pathways with four means and 100 random starts. After classifying samples, t-tests were performed using the R function t.test on known HER2/AKT/PI3k/mTOR targeting drugs and EGFR/MEK targeting drugs from the drug response assay described above between the cell lines identified as AKT/HER2 high and AKT/HER2 low, and between the cell lines identified as EGFR/BAD high and EGFR/BAD low. P-values were corrected using an FDR correction and identified drugs that showed a significantly different drug response among the growth factor subgroups. When determining how growth phenotypes and ER, PR, and HER2 status performed in assessing drug response, mean drug response across all available cell lines as the cutoff were used. Cell line drug sensitivity value above this cutoff was considered as “sensitive” and otherwise “resistant”.

Statistical analyses

The prcomp function from the stats R package was used to compute the principal components in TCGA breast cancer patient RNA-Seq data. The Spearman rank-based pairwise correlation method was used for all principal component-based correlations, pathway predictions, and protein correlations. The cor.test function from the stats R package was used to calculate p-values for each correlation (Hollander, Wolfe & Chicken, 2013, Best, Roberts, 1975). Student’s t-tests were used to find the differences in principal component values based on IHC-based subtypes and mutation status within GFRN phenotypes; pathway activity based on mutation status and drug; sensitivity differences based on pathway activity, and gene expression boxplots. The heatmap.2 function from the ggplots R package and the Heatmap function from the ComplexHeatmap R package were used for generating pathway activity and pathway activity–drug response correlation heatmaps (Wickham, 2010, Gu, Eils & Schlesner, 2016). The lm function from the stats R package was used to model principal component values in TCGA using clinical subtypes, intrinsic subtypes, and GFRN subgroups to determine R^2 values. Models were compared using the anova function from the stats package to determine significance of adding additional features to the models. All analyses were conducted in R and the code is available at https://github.com/mumtahena/GFRN_signatures (The R Core Team, 2014).

Results

Two dominant phenotypes in breast cancer patients and cell lines

Gene expression signatures were developed and validated for the following GFRN pathways: AKT, BAD, EGFR, HER2, IGF1R, KRAS (G12V mutation), and RAF1. Signatures were generated in normal human mammary epithelial cells (HMECs) by expressing these genes using recombinant adenoviruses. The control samples received green fluorescent protein (GFP) adenovirus. The overall goal of this approach was to capture the downstream transcriptional events specific for each expressed GFRN gene, or the gene expression signatures, and to use these signatures to estimate pathway activity in cell lines and patient samples. To determine if adenovirus infection led to pathway activation for each overexpressed gene, protein levels of gene products and their downstream targets were measured the using western blotting (Figure 3.2). Next, RNA-Seq was performed on multiple replicates of HMECs overexpressing GFRN genes and GFP controls. These data were used to generate pathway-based gene expression signatures for each overexpressed gene using the previously published ASSIGN pathway profiling approach (Figure 3.3) (Shen et al., 2015). Briefly, ASSIGN prioritized genes that best discriminated GFP control samples from samples overexpressing GFRN genes to generate gene expression signatures. Next, ASSIGN was used to estimate the activation of each GFRN member (AKT, BAD, EGFR, HER2, IGF1R, KRAS (G12V), and RAF1) in 1,119 breast cancer patient samples from TCGA and 55 samples from the ICBP panel of breast cancer cell lines. ASSIGN was used to measure highly correlated GFRN pathway activity more accurately in patient samples with signatures generated in HMECs since ASSIGN estimates correlated pathway activities robustly by adapting pathway signatures into specific disease context. The robustness of each pathway signature was validated with (1) leave-one-out cross-validation (LOOCV), (2) relevant reverse phase protein array (RPPA) scores, (3) gene expression data for the overexpressed oncogenes, and (4) mutation data (Figure 3.4). After validating the GFRN signatures, gene set enrichment analysis was performed to identify enriched signaling patterns within each signature (Rahman et al., 2017).

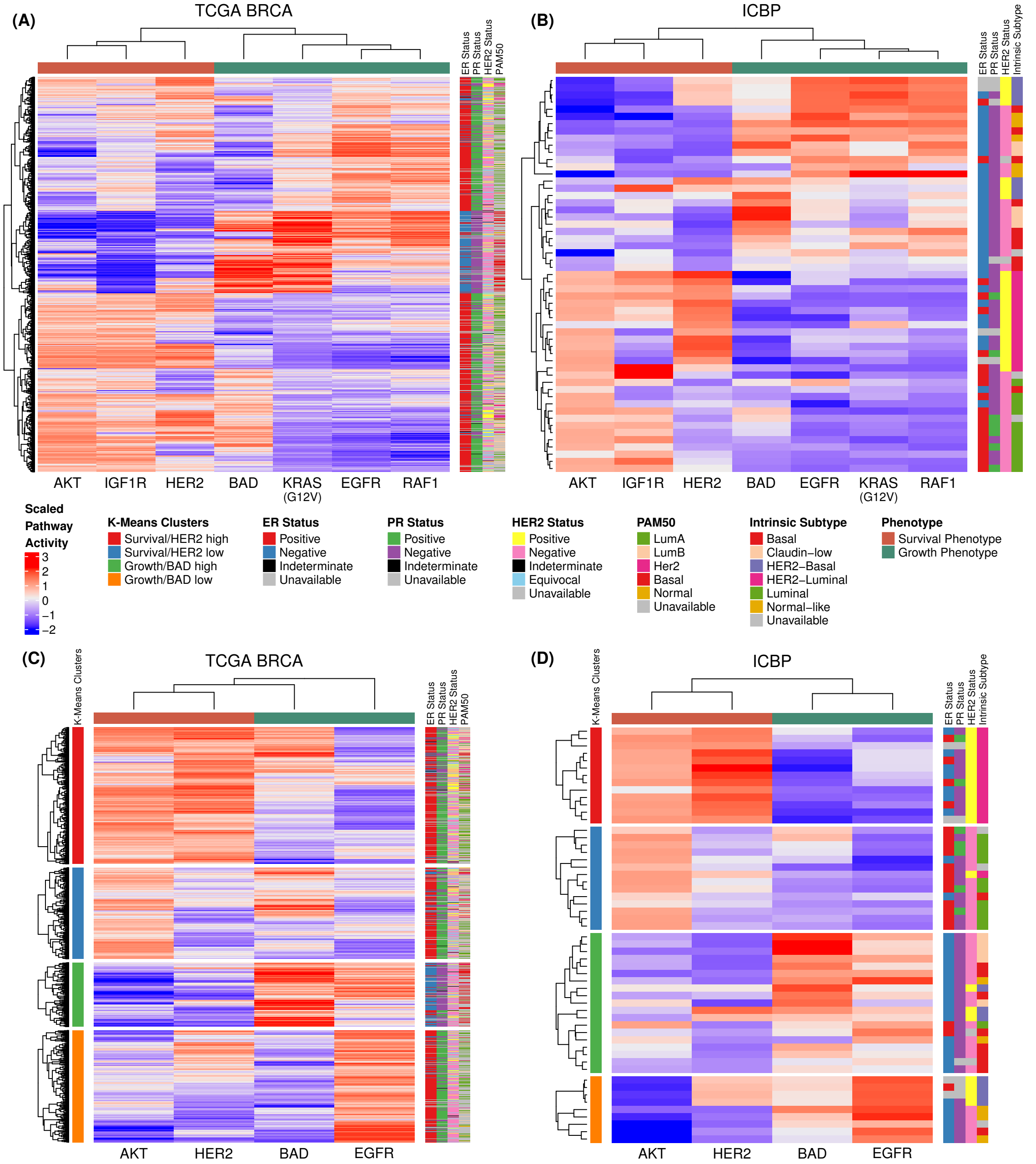


Figure 3.5. Analysis of pathway activity and intrinsic subtypes. in **a.** 1,119 TCGA breast cancer samples and **b.** 55 ICBP breast cancer cell lines. HER2, IGF1R, and AKT and BAD, EGFR, KRAS (G12V), and RAF1 pathway activities reveal two distinct clusters that were negatively associated. GFRN characterization reveals a dichotomy in TCGA breast cancer patients, high BAD/EGFR/KRAS/RAF1 (growth phenotype; column color label shown in aquamarine) and high HER2/IGF1R/AKT (survival phenotype; column color label shown in coral). Subtypes determined by immunohistochemistry and intrinsic subtyping are shown on the right side row color labels. **c.** K-means clustering of TCGA samples identifies subsets of samples within the survival phenotype that have high HER2 activation and low HER2 activation, and subsets of samples within the growth phenotype that have high BAD activation and low BAD activation (shown in the left side row color labels). **d.** These clusters are also seen in ICBP

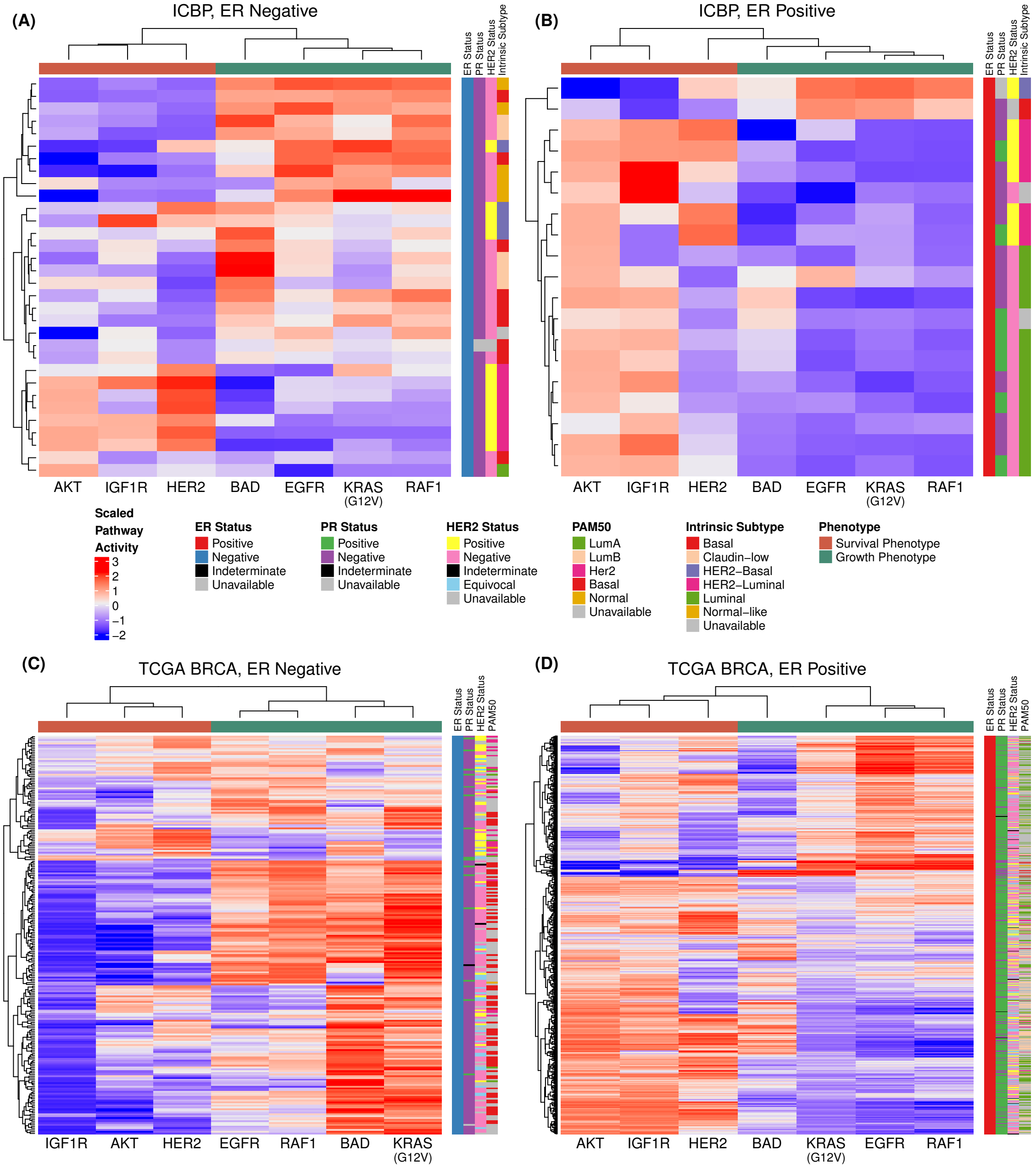


Figure 3.6. Pathway activity estimates between ER+ and ER- samples in breast cancer cell lines and patient data. a. 19 ER- breast cancer cell lines from ICBP, **b.** 32 ER+ breast cancer cell lines from ICBP. **c.** 230 ER- breast cancer patient samples from TCGA, and **d.** 785 ER+ breast cancer patient samples from TCGA. The growth phenotype is represented in aquamarine above the heat map, and the survival phenotype in coral. Subtypes determined by immunohistochemistry (ER, PR, and HER2), intrinsic subtyping, and PAM50, are label in the right side of the heatmap.

Finally, unsupervised hierarchical clustering of the pathway activity estimates for all GFRN signatures in both ICBP cell lines and TCGA patient data resulted in a dichotomous pattern (Figure 3.5a, b). The HER2, IGF1R, and AKT pathways formed a cluster, as did the remaining BAD, EGFR, KRAS, and RAF1 pathways (Figure 3.5a, b). There was some overlap between the two clusters, likely due to the known crosstalk and compensation that occurs between the PI3K and MAPK pathways (Mendoza, Er & Blenis, 2011). In general, however, when one set of pathways was high, the other set was low, which shows that samples expressed a dominant phenotype of GFRN activity. These results strongly suggest a pathway-level dichotomization of the GFRN, which is represented by two primary phenotypes: (1) activation of the HER2/IGF1R/AKT pathways or "survival phenotype"; (2) activation of the BAD/EGFR/KRAS/RAF1 pathways or "growth phenotype."

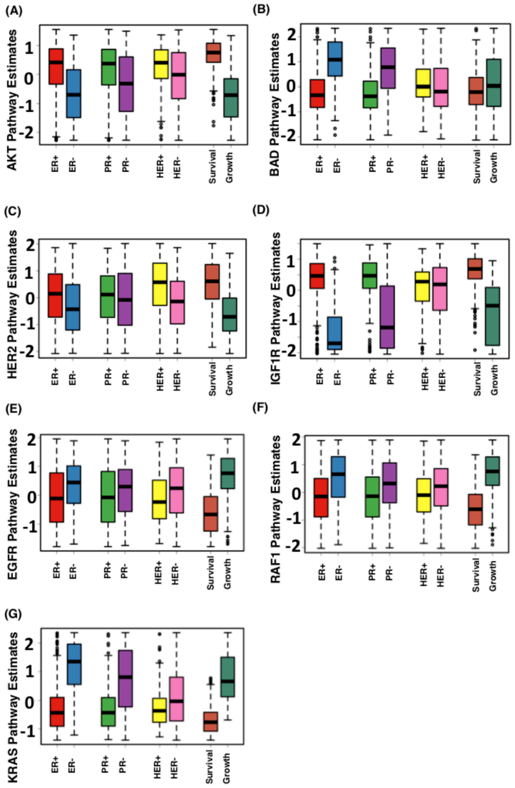


Figure 3.7. Pathway activation estimates across clinical subtypes. (IHC-based, N=1012) in TCGA breast cancer data for **a.** the AKT pathway **b.** the BAD pathway **c.** the HER2 pathway **d.** the IGF1R pathway **e.** the EGFR pathway **f.** the RAF1 pathway **g.** the KRAS pathway.

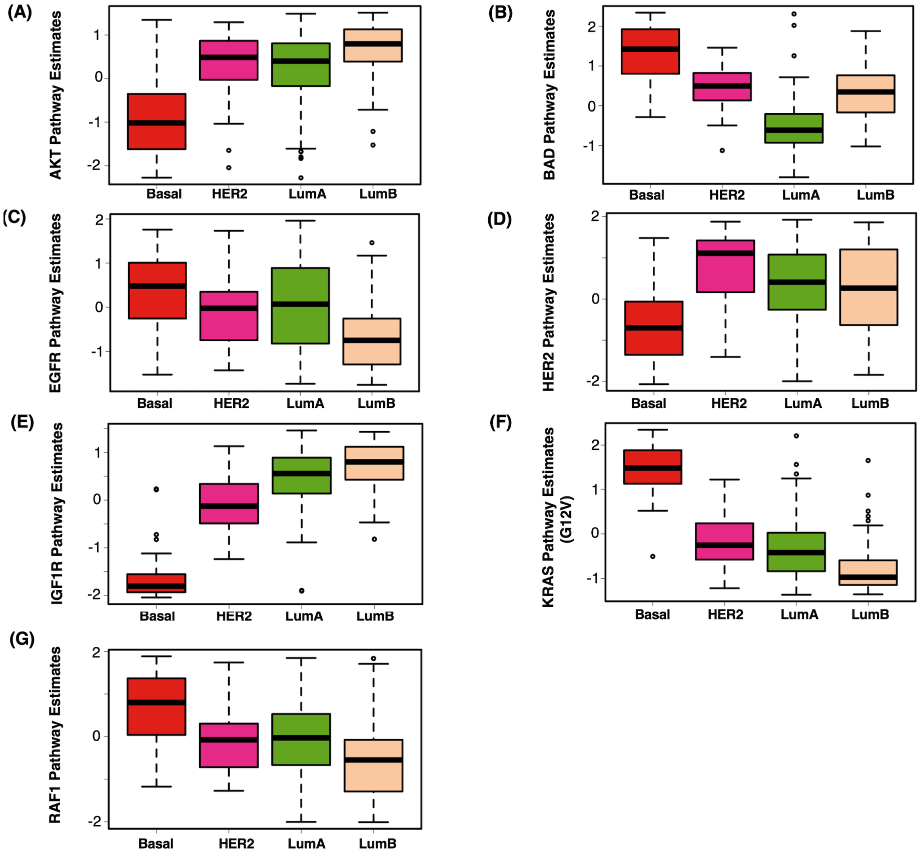


Figure 3.8. Pathway activation estimates across intrinsic subtypes. (PAM50 based, N=510) in TCGA breast cancer data for **a.** the AKT pathway **b.** the BAD pathway **c.** the EGFR pathway **d.** the HER2 pathway **e.** the IGF1R pathway **f.** the KRAS pathway **g.** the RAF1 pathway estimates.

After identifying the two main dichotomous GFRN phenotypes, these phenotypes were investigated for how they related to classic IHC-based subtypes, intrinsic subtypes, and additional heterogeneity present within each phenotype (Figure 3.5). To investigate if these phenotypes were independent of ER status, pathway activity estimates were clustered for ER+ and ER- samples separately for both ICBP and TCGA samples. The pathway activity bifurcation pattern, as represented by GFRN phenotypes, was consistent within ER+ and ER- samples, indicating GFRN phenotypes are partially independent of ER status (Figure 3.6). The variability between histological and intrinsic subtypes can also been seen in the heatmap sidebars for TCGA and ICBP data (Figure 3.5a–d), and in boxplots of pathway activity estimates across clinical and intrinsic subtypes in TCGA (Figure 3.7 and Figure 3.8). Samples classified as the survival phenotype included samples from all histological and intrinsic subtypes. Of the 596 TCGA tumors from the survival phenotype, 84.74% were ER+, 72.99% were PR+, 18.12% were HER2+, and 26.51%, 17.79%, 6.88%, and 0.34% were of luminal A, luminal B, HER2-enriched, and basal subtypes, respectively. For the growth phenotype (n = 523), even more heterogeneity in ER, PR, and HER2 status was observed (ER+, 53.54%; ER-, 37.67%; PR+, 46.85%; PR-, 43.98%; HER2+, 10.33%; HER2-, 56.41%; basal, 17.78%; Her2 enriched, 3.06%; luminal A, 13.96%; and luminal B, 4.02%). Hence, clinical and intrinsic subtypes varied in each phenotype cluster, and the GFRN phenotypes provide additional information which complements existing breast cancer clinical and intrinsic subtypes in both patient and cell line data (Perou, 2010, Sørlie et al., 2001, Sotiriou et al., 2003, Perou, C. M. et al., 2000).

HER2 activity differences were also observed within the survival phenotype, and differences in BAD activity within the growth phenotype. To further classify samples specifically on these differences, k-means clustering was performed on the AKT, BAD, EGFR, and HER2 pathway activity predictions in ICBP and TCGA. The four resulting clusters separated the survival phenotype into two subsets of samples that had either high or low HER2 activity, and the growth phenotype into two subsets of samples that had either high or low BAD activity. These patterns were observed in both TCGA and ICBP datasets (Figure 3.5c, d). Again, subtype plotted against these four subgroups as presented in the sidebars reveal there is additional heterogeneity within ER and PR status that is captured using GFRN subgroups. Of note, a survival analysis of the four subgroups in TCGA did not show significant differences in survival ($\lambda^2 = 5.5$, p-value = 0.141; Figure 3.9). This indicates that these subgroups may not relate to survival directly. Instead, these subgroups discriminate aberrant pathway activity that may help select patient subgroups likely to respond to specific drugs targeting those pathways. GFRN phenotypes complement ER status and current subtyping methods, but are more biologically focused than current intrinsic subtypes and are useful in addition to current IHC-based subtypes.

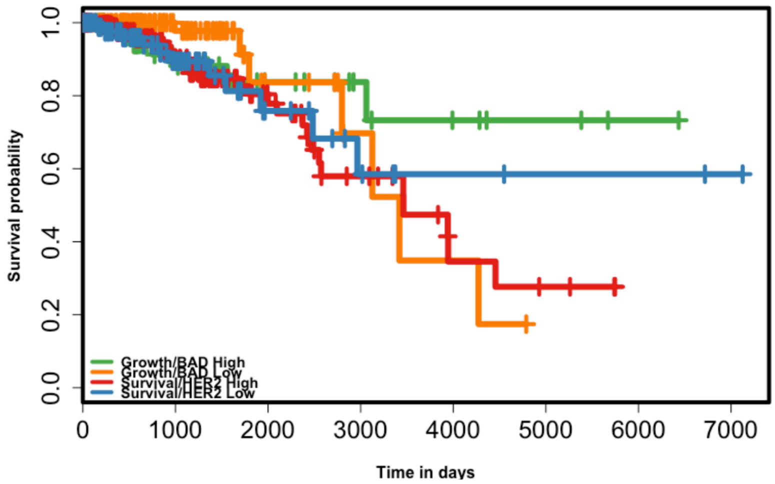


Figure 3.9. Survival analysis of the four subgroups in TCGA BRCA samples (N=1,119). (Kaplan-Meier survival analysis for the four identified subgroups using the Peto and Peto modification of Gehan-Wilcoxon test did not show significant differences across the subgroups (λ^2=5.5, p=0.141).

GFRN phenotypes and subgroups contribute to variation found in TCGA breast cancer gene expression data

In order to determine if the GFRN phenotypes and subgroups contributed to heterogeneity in the breast cancer data using an unbiased approach, an unsupervised PCA was performed on 1,119 breast cancer RNA-Seq samples from TCGA. PCA is a dimension reduction method capable of identifying uncorrelated sources of variation within a dataset as principal components (PCs) (Pearson, 1901, Hotelling, 1933). The first five PCs identified in this dataset represented the most significant amount of variability explaining 34.3% of the total variance. The remaining components, each accounting for less than 4% of the total variation, were not investigated due to their minor contribution to total variance. Of note, PC 1 was significantly associated with average gene expression of the samples (Spearman's correlation -0.786, p-value < 0.0001), potentially reflecting technical and non-disease-related sample variation (Figure 3.10). However, PC 1 was included in analyses to demonstrate its performance. To explain variability as presented by PC values, currently used histological (ER, PR, and HER2) and intrinsic subtypes were compared to GFRN-based approaches. First, each classification approach was investigated if it explained variability in each PC. When comparing PC values, significant differences were found between ER+ and ER- samples and PR+ and PR- samples for PCs 1 through 5, between HER2+ and HER2- samples for PCs 3, 4, and 5, across intrinsic subtypes for PCs 1 through 5 (ANOVA, p-value < 0.0001), between growth and survival phenotypes for PCs 2 through 5, and across four GFRN subgroups for PCs 1 through 5 (ANOVA p-value < 0.0001). These results indicate that significant variation underlying TCGA breast cancer data may be contributed from multiple sources, including GFRN phenotypes, subgroups, and histological and intrinsic subtypes.

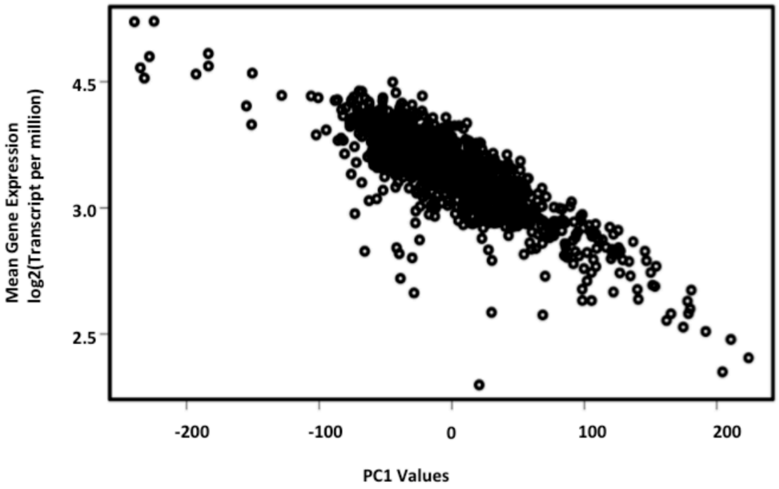


Figure 3.10. Correlation between mean gene expression values for all samples and PC 1 values from breast cancer (BRCA) TCGA RNA sequencing samples (Spearman's correlations: -0.786, p-value < 0.0001.

Second, a linear modeling approach was used to model the first five PCs with GFRN subgroups, intrinsic subtypes (PAM50), and histological (ER, PR, and HER2) subtypes. Variance explained by each model was compared in terms of R^2 values. We included 355 TCGA tumor samples for which all of these variables were available. ER ($R^2 = 0.56$) and PR ($R^2 = 0.407$) status explained a significant proportion of PC 2 but explained less than 10% of the total variability in the other PCs. HER2 status alone explained less than 4% of the variability for any of the PCs. Both GFRN subgroups, and intrinsic subtypes, explained additional variability in PCs 1–5. For all five PCs, adding the GFRN subgroups or intrinsic subtypes to clinical subtypes increased the R^2 values of the model (p-value < 0.01 for all models tested). Specifically, adding GFRN subtypes to a model of PCs explained an additional 10–35% (p-value < 0.00001) of the variation when compared to a model of ER status alone while PAM50 explained only 4–20% of the variation.

On a more granular level, GFRN subgroups explained an additional 13.5% (p-value < 0.00001) of the variability for PC 2, which was not explained by ER status alone.

For PC 3, GFRN subtypes explained an additional 35% of the variation when compared to a model of ER status alone (ER R^2, 0.052; ER+ GFRN subtype R^2, 0.398; p-value < 0.00001) and intrinsic subtypes only explained an additional 20% of the variation compared to the same model of ER status alone (ER+ intrinsic subtype R^2, 0.254; p-value < 0.00001). Overall, the models that contained GFRN subgroups explained a larger percentage of the variance of PC 1, 3, and 4, and models that contained intrinsic subgroups explained a larger percentage of the variance of PCs 2 and 5. These significant R^2 and p-values confirm the non-redundancy of GFRN subgroups in relation to commonly used clinical features in breast cancer. Additionally, GFRN subgroups explain additional variance in models of PCs 1, 3, and 4 compared to models containing intrinsic subgroups.

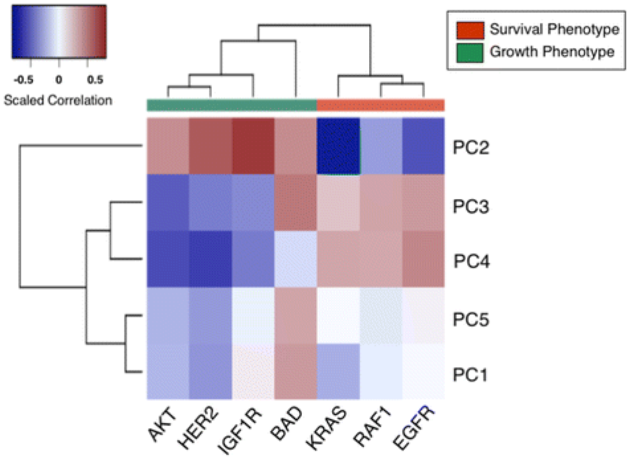


Figure 3.11. Principal component analysis across TCGA breast tumors. Correlation heatmap between principal component (PC) values from PCs 1 through 5 and ASSIGN GFRN pathway estimates from TCGA breast cancer RNA-Seq data. Red colors represent a positive correlation and blue colors represent a negative correlation.

Next, the variability contributed by GFRN subgroups was investigated in relation to biological signals, or pathway activity in this case. PC values for PCs 1 through 5 were correlated with the GFRN pathway activation estimates from TCGA (Figure 3.11). Again, a striking bifurcated pattern was found in the correlations between pathway activity and PCs in this independent variability analysis. PC 2 was positively correlated with EGFR, KRAS, RAF1, and BAD activation and negatively correlated with HER2, IGF1R, and AKT activation. Therefore, PC 2 is demonstrating characters of the growth phenotype. PCs 3 and 4 were positively correlated with HER2, IGF1R, and AKT activation and negatively correlated with EGFR, KRAS, RAF1, and BAD activation, thus representing growth phenotype characteristics (Figure 3.11). Both PC 1 and PC 5 were negatively correlated with EGFR and RAF1 activation but positively correlated with BAD activation. Since intrinsic subtypes are derived empirically without pointing to any specific biological phenomenon, a correlation to intrinsic subtypes could not be performed.

In summary, these novel GFRN subgroups explained a significant amount of variability in TCGA RNA-Seq data. The GFRN subgroups described variation beyond ER, PR, and HER2 status in all cases, and beyond intrinsic subtypes for three out of five cases. These results suggest that variability in breast cancer data can be further explained in terms of the GFRN pathway activity. Therefore, GFRN subgroups can augment current breast cancer subtyping methods by encompassing additional heterogeneity not captured by traditional approaches. This pathway-based approach may further explain specific variation in terms of pathway activity, which may point to identifying therapeutic targets.

Breast cancer growth phenotypes bifurcate in expression of mitochondrial apoptotic proteins

Next, differences between the survival and growth phenotypes were examined at the biological level, specifically in terms of mitochondrial-mediated intrinsic apoptosis mechanisms. Although cytotoxic anticancer agents induce cell death through various mechanisms, including intrinsic or extrinsic apoptosis, necrosis, autophagy, or mitotic catastrophe (Ricci, Zong, 2006, Fulda, Debatin, 2006), we focused on mitochondrial-mediated intrinsic apoptosis mediated by BCL-2 family proteins for the following reasons. First, BCL-2 family members, which regulate the commitment to mitochondrial apoptosis by balancing pro-apoptotic proteins such as BAD and BIM, and anti-apoptotic proteins such as BCL-2 or MCL-1 (Czabotar et al., 2014), have been shown to contribute to the formation, progression, and therapeutic response in breast and other cancers (Vo, Letai, 2010, Williams, Cook, 2015).

Second, particular GFRN signaling pathways, such as those found in the survival and growth phenotypes, have the potential to induce apoptosis resistance by dysregulating BCL-2 family proteins, suggesting that targeting GFRN members may lead to increased apoptosis (Datta et al., 1997, Franke et al., 2003, Townsend et al., 1998, Carpenter, Lo, 2013, Weston et al., 2003, Ley et al., 2003, Deng et al., 2007, Nalluri et al., 2015, Boucher et al., 2000, Booy, Henson & Gibson, 2011). Third, several therapeutic strategies targeting anti-apoptotic BCL-2 family members are currently under investigation; therefore, understanding which BCL-2 proteins each phenotype is expressing may provide insight into additional treatment strategies for breast cancer (Letai, 2008, Montero et al., 2015, Vogler, 2014, Hassan et al., 2014).

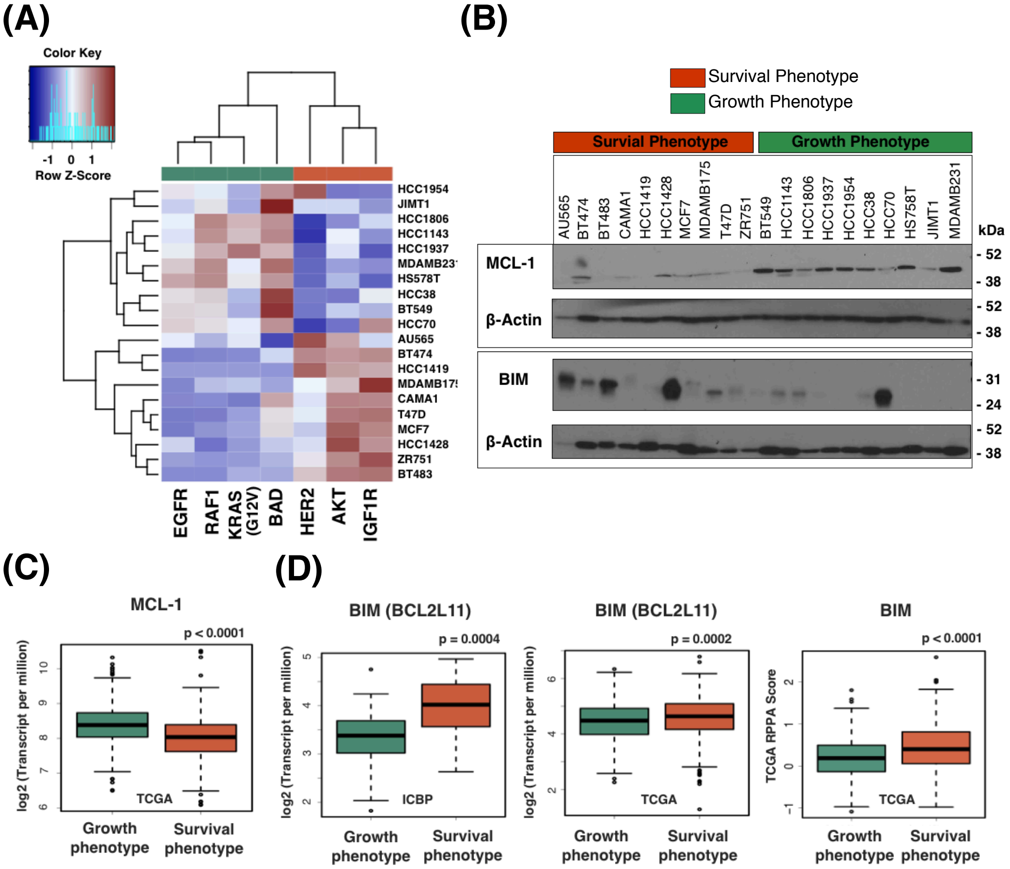


Figure 3.12. Survival and growth phenotypes differ in cell survival mechanisms. a. The heatmap represents scaled activation values across 20 breast cancer cell lines used in this analysis for each GFRN pathway. **b.** Western blot analysis for MCL-1, BIM, and B-actin control across 20 breast cancer cell lines of either the survival phenotype or growth phenotype. **c., d.** Boxplots between samples classified as the survival phenotype or growth phenotype for **c** MCL-1 gene expression (log2 (Transcript per million)) in TCGA data, d BIM gene expression (log2 (Transcript per million)) in TCGA and ICBP data, and protein expression (RPPA score) in TCGA data. Student t-tests were performed to determine significance.

Here, Western blotting was used to investigate whether protein expression of particular BCL-2 family members differed in breast cancer cell lines classified as the survival or growth phenotypes (Figure 3.12). The pro-apoptotic protein BIM and anti-apoptotic protein MCL-1 were probed across ten breast cancer cell lines of the survival phenotype (eight ER+, two ER-), and ten cell lines of the growth phenotype (ten ER-) (Appendix A). Higher levels of MCL-1 were found in cell lines of the growth phenotype, and higher levels of BIM were found in the survival phenotype (Figure 3.12b). To determine if differences in MCL-1 and BIM protein expression between the survival and growth phenotypes were due to other properties, such as ER status, a Western blot assay was performed using cell lines with additional heterogeneity in ER status. Although limited by the number of ER+ cell lines of the growth phenotype, 12 cell lines belonging to the survival phenotype (five novel ER+, three ER+ repeats from previous assay, and four novel ER-) and seven cell lines from the growth phenotype (one novel ER+, two novel ER-, and four ER- repeats) were included. The protein expression of MCL-1 and BIM were not strictly dependent on the ER status (Figure 3.13).

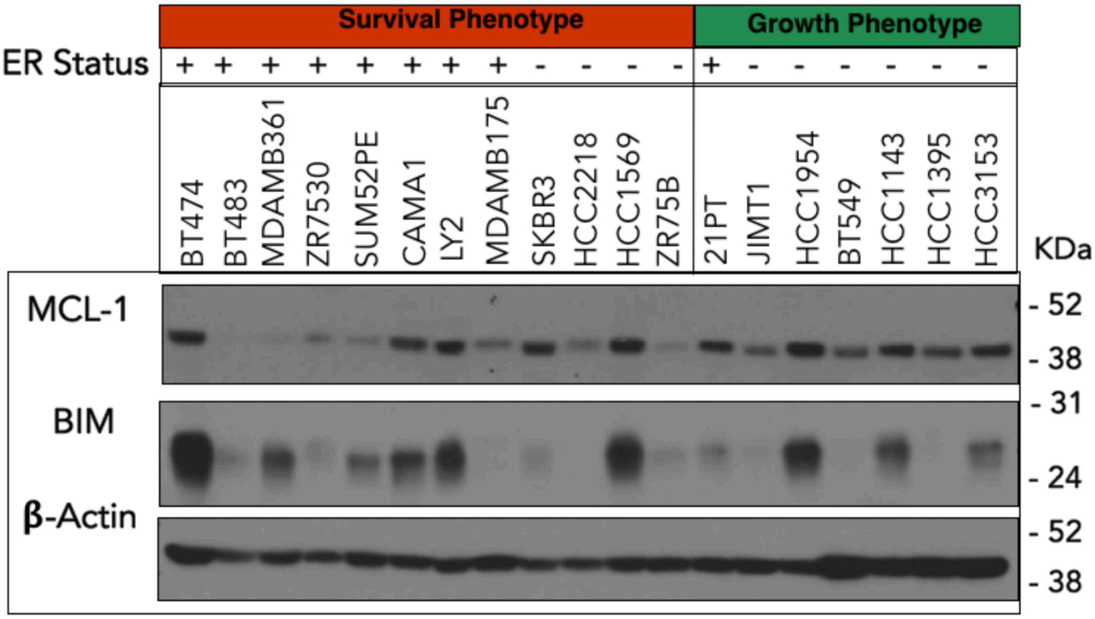


Figure 3.13. Independent western blot assay for MCL-1 and BIM proteins between breast cancer cell lines from the survival and growth phenotypes. Lysates from 12 cell lines from the survival phenotype (8 ER+ and 4 ER-) and 7 cell lines from the growth phenotype (1 ER+ and 6 ER-) were probed for anti- and pro-apoptotic proteins, MCL-1 and BIM, and compared to β-actin (loading control).

To understand if similar results could be found in patient tumors, the expression of BCL-2 family member genes was examined, and MCL-1 gene expression was found to be higher in the growth phenotype of TCGA patient tumors (n = 523) versus the survival phenotype (n = 596, p < 0.0001) (Figure 3.12c). These results were consistent with previous studies showing that EGFR signaling can upregulate gene expression of MCL-1 (Townsend et al., 1998, Nalluri et al., 2015, Boucher et al., 2000, Booy, Henson & Gibson, 2011). In addition to MCL-1 dysregulation, breast cancer cell lines of the growth phenotype expressed lower levels of the pro-apoptotic protein BIM (Figure 3.12d). In support of this assessment, lower levels of BIM (BCL2L11) gene expression were found in ICBP breast cancer cell lines ($p = 0.0004$) and TCGA tumors ($p = 0.0002$), and RPPA protein expression was lower in TCGA tumors ($p < 0.0001$) (Figure 3.12d). These results concur with literature showing that EGFR signaling through ERK activation can lead to repression of BIM (Weston et al., 2003, Ley et al., 2003, Deng et al., 2007). Also, the co-occurrence of high MCL-1 levels and low BIM levels in the growth phenotype are likely due to MCL-1's known ability to bind and neutralize BIM, which leads to prevention of apoptosis death effector activation (Vo, Letai, 2010, Wuillème-Toumi et al., 2007). In summary, these results show an interesting mitochondrial apoptotic pathway induction that is dependent on GFRN activity. Specifically, breast tumors classified as the growth phenotype may overexpress MCL-1 and inhibit BIM expression to achieve cell survival. These findings illustrate that breast cancer phenotypes, defined by activation of specific growth factor receptor pathways, express different apoptotic proteins and may resist apoptosis differently.

GFRNs predict drug response in breast cancer

Since there was a clear dichotomy in the GFRN signaling mechanisms between the survival and growth phenotypes, these phenotypes were investigated in relation to drug response in breast cancer cell lines. Pathway activation estimates were correlated with drug response data for 90 drugs from the ICBP breast cancer cell line panel.

Importantly, a consistent bifurcation pattern was observed for drug response in the cell line data that matched the observed pathway-level bifurcation. Specifically, cancer cells classified as expressing the survival phenotype were sensitive to therapies that target AKT, PI3K, HER2, and mTOR (Figure 3.14a). Additionally, these cell lines were more resistant to chemotherapies and targeted therapies that block EGFR and MEK. In contrast, cancer cells expressing the growth phenotype were sensitive to chemotherapeutics such as docetaxel, paclitaxel, and cisplatin. These cell lines were also sensitive to EGFR- and MEK-targeted therapies, but more resistant to AKT, PI3K, HER2, and mTOR inhibitors (Figure 3.14a).

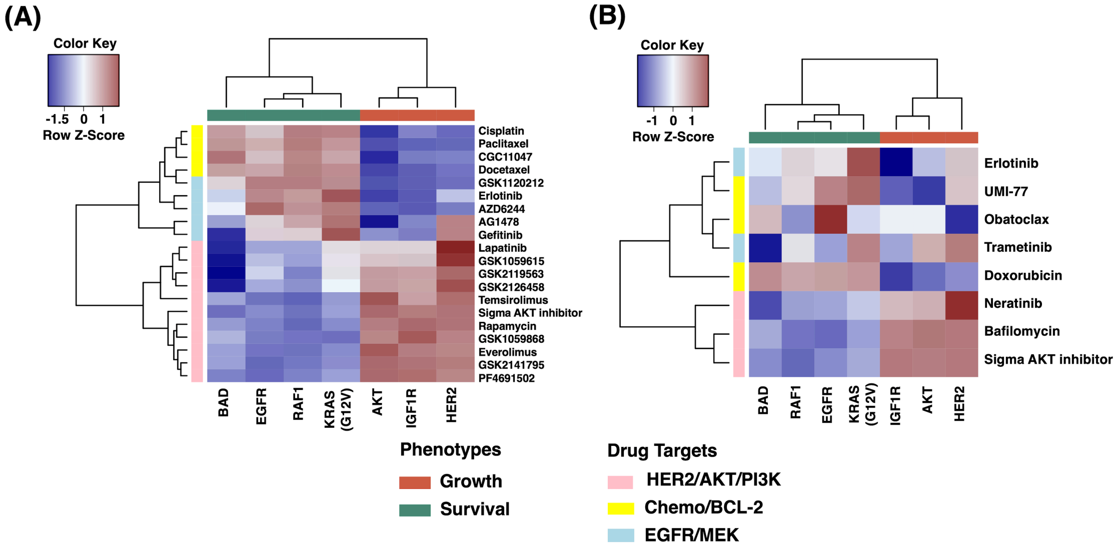


Figure 3.14. Growth factor receptor network phenotypes reflect dichotomous drug response in breast cancer cell lines. Colors correspond to scaled Spearman correlations between specific pathway activation estimates generated with ASSIGN and drug sensitivity (-logGI50) across **a.** 55 breast cancer cell lines from the ICBP panel and **b.** 23 breast cancer cell lines in an independent drug assay. Red represents positive correlation and blue represents negative correlation. Pathways cluster across the x-axis as AKT growth phenotype (coral color) and EGFR growth phenotype (green). Drug classes are represented along the y-axis: pink, HER2/AKT/PI3K/mTOR-targeted therapies; yellow, chemotherapies/BCL-2 targeting therapies; and blue, EGFR/MEK-targeted therapies.

This dichotomy in drug response of the survival and growth phenotypes was further tested in an independent drug response assay. Eight drugs on a panel of 23 breast cancer cell lines were tested, and cell viability was tested upon drug treatment by measuring ATP levels. Drugs included were obatoclax (BCL-2, BCL-XL, BCL-W, BAK inhibitor), UMI-77 (selective MCL-1 inhibitor), erlotinib (EGFR inhibitor), doxorubicin (topoisomerase II inhibitor), trametinib (MEK inhibitor), neratinib (pan-HER tyrosine kinase inhibitor), Sigma-Aldrich AKT1/2 inhibitor (dual AKT1/2 inhibitor), and bafilomycin (apoptosis inducer that inhibits PI3K/AKT signaling and autophagy inhibitor) at different doses (Table 3.1). Again, a discrete pattern was observed between the survival and growth phenotypes that translated to a bifurcated drug response pattern (Figure 3.14b). Responses to the chemotherapy (doxorubicin) and the EGFR pathway inhibitor (erlotinib) were high for the growth phenotype. In contrast, cancer cell lines classified as the survival phenotype responded well to drugs targeting components of the PI3K pathway, such as Sigma-Aldrich AKT1/2 inhibitor, neratinib, and bafilomycin.

In addition to the bifurcation of GFRN and drug response, breast tumor cells of the growth phenotype showed a higher response to the specific MCL-1 inhibitor UMI-77 (Figure 3.14b). This is consistent with the findings that samples within the growth phenotype have higher MCL-1 expression than the survival phenotype. Response to obatoclax could not be clearly distinguished based on these phenotypes, likely due to its nonspecific binding to pro-survival proteins, including BCL-2, BCL-XL, and MCL-1 (Goard, Schimmer, 2013). Overall, the GFRN phenotype-based drug response predictions were validated in this independent drug response assay. Additionally, drug sensitivity of emerging therapies such as UMI-77, neratinib, and bafilomycin showed differences between the two phenotypes, further highlighting the close relationship between GFRN signaling activity and response to therapies directed at pathways in this network.

When GFRN phenotype subgroups were considered, several drugs in the ICBP drug response assay showed significantly different drug response profiles in the subgroups found in each GFRN phenotypic arm. For example, the PI3K and mTOR inhibitor GSK1059615 and HER2/EGFR-targeting drug lapatinib were more effective in cell lines within the survival phenotype showing higher HER2 activity ($p = 0.009$ and $p < 0.000001$, respectively; Figure 3.15a, b). Additionally, ICBP cell lines expressing the growth phenotype responded better to EGFR-targeting drugs AG1478 and gefitinib in the EGFR/BAD low cluster compared to the EGFR/BAD high cluster ($p = 0.001$ and $p = 0.001$, respectively; Figure 3.15c, d).

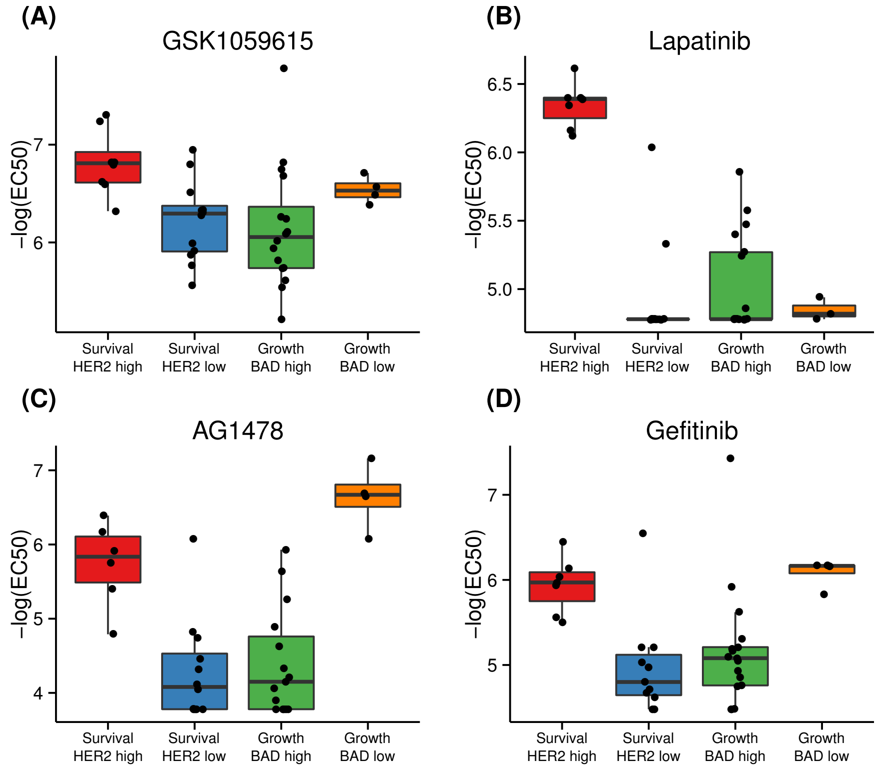


Figure 3.15. Differential drug response identified in GFRN phenotype heterogeneity. Boxplots of –log(EC50) drug response data from four drugs in the drug assay that show a differential drug response within growth factor phenotypes. **a.** GSK1059615, a PI3K and mTOR inhibitor, caused an increase in response in samples within the survival phenotype classified as having high HER2 activity. **b.** Lapatinib, a HER2 inhibitor, stimulated a stronger response in samples within the survival phenotype with high HER2 activity. **c.** AG1478 and **d.** gefitinib, EGFR inhibitors, caused an increased response in samples within the growth phenotype classified as having low BAD activity.

To determine if this bifurcation pattern was independent of clinical and intrinsic subtyping approaches, the correlations between pathway activation and drug response for ER+ and ER- and HER+ and HER- ICBP cell lines were clustered separately. Again, cell lines with high AKT/IGF1R/HER activity, i.e., the survival phenotype, were more sensitive to HER2/AKT/PI3K-targeted drugs even within ER- and HER- cell lines (Figure 3.16) In ER+ and HER+ cell lines, many PI3K/AKT/HER2-targeting drugs are more effective in the survival phenotype, as expected. However, there was additional drug response heterogeneity within ER+ samples that is associated with variations in BAD and HER2 pathway activity. These subgroups are thus helpful to further classify samples for better drug response prediction. To assess drug response across ER, PR, and HER2 status and intrinsic subtypes, it was found that out of 90 drugs studied in ICBP only 13 (14.4%), 12 (13.3%), and 19 (21.1%) showed significant differences in drug response based on ER, PR, and HER2 status, respectively, but growth/survival phenotypes were significant for 27 (49%). As further evidence, while HER2 positive status is a biomarker for effective HER2-targeted therapy, drug sensitivity does not solely depend on HER2 status. For example, while HER2 status performs much better in differentiating lapatinib's response than ER and PR status ($p < 0.0001$), some HER2- cell lines, such as HCC70 and 184A1, may respond to lapatinib. The subgroup analysis showed the survival/HER2 high subgroup to be more sensitive to lapatinib than any other subgroup (Figure 3.15b). In contrast, intrinsic subgroup analysis showed, in general, that the luminal subtype was more sensitive, but significant variability in lapatinib sensitivity exists within the luminal subtype. Other detailed examples describing comparisons between the GFRN phenotypes and other methods are included in Figure 3.15. In conclusion, the GFRN phenotypes provide additional information to current approaches; GFRN phenotypes and subgroups could be used to further stratify samples and may help select more appropriate candidates for effective drug response.

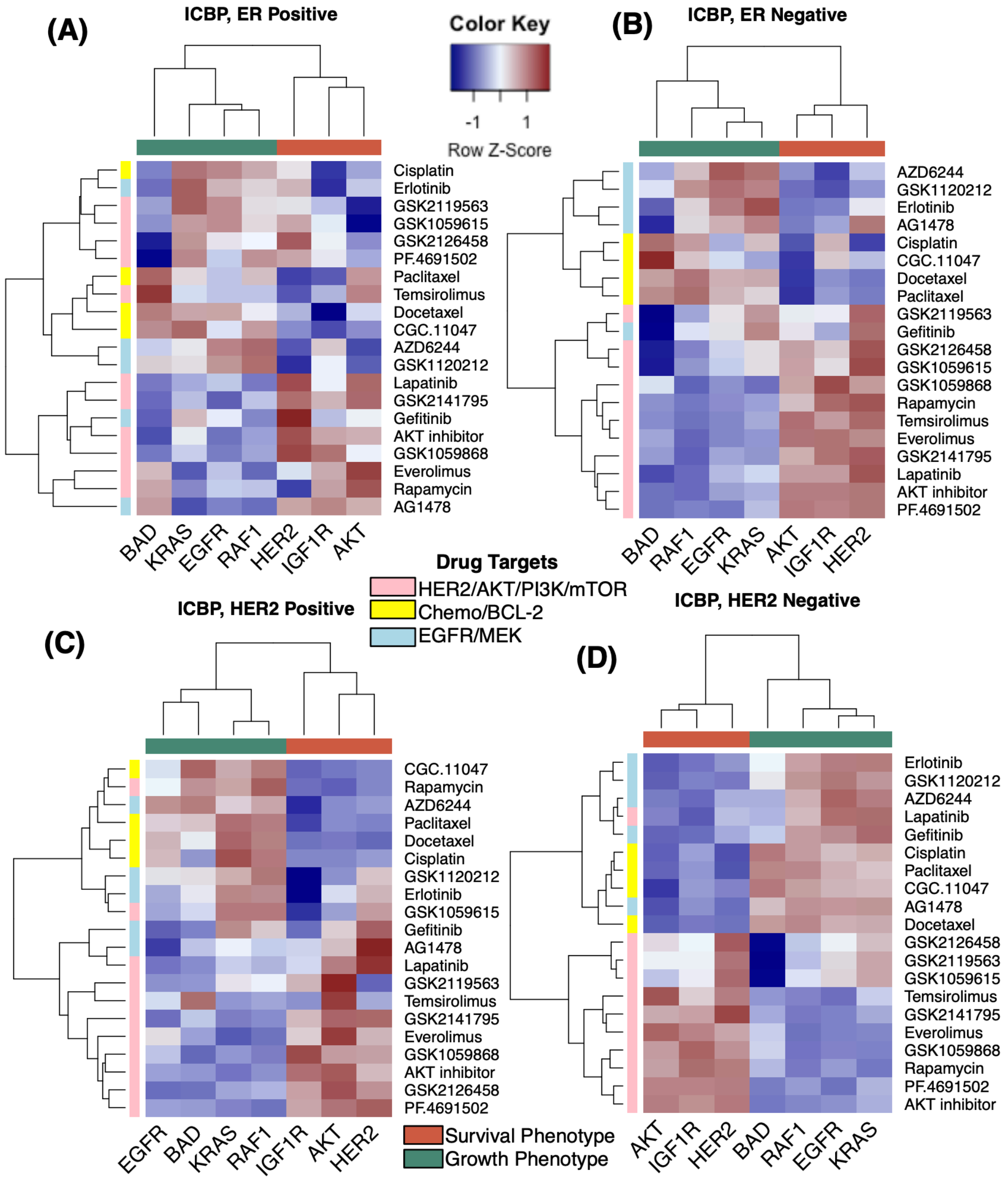


Figure 3.16. Correlations between pathway activation estimates and drug response values between ER+ and ER- and between HER+ and HER2- samples in breast cancer cell lines. Colors correspond to scaled Spearman correlations between specific pathway activation estimates generated with ASSIGN and drug sensitivity (-logGI50) across **a.** 18 ER+ breast cancer cell lines, **b.** 32 ER- breast cancer cell lines from the ICBP panel, **c.** 18 HER2+ breast cancer cell lines, and **d.** 32 HER2- breast cancer cell lines from the ICBP panel. Red represents positive correlation and blue represents negative correlation. Pathways cluster across the x-axis as (coral color) survival phenotype and (green)

growth phenotype. Drug classes are represented along the y-axis as pink (HER2/AKT/PI3K/mTOR targeted-therapies), yellow (chemotherapies/BCL-2 targeting therapies), and blue (EGFR/MEK targeted-therapies).

Discussion

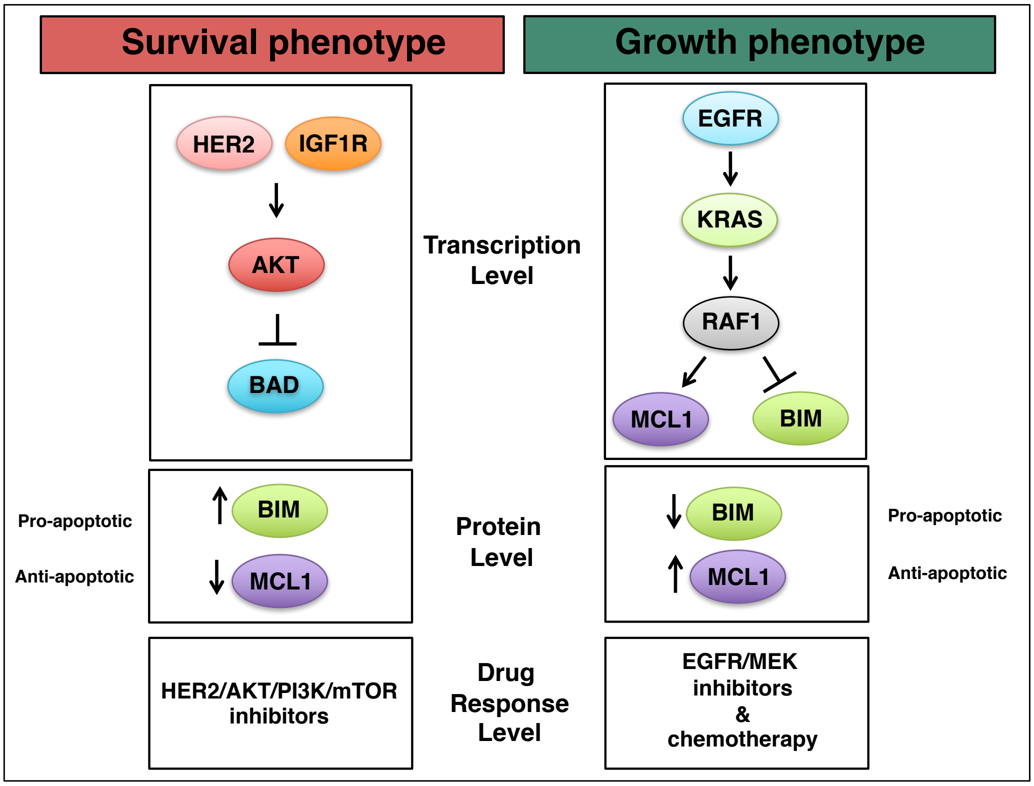


Figure 3.17. Summary of the survival and growth phenotypes in breast cancer. The survival phenotype is characterized by high HER2, IGF1R, and AKT pathway activation, high expression of pro-apoptotic BIM, low expression of anti-apoptotic MCL-1, and response to HER2, AKT, PI3K, and mTOR inhibitors. The growth phenotype is characterized by high EGFR, KRAS, and RAF1 activation, high expression of MCL-1, low expression of BIM, and response to EGFR/MEK-targeted therapies and chemotherapies.

Targeted therapies directed against the key members of the growth factor receptor network (GFRN), such as EGFR, PI3K, AKT, and mTOR inhibitors, are currently in preclinical development, clinical trials, or approved for use in breast cancer (Paplomata, O'Regan, 2014). However, predicting patients' responses to therapies is challenging due to difficulties in measuring complex signaling events in tumors. Here, this issue was addressed by investigating global GFRN activity in breast cancer using these novel signatures. Two discrete patterns of GFRN pathway activity, or phenotypes, were found

(Figure 3.17). The survival phenotype was characterized by the activation of the HER2, AKT, and IGF1R pathways, and the growth phenotype by the activation of the EGFR, KRAS, RAF1, and BAD pathways. Additional subgroups were also found within the survival and growth phenotypes, including HER2 high and low activity groups within the survival phenotype and BAD high and low activity groups within the growth phenotype. Although these discrete phenotypes were named the survival and growth phenotypes for simplicity, GFRN pathways comprising both groups can contribute to growth and survival. To the best of our knowledge, this is the first study to characterize GFRN activity using signature-based representations of activity across multiple pathways.

These discrete subgroups displayed differences in response to targeted therapies and chemotherapies in breast cancer cell lines. For example, conventional chemotherapies such as docetaxel, paclitaxel, and doxorubicin were more effective for the growth phenotype than the survival phenotype. Sensitivity to PI3K, HER2, AKT, and mTOR inhibitors and resistance to conventional chemotherapies were also found in the survival phenotype. Among the subgroups, the survival phenotype/high HER2 subgroup was hypersensitive to lapatinib, a HER2 and EGFR dual inhibitor. Similarly, the survival phenotype/high HER2 subgroup was more sensitive to GSK1059615, a PI3K/mTOR inhibitor than the survival phenotype/low HER2 subgroup. Cell lines of the growth phenotype responded better to EGFR and MEK inhibitors and to conventional chemotherapies. The growth phenotype/low BAD subtype was more sensitive to both AG1478 and gefitinib (EGFR inhibitors) than the growth phenotype/high BAD subtype. Overall, the GFRN pathway-based phenotyping contributed to information related to drug response.

Analysis of these novel phenotypes in breast cancer cell lines and tumors also revealed interesting differences in intrinsic apoptosis. For example, breast cancer cell lines and tumors of the growth phenotype had higher levels of the anti-apoptotic protein MCL-1 and lower levels of the critical pro-apoptotic protein BIM. These results are consistent with the notion that the MAPK pathway can activate MCL-1 expression and that activation of ERK1/2 and the MAPK pathway can repress BIM (Townsend et al., 1998, Weston et al., 2003, Ley et al., 2003, Deng et al., 2007). An independent drug assay also showed that the growth phenotypic cell lines responded better to a MCL-1 inhibitor (UMI-77). These results suggest that the patients with growth phenotypic expression may benefit from treatments that increase BIM, i.e., MCL-1 inhibitors, in combination with chemotherapies, EGFR inhibitors, or other inhibitors of the MAPK pathway (Akiyama, Dass & Choong, 2009, Faber et al., 2011). Therefore, targeting GFRN members may be an effective therapeutic strategy for inhibiting GFRN pathways and increasing apoptosis (Letai, 2008). These results highlight that mapping phenotypes, such as growth networks in breast tumors, can be exploited to guide the use of targeted therapies. This study was limited to how GFRN activity related to drug response and cellular intrinsic apoptosis, but it is understood that this is not the sole mechanism by which cancer cells die, and other cell death mechanisms, such as necrosis, autophagy, and mitotic catastrophe, should also be considered. In addition, as the use of cell lines is limited, a larger-scale analysis of apoptotic pathways dysregulation in patient tumor cells of all subtypes will be informative in further detailing how these pathways signal in cancer. These phenotypes may correlate with other subtyping properties, and may also be confounded by properties of intrinsic subtyping.

Importantly, these newly discovered breast cancer survival and growth phenotypes are biologically relevant and offer a direct method for probing and targeting the GFRN in breast tumors. In addition, these phenotypes complement widely used clinical and intrinsic subtypes, and stratification of cancers by these phenotypes leads to enhanced drug response predictions compared to classifying cancers by clinical subtyping approaches. This is most likely because oncogenic pathway activation was measured more comprehensively than relying on single protein measurements. In addition, this approach considers crosstalk between members of the GFRN and correlates with biological processes such as cell survival. This pathway-based approach for identifying phenotypes allows for exploration of additional heterogeneity occurring within the identified phenotypes, which can further improve the ability to stratify breast cancers by pathway activity, which then can be used to predict drug response. Although this method has added to current approaches for predicting drug response in breast cancer, most experiments were performed in breast cancer cell lines with particular classes of drugs; additional drug testing should be performed in breast cancer patient cells in order to confirm these phenotypes.

In summary, a novel genomic pathway-based approach of characterizing the interactive GFRN activation in breast cancer was used to discover two discrete GFRN phenotypes with significant differences in cell survival mechanisms and drug response in breast cancer. These phenotypes captured the distinct bifurcation pattern seen in gene expression, the GFRN pathway activity, mitochondrial apoptotic network protein expression, and drug response (Figure 3.17). While ER, PR, HER2 status and, more recently, intrinsic subtype are used to guide breast cancer treatment, these subtyping or classifying approaches may not describe signaling pathway dysregulation in tumor cells. Pathway activity data provide additional information about tumor cells that can be leveraged to predict drug response. Characterizing individual tumors into these phenotypes can help determine which patients will benefit from a treatment and select the appropriate subpopulations for clinical trials. Importantly, these seven pathways did not capture all of the heterogeneity of the samples and inclusion of other pathways may have additional benefits. Although feasible, additional investigation is needed before these phenotypes can be used in clinical trials for patient selection, including the testing of these phenotypes in patient primary tumor cells.

Conclusion

A discriminating bifurcation pattern of key GFRN pathways was identified in breast tumors that expands beyond histological and clinical subtypes. These phenotypes correlated with unique apoptotic and drug response mechanisms. The ability to measure signaling events more accurately in patient tumors advances understanding of the biological basis of cancer. These results may lead to more effective and individualized treatment selection in patients with breast cancer.

Acknowledgments

We thank Laurie Jackson for generation of gene expression data and Bai Luo for assisting with the drug response assay.

Funding

MR was funded in part by a National Library of Medicine training fellowship (T15LM007124). AHB, WEJ, DFJ, SMM, LH, and JG were funded by (U01CA164720).

Availability of Data and Materials

The datasets supporting the conclusions of this article and instructions for how to download them are available in the GitHub repository titled "GRFN_signatures" found at https://github.com/mumtahena/GFRN_signatures. Gene expression signatures can be found at the GEO under accessions GSE83083 and GSE59765.

Author Contributions

AHB and WEJ conceived of the study; AHB, WEJ, MR, JWG, LH, and SMM designed the study; SRP set up the initial bioinformatics pipeline; MR, SMM, DFJ, and SRP performed bioinformatics and data analysis; SM, GS, SWR, and JAM performed the experimental work. MR, SM, DFJ, AHB, and WEJ wrote the manuscript; SRP, JAM, SWR, LWG, and JG provided crucial manuscript feedback and suggestions. All authors read and approved the final manuscript.

Competing Interests

The authors declare that they have no competing interests.

Ethics Approval and Consent to Participate

All research involving human samples has been approved by the University of Utah Institutional Review Board. All research conformed to principles of the declaration of Helsinki. With informed consent, breast tissue samples were collected from patients at the University of Utah at time of surgery for Human Mammary Epithelial Cell preparations.

Chapter 4. Pathway signature profiling of tuberculosis RNA-Seq data

Introduction

Tuberculosis (TB) is an infectious disease that is among the top ten causes of death worldwide (World Health Organization, 2016). Active TB disease is treated with a 6 to 9 month course of antibiotics (Dorman, Chaisson, 2007). In India, the treatment success rate for new and relapse patients in 2016 was 69% (World Health Organization, 2016). Predicting and understanding why some patients eventually fail TB treatment could help personalize TB treatment and improve treatment outcomes.

Previously, gene expression biomarkers have been developed to detect patients with active TB disease, patients that are at risk of TB treatment failure, or patients that have a latent TB infection that is likely to progress to active TB disease (Zak et al., 2016, Bloom et al., 2013, Suliman et al., 2018, Thompson et al., 2017, Leong et al., 2018). Gene lists can be analyzed using tools such as Gene Set Variation Analysis (GSVA), Single Sample Gene Set Enrichment analysis (ssGSEA), or Adaptive Signature Selection and Integration (ASSIGN) to create a single score that represents the activity of the set of genes, which can then be used as a predictor (Hänzelmann, Castelo & Guinney, 2013, Shen et al., 2015, Barbie et al., 2009). These gene signature scores can also be used to stratify samples into groups that show similar TB signature activity. These groups can be used to understand the heterogeneous response to TB and help identify the pathways and underlying biology of TB disease progression.

To assist researchers in applying a large set of TB signatures to available datasets, a set of 30 previously published signatures of TB disease was collected. This set of signatures has been included in the TB Signature Profiler, a novel R package that allows users to quickly and easily perform pathway enrichment analysis using a set of signatures and multiple methods in an easy to use analysis framework for profiling and visualizing these pathways using simple, user friendly R functions.

The TB Signature Profiler was applied to a novel TB dataset to understand gene expression differences between TB patients who successfully cleared active TB disease and those who failed treatment. Samples from TB patients were collected and monitored over time. After the course of treatment, baseline samples from patients that successfully treated TB and those that failed treatment were sequenced. Additionally, samples from treatment failure patients were also sequenced at a two-month mid-treatment timepoint. Decreased predicted TB pathway activity was observed in the month two treatment failure samples when compared to baseline samples. Additionally, treatment failure samples from patients that reported missed treatment doses showed higher TB signature activity when compared to patients that reported adherence to the prescribed treatment. No previously published signature was able to accurately predict treatment failure at baseline, and no significant differentially expressed genes that could stratify treatment failure samples were found. These results serve as an example of the kind of analysis that can be performed using the TB Signature Profiler.

Methods

Sample Processing and Sequencing

Patients with active TB were monitored over the course of TB treatment. Samples were collected at baseline, and at two-month timepoints. After treatment, subjects were categorized as either control (successful TB treatment) or failure (TB treatment failure). RNA sequencing libraries were prepared for samples from 21 baseline control samples, 20 baseline failure samples, and 20 month two failure samples. Multiplex Illumina sequencing was performed on the samples using 100 base pair paired end reads to yield an average of 37 million read pairs per sample.

RNA-Seq Data Analysis

Quality control was performed on the raw sequencing FASTQ read files using FastQC and MultiQC (Babraham Bioinformatics, 2011, Ewels et al., 2016). Reads were aligned to the human reference genome (hg19) using Rsubread, version 1.30.5 (Liao, Smyth & Shi, 2013). Samples had an average alignment percentage of 82% (range 62-89%). Read counts for each gene were calculated using the `featureCounts()` function from the Rsubread package and gene annotations from the UCSC refGene database (Karolchik et al., 2004). An average of 70% of the reads were successfully assigned to a gene (range 55-75%). The count matrix was normalized and scaled by calculating fragments per kilobase of transcript per million mapped reads (FPKM) and transcripts per million (TPM) values. Normalized values were log transformed for downstream analysis. The count matrix, normalized matrices, and annotation information were loaded into R and stored in a SummarizedExperiment object for downstream analysis (Huber et al., 2015). To check for obvious outlier samples, random subsets of genes were visualized using a heatmap. One baseline failure sample showed consistently anomalous expression that could not be corrected and was excluded as an outlier.

For additional analysis comparing the failure data to LTBI samples, the failure dataset was combined with the India TB vs. LTBI RNA-Seq dataset available from GEO at GSE101705 (Leong et al., 2018). Log normalized TPM values were combined and batch correction using ComBat was performed to remove the sequencing batch effect between the two datasets (Johnson, Li & Rabinovic, 2007).

Collection of Published TB Signatures

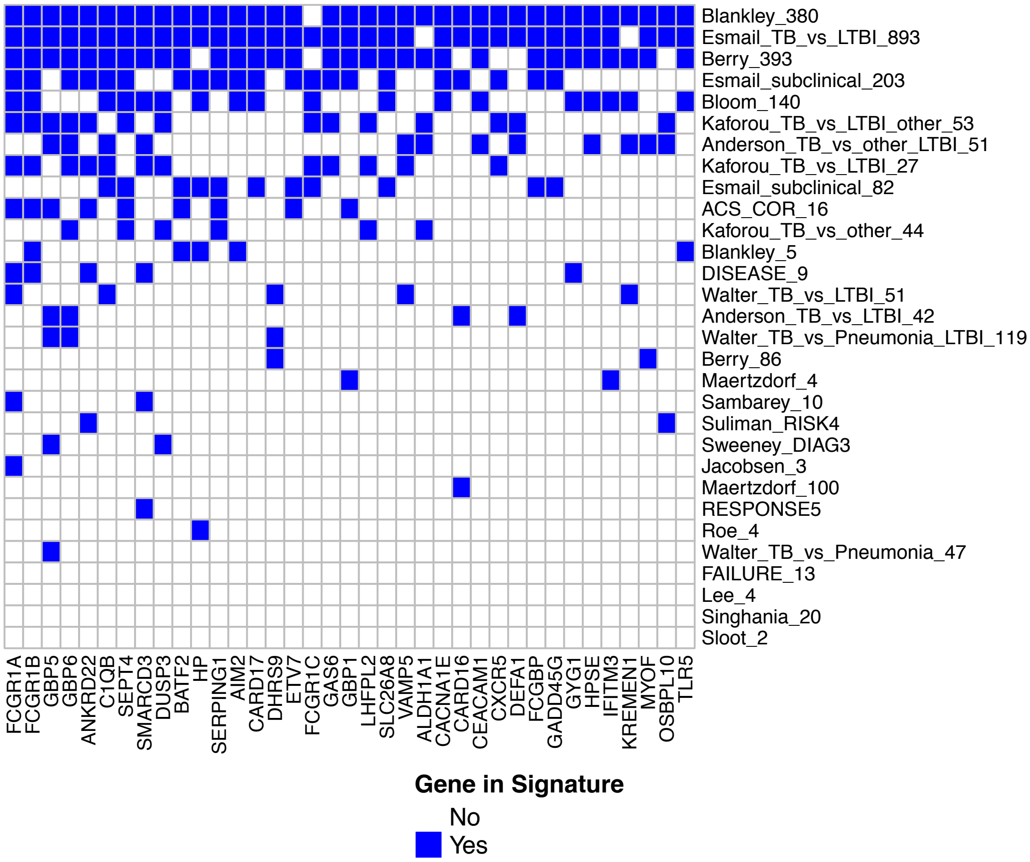


Figure 4.1. Overlap of genes in the TB signature cohort listed in 5 or more signatures. Of the 1,392 unique genes in the 30 signatures, 37 are listed in 5 or more signatures. The majority of these signature genes are contained in the large Esmail 893 gene, Berry 393 gene, and Blankley 380 gene signatures.

A set of previously published gene signatures of TB disease and TB disease progression were collected. Signatures designed to distinguish TB disease vs LTBI or healthy samples include the 16 gene "ACS_COR" signature (Zak et al., 2016), the 393 gene Berry signature (Berry et al., 2010), the 380 gene Blankley signature (Blankley et al., 2016), the 893 gene Esmail signature (Esmail et al., 2018), the 3 gene Jacobsen signature (Jacobsen et al., 2007), the 27 gene Kaforou signature (Kaforou et al., 2013), the 4 gene Lee signature (Lee et al., 2016), the 4 gene Suliman “RISK4” signature (Suliman et al., 2018), the 51 gene Walter signature (Walter et al., 2016), and the 42 gene Anderson signature (Anderson et al., 2014). Signatures that distinguish TB disease vs LTBI or other diseases include the 51 gene Anderson signature (Anderson et al., 2014), the 86 gene Berry signature (Berry et al., 2010), the 140 gene Bloom signature (Bloom et al., 2013), the 53 gene Kaforou signature (Kaforou et al., 2013), the 44 gene Kaforou signature (Kaforou et al., 2013), the 100 gene Maertzdorf signature (Maertzdorf et al., 2012), the 4 gene Maertzdorf signature (Maertzdorf et al., 2016), the 4 gene Roe signature (Roe et al. 2016), the 20 gene Singhania signature (Singhania et al., 2018), and the 3 gene Sweeney “DIAG3” signature (Sweeney et al., 2016). The Blankley 5 gene signature distinguishes active TB disease from healthy, LTBI, or post treatment samples (Blankley et al., 2016). The 9 gene “DISEASE” signature, the 13 gene “FAILURE” signature, and the 5 gene “RESPONSE5” signature predict treatment failure and response to treatment (Thompson et al., 2017). The 203 gene and 82 gene Esmail signatures were designed to distinguish subclinical TB disease and LTBI (Esmail et al., 2018). The 10 gene Sambarey signature identifies TB disease from LTBI samples in the context of HIV infection (Sambarey et al., 2017). The 2 gene Sloot signature predicts TB disease progression in the context of HIV (Sloot et al., 2015). Finally, the 47 gene and 119 gene Walter signatures identify TB disease in the context of Pneumonia (Walter et al., 2016). All signature gene lists were compared to the gene annotations in the UCSC hg19 human reference genome. Signature genes that did not have a corresponding gene included in the hg19 gene annotation or genes that mapped to duplicate gene annotations in hg19 were removed. The 30 signatures consist of 1,392 unique TB associated genes. The majority of the genes (878, 63%) are listed in a single signature and 97.3% (1,355 genes) are listed in four or fewer signatures, with 37 (2.7%) genes listed in five or more signatures (Figure 4.1). Genes that occur frequently in the signatures include FCGR1A, FCGR1B, GBP5, and GBP4 (in 12, 10, 10, and 10 of the signatures, respectively).

Differential Expression Analysis

Limma was used to identify differentially expressed genes in the baseline samples (Ritchie et al., 2015). Log transformed TPM values and the default limma parameters were used. An FDR corrected p-value of 0.05 was used to determine if a gene was differentially expressed between baseline control and baseline failure samples. Available covariates were added to the limma differential expression model to see if correcting for other sources of variation would produce differentially expressed genes. Smoking status, diabetes status, cough duration before treatment, random blood sugar, number of alcoholic drinks per day, age, sex, time to positive diagnosis, and smear result were each added to a limma model along with control vs. failure (limma model ~control_vs_failure + covariate). DESeq2, another method for differential expression analysis, was performed to try to identify differentially expressed genes (Love, Huber & Anders, 2014). Raw count values and the default DESeq2 parameters were used.

Gene Set Analysis

Several methods were used to perform gene set enrichment analysis. Single sample GSEA (ssGSEA) is an extension to the GSEA algorithm that provides a gene set enrichment score for each sample in a dataset given a gene list (Barbie et al., 2009). This is accomplished by ranking genes by absolute expression and calculating an enrichment score for the genes in the signature based on this ranking. (Barbie et al., 2009). Gene Set Variation Analysis (GSVA) calculates a similar statistic, but by first calculating an expression-level statistic using kernel estimation (Hänzelmann, Castelo & Guinney, 2013). Adaptive Signature Selection and Integration (ASSIGN) calculates signature activity scores using a Bayesian estimation framework to adapt signature genes to the specific context of the tested samples (Shen et al., 2015). Methods for performing ssGSEA and GSVA are provided in the GSVA package available on Bioconductor (Huber et al., 2015). ASSIGN is available as a standalone package on Bioconductor (Shen et al., 2015). Gene Set Enrichment Analysis (GSEA) was used to create enrichment score plots of signatures in a ranked list of genes based on a given phenotype (Subramanian et al., 2005).

Visualization

The ComplexHeatmap R package was used to create an annotated heatmap of pathway signature activities (Gu, Eils & Schlesner, 2016). Each row of the heatmap represents a TB signature, and each column represents a sample. Pathway activity scores were scaled to highlight the differences in pathway activity across samples and hierarchical clustering was used to identify samples that showed similar patterns of expression. Annotation information was added to the top of the heatmap as a color bar.

Boxplots of signature activity were created using the ggplot2 R package (Wickham, 2010). Pathway activity scores for each signature were grouped based on annotation information.

The ComplexHeatmap R package was used to create annotated heatmaps of individual signature activity (Gu, Eils & Schlesner, 2016). Each row of the heatmap represents a gene in the TB signature, and each column represents a sample. Gene expression was scaled to highlight expression differences between samples. Annotation information was added to the top of the heatmap. Below the annotation information, pathway activity scores were added. The plotROC package was used to calculate AUC values, confidence intervals, and create ROC curves (Sachs, 2017).

Software Availability

Methods for performing gene set analysis using ssGSEA, GSVA, and ASSIGN, and visualizing the results using boxplots and heatmaps of pathway activity predictions were packaged into the TB Signature Profiler R package. The software utilizes the SummarizedExperiment framework to store raw expression data, annotations, and results within a single object. Raw expression data is stored as a set of multiple matrices called assays that must contain identical dimensions. Along with the assay data the user can provide sample annotation and gene annotation data that can be stored in the colData and rowData slots of the SummarizedExperiment object, respectively (Huber et al., 2015). Users run the gene set analysis using the `runTBsigProfiler()` function providing an input SummarizedExperiment object, the assay to use for profiling, and the algorithms to use for gene set analysis. The results of this analysis are per sample gene set enrichment scores that are stored in the colData slot of the SummarizedExperiment object that is returned by the `runTBsigProfiler()` function. The resulting gene set enrichment results can be visualized as described above using the `signatureHeatmap()`, `signatureBoxplot()`, and `signatureGeneHeatmap()` functions. The software is available on GitHub https://github.com/compbiomed/TBSignatureProfiler.

Results

Previously published TB signatures show decreased TB signature activity at month two in TB failure samples

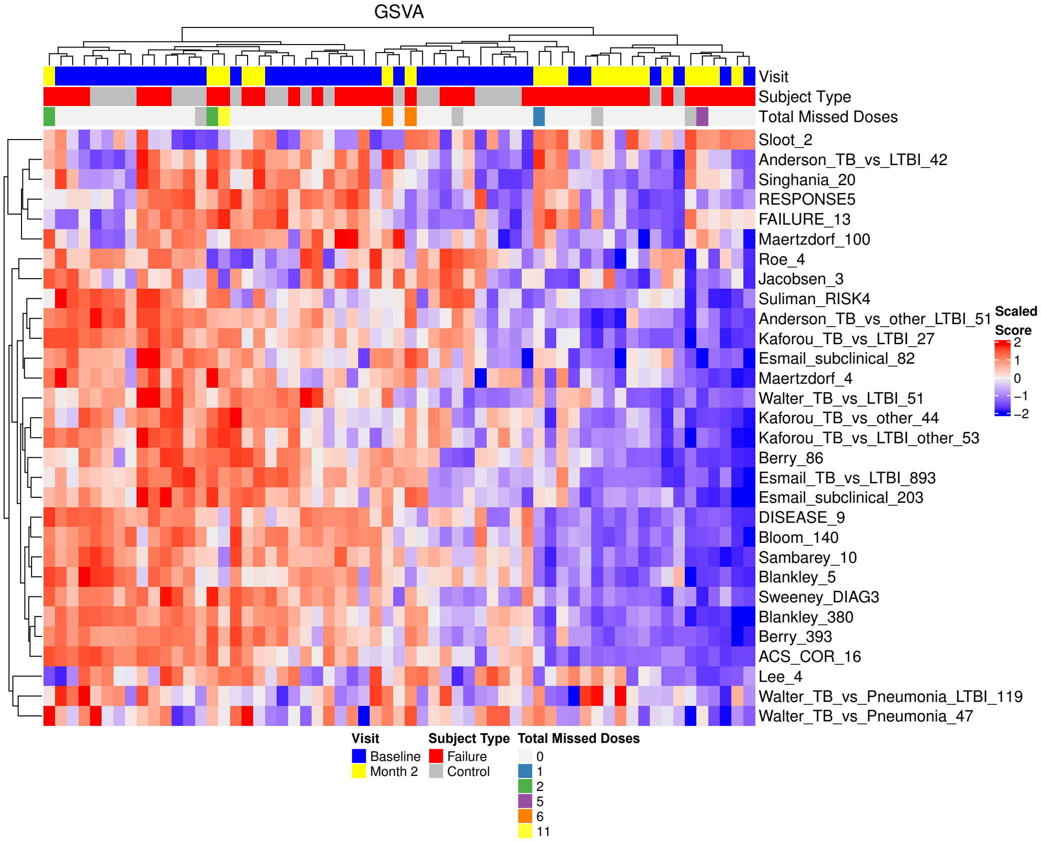


Figure 4.2. Scaled GSVA pathway activity scores for baseline failure, baseline control, and month two failure samples. GSVA pathway activity scores are elevated in baseline control and failure samples and appear to decrease in the month two samples. Month two samples that have elevated TB signature activity tend to be from patients that reported missing doses during treatment.

Using the TB Signature Profiler, GSVA scores were produced for each sample (Figure 4.2). Since baseline samples come from patients with active TB disease, TB pathway activity scores are elevated in baseline samples. At month two, pathway activity scores in 12 of the 30 tested pathways show significantly decreased activity levels when compared to all baseline samples (FDR corrected p-value < 0.05 in Blankley_5, Bloom_140, Roe_4, Sambarey_10, DISEASE_9, Blankley_380, Kaforou_TB_vs_LTBI_27, ACS_COR_16, Anderson_TB_vs_other_LTBI_51, Berry_393, Jacobsen_3, and Suliman_RISK4). These decreased pathway activity scores could indicate an initial response to TB treatment, despite the fact that these patients eventually progress to treatment failure.

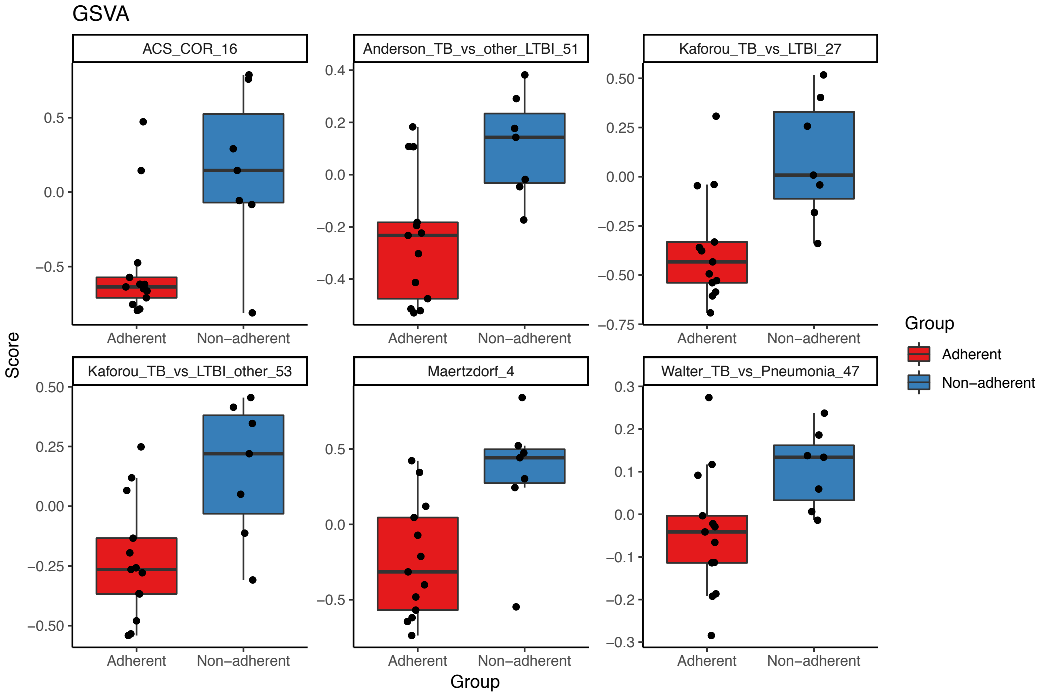


Figure 4.3. Pathway activity scores from month two failure samples. Boxplots are split into adherent and non-adherent groups based on the total number of missed doses reported on the DOTS card. Significant pathway activity differences are observed in 6 of the 30 tested pathways (FDR corrected p-value < 0.05), indicating that patients that adhere to the treatment protocol are showing decreased TB activity when compared to those that are non-adherent.

In some month two samples, pathway activity levels have not decreased, causing these samples to cluster with the baseline samples with higher pathway activity levels. These elevated levels of TB pathway signaling tend to occur in patients that have missed doses in their treatment (as reported on the DOTS card), indicating that this difference could be due to patients not adhering to the treatment protocol. Significant differences between adherent and non-adherent patients were observed in 6 of the 30 pathways (Figure 4.3). If the non-adherent samples are removed, 16 of the 30 pathways identify a significant difference between baseline and month 2 failure samples (FDR corrected p-value < 0.05 in ACS_COR_16, Anderson_TB_vs_other_LTBI_51, Berry_393, Blankley_380, Blankley_5, Bloom_140, DISEASE_9, Jacobsen_3, Kaforou_TB_vs_LTBI_27, Kaforou_TB_vs_LTBI_other_53, Kaforou_TB_vs_other_44, Roe_4, Sambarey_10, Suliman_RISK4, Sweeney_DIAG3, and Walter_TB_vs_LTBI_51).

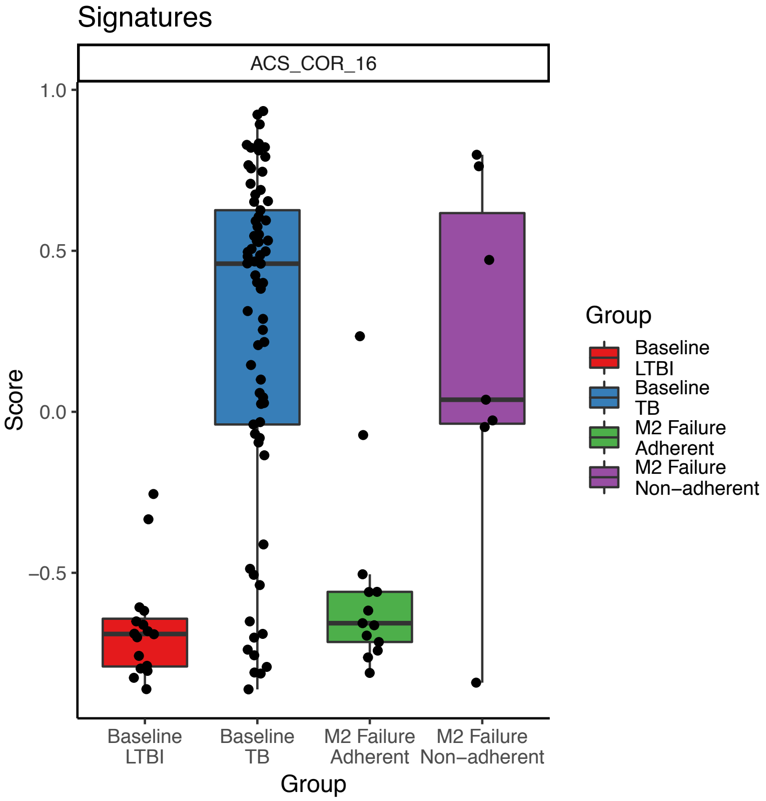


Figure 4.4. Boxplot of ACS_COR signature scores in combined India failure and Leong et al. India datasets. LTBI samples show decreased pathway activity estimates for ACS_COR when compared to baseline TB and month 2 failure non-adherent samples, but show similar pathway activity scores when compared to the month 2 adherent samples.

To compare the adherent and non-adherent samples to additional Indian TB samples, the failure data was combined with a previously published dataset (Leong et al., 2018). The Leong et al. dataset contains 28 samples with active TB disease and 16 LTBI samples, which have previously been shown to have differences in pathway activity signatures. After combining this data with the India failure dataset and adjusting for batch effects using ComBat, GSVA was used to profile the samples using the ACS_COR 16 gene signature. Baseline samples with active TB disease showed elevated ACS_COR signature scores when compared to the LTBI samples and the non-adherent month 2 samples (Figure 4.4). Adherent month two samples show decreased ACS_COR signature scores similar to those of LTBI samples.

Existing signatures of TB fail to distinguish TB treatment failures at baseline

Signature	**AUC (95% CI)**
ACS_COR (16 gene)	0.543 (0.357-0.721)
Anderson TB vs. LTBI (42 gene)	0.610 (0.431-0.776)
Anderson TB vs. other/LTBI (51 gene)	0.507 (0.329-0.688)
Berry (393 gene)	0.583 (0.402-0.760)
Berry (86 gene)	0.576 (0.395-0.752)
Blankley (380 gene)	0.679 (0.505-0.833)
Blankley (5 gene)	0.543 (0.276-0.643)
Bloom (140 gene)	0.648 (0.469-0.807)
DISEASE (9 gene)	0.562 (0.381-0.736)
Esmail subclinical (203 gene)	0.576 (0.393-0.752)
Esmail subclinical (82 gene)	0.510 (0.305-0.679)
Esmail TB vs LTBI (893 gene)	0.579 (0.395-0.748)
FAILURE (13 gene)	0.548 (0.362-0.721)
Jacobsen (3 gene)	0.643 (0.459-0.812)
Kaforou TB vs LTBI (27 gene)	0.576 (0.393-0.750)
Kaforou TB vs LTBI/other (53 gene)	0.550 (0.364-0.736)
Kaforou TB vs other (44 gene)	0.614 (0.443-0.783)
Lee (4 gene)	0.662 (0.476-0.833)
Maertzdorf (100 gene)	0.543 (0.360-0.721)
Maertzdorf (4 gene)	0.600 (0.417-0.771)
RESPONSE5 (5 gene)	0.538 (0.355-0.717)
Roe (4 gene)	0.521 (0.290-0.664)
Sambarey (10 gene)	0.567 (0.381-0.741)
Singhania (20 gene)	0.657 (0.486-0.817)
Sloot (2 gene)	0.571 (0.388-0.743)

Suliman RISK4 (4 gene)	0.593 (0.405-0.769)
Sweeney DIAG3 (3 gene)	0.552 (0.371-0.731)
Walter TB vs LTBI (51 gene)	0.505 (0.317-0.676)
Walter TB vs Pneumonia (47 gene)	0.514 (0.333-0.700)
Walter TB vs Pneumonia/LTBI (119 gene)	0.555 (0.362-0.738)

Table 4.1. AUC Values and 95% confidence intervals for pathway activity predictions using GSVA scores to predict failure samples. All existing signatures show poor predictive power for pathway failure. All 95% confidence intervals contain 0.5 with the exception of the Blankley 380 gene signature.

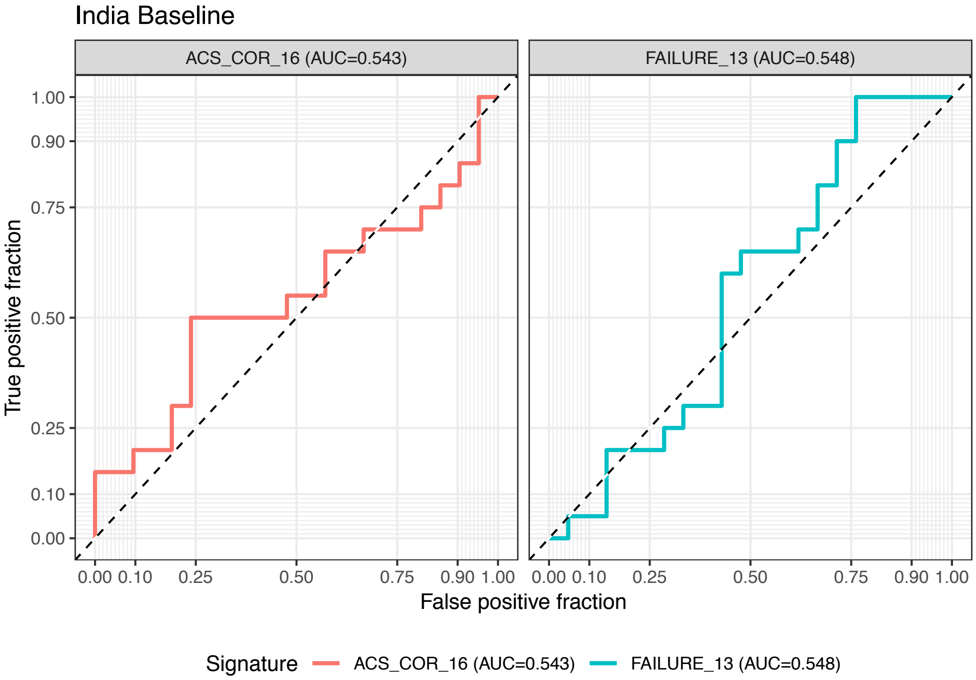


Figure 4.5. ROC Curves of ACS_COR and FAILURE signatures in baseline samples. Previously published signatures of TB activity and treatment failure fail to distinguish TB samples vs controls at baseline in the India data. Left: ACS_COR AUC=0.543 (95% CI: 0.362-0.724). Right: FAILURE_13 AUC=0.548 (95% CI: 0.364-0.721).

None of the GSVA scores for the 30 TB signatures show a significant difference in pathway activity levels at baseline (FDR corrected p-value > 0.6 for all pathways). When visualized using a heatmap, pathway activity levels at baseline do not separate by treatment failure vs control (Figure 4.2). The AUC values for all thirty signatures show poor predictive ability to distinguish the samples at baseline (Table 4.1). Previously, a thirteen gene signature of treatment failure was produced and shown, along with the ACS_COR signature, to accurately predict treatment failure in TB samples (Thompson et al., 2017). At baseline, these genes fail to separate baseline failure and baseline control samples in this cohort (ACS_COR: AUC=0.543 (95% CI 0.357-0.721) Figure 4.5 left, FAILURE: AUC=0.548 (95% CI 0.362-0.721) Figure 4.5 right, Figure 4.6). To test if this failure signature was enriched in our treatment failure samples at baseline, a GSEA analysis was performed. The genes show no enrichment when compared to all genes ordered by their difference between control and failure samples (FDR corrected p-value > 0.99 Figure 4.7).

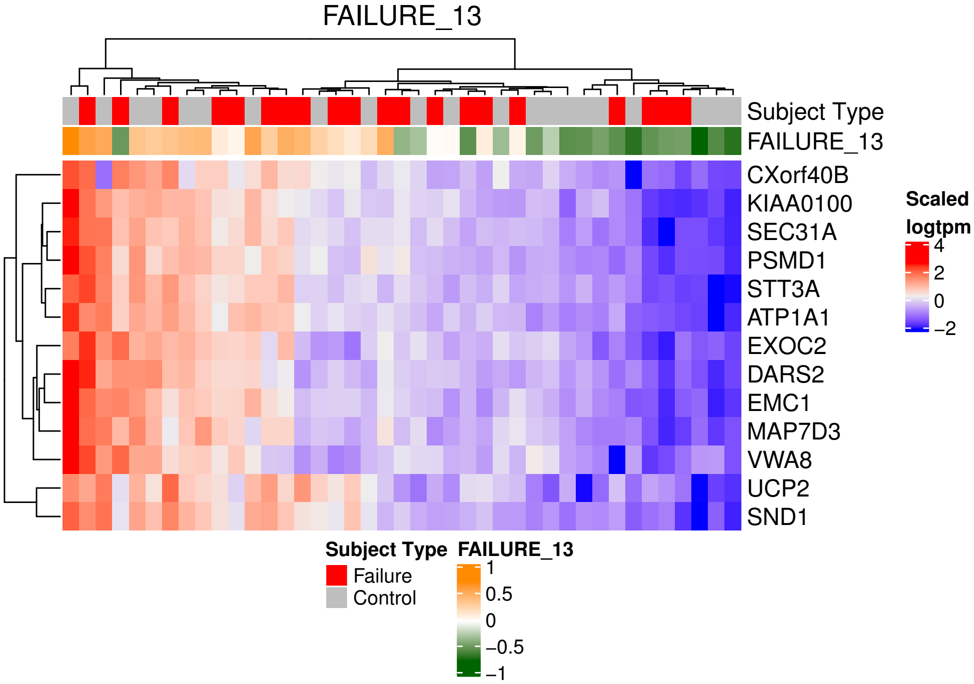


Figure 4.6. Heatmap of row scaled log(TPM) gene expression data for the FAILURE 13-gene signature in baseline India samples. Gene expression differences in the 13 gene signature do not separate failure and control samples. The top color bar indicates the subject type and the second color bar indicates the predicted pathway activity scores from GSVA from the FAILURE signature. The separation in this heatmap is not associated with sequencing batch or any other annotation information available for these samples.

To confirm the pathway activity methods employed by the TB Signature Profiler can be used to effectively predict pathway activity differences using this signature, a reanalysis of the dataset in Thopmson et al. was performed. The Thompson et al. cohort contains samples from patients with active TB at several time points (baseline, day seven, week 4, and week 24) that are categorized into groups of “Not Cured”, “Possibly Cured”, “Probably Cured”, and “Definitely Cured”. Baseline “Not Cured” samples (n=7) show a significantly lower failure signature score when compared to baseline “Definitely Cured” samples (n=71). To determine if the failure signature is an effective predictor of treatment failure at baseline in the Thompson et al. cohort, AUC values were calculated (Figure 4.8 right, AUC=0.938, 95% confidence interval 0.865-0.988).

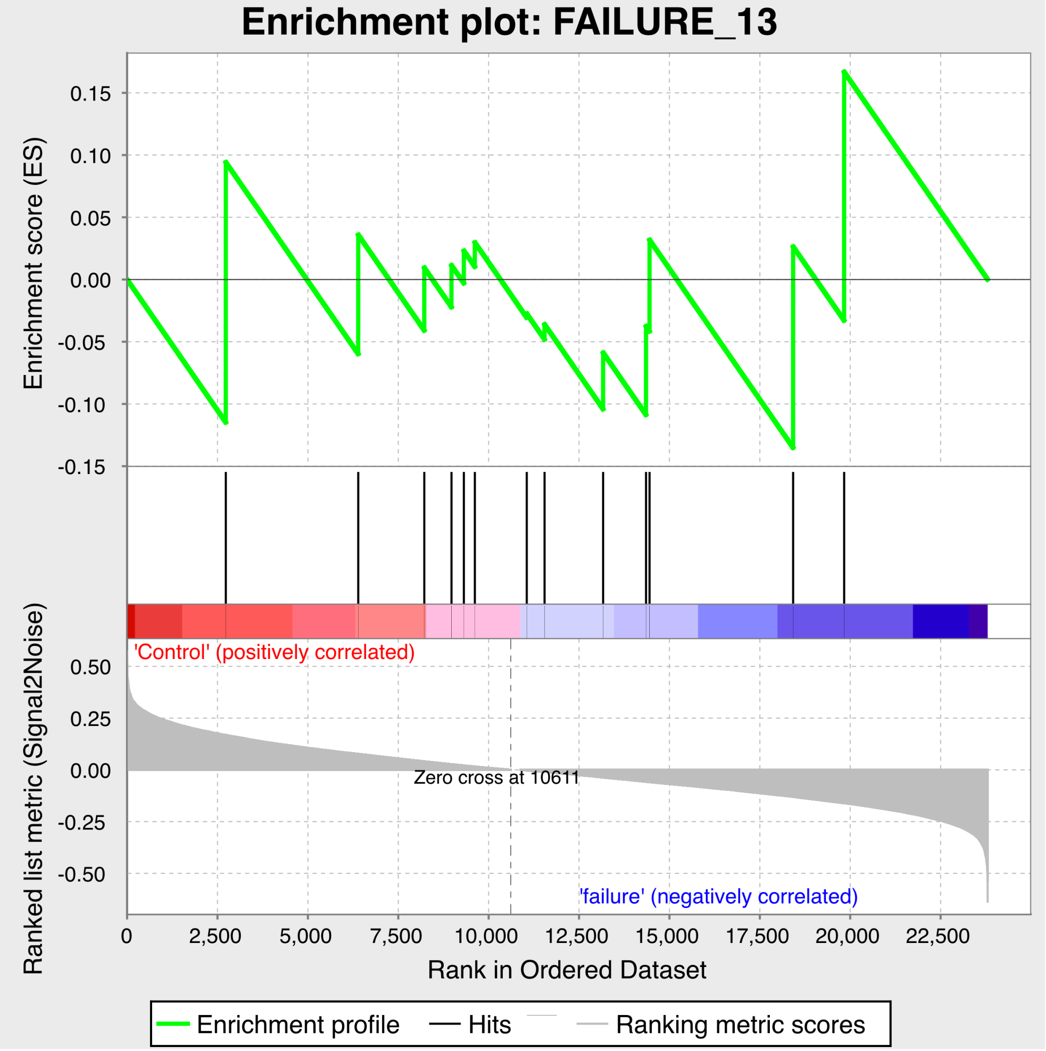


Figure 4.7. GSEA enrichment of 13-gene failure signature on baseline samples using gene set enrichment analysis. Enrichment score indicates enrichment of the gene list of interest in all genes sorted by their difference between control and failure samples. FAILURE genes are not significantly enriched in up- or down- regulated genes (FDR corrected p-value > 0.99).

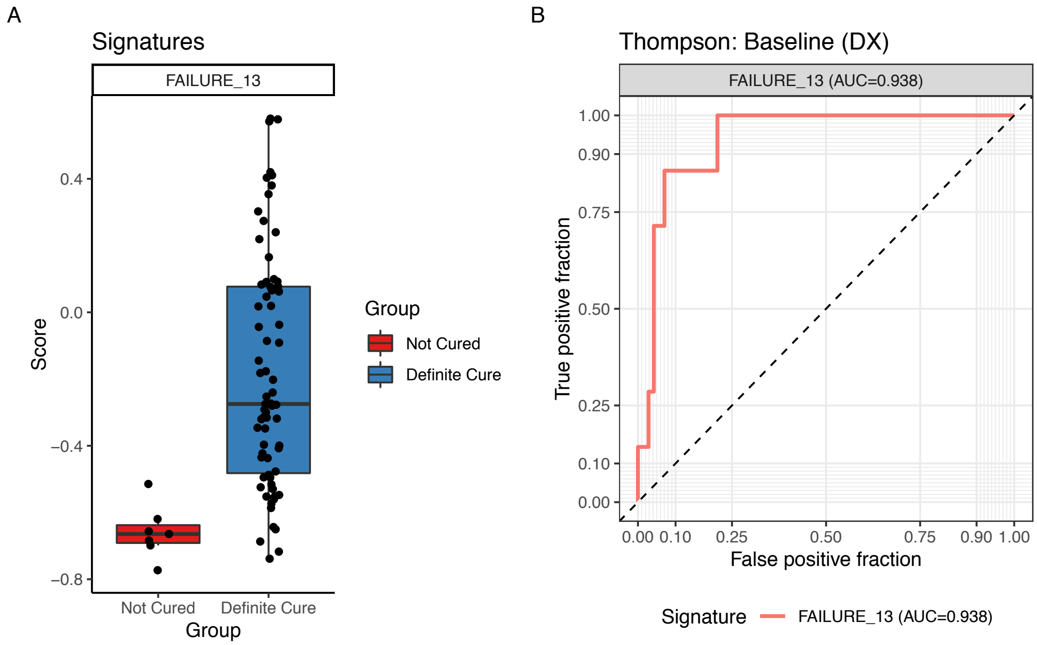


Figure 4.8. GSVA pathway activity scores for FAILURE signature in baseline Thompson et al. samples. a. Boxplot of GSVA pathway activity scores at baseline for not cured and definite cure samples. **b.** ROC curve for GSVA scores to predict treatment failure at baseline in the Thompson et al. cohort AUC=0.938 (95% CI: 0.865-0.988).

To test if the FAILURE signature is overfit to the Thompson dataset, "Not Cured" and "Definite Cure" labels were randomly shuffled across the 78 baseline samples used to create the signature. Limma was used to identify 13 genes that best separate the two groups of samples using the shuffled labels. The GSVA score was then calculated for the shuffled labels and the AUC was calculated. This process was repeated 10,000 times. The median AUC of the iterations was 0.843 with 23.1% of the AUC values being 0.938 or higher, indicating this data produces a signature that is overfit.

No significantly differentially expressed genes separated baseline controls and TB treatment failures

Since none of the existing TB signatures effectively predicted TB treatment failure at baseline in the India cohort, differential expression analysis was performed to identify genes that showed significant differences in expression at baseline between control and failure samples. Using limma, no genes were found to be significantly differentially expressed at a corrected p-value < 0.05. To ensure no additional sources of variability were affecting the analysis, covariates from available sample annotations were added to the limma model including smoking status, diabetes status, cough duration before treatment, random blood sugar, number of alcoholic drinks per day, age, sex, time to positive diagnosis, and smear result. Again, no genes reached a significance threshold p < 0.05 with any of the covariates added to the model. No differentially expressed genes at an FDR corrected p-value < 0.05 were identified.

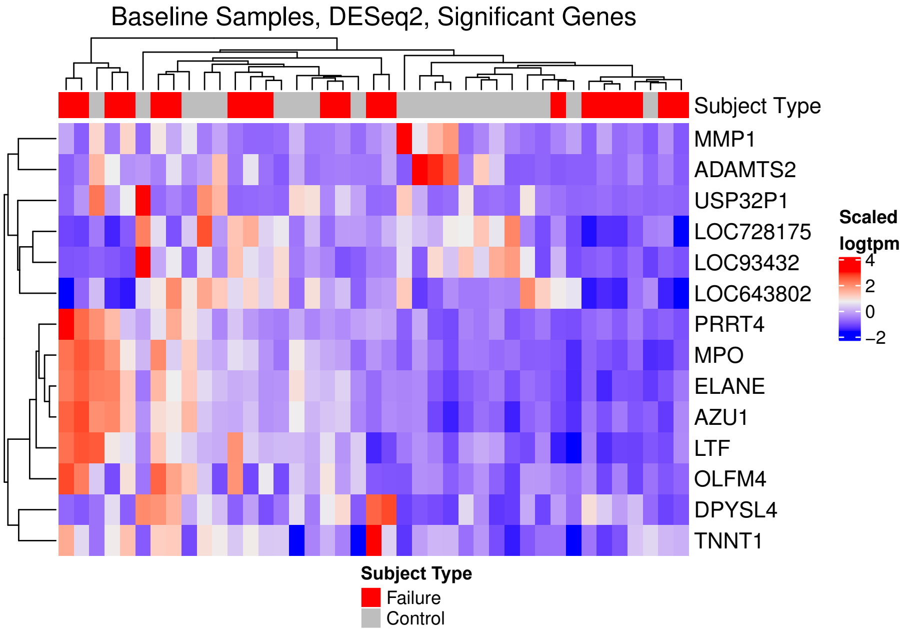


Figure 4.9. Differentially expressed genes at baseline as identified by DESeq2. Differential expression analysis using DESeq2 was performed on the baseline samples to identify genes that differentiate failure and control samples. Fourteen genes were identified using an FDR corrected p-value < 0.05 and a minimum absolute fold change of 2. When clustered using hierarchical clustering, these genes do not separate the data.

To validate this result, DESeq2, another method for differential expression analysis, was performed to try to identify differentially expressed genes. 1,699 differentially expressed genes were identified (FDR corrected p-value < 0.05). Of these 1,699 genes, only 14 genes were found to be differentially expressed with a fold change greater than two in either direction. When visualized, these significant genes failed to separate the differentially expressed genes, indicating that these genes do not represent a consistent potential biomarker for disease failure in the India cohort (Figure 4.9).

Discussion

We have created the TB Signature Profiler, an R package for calculating and visualizing pathway activity scores using currently available tools including GSVA, ssGSEA, and ASSIGN. Because our package leverages the SummarizedExperiment framework, users can easily store their raw gene expression data, annotation information, and pathway activity scores together in a single R object, which can then be visualized. The `signatureHeatmap()` function can be used to plot a heatmap of pathway activity scores along with annotations (Figure 4.2). The `signatureBoxplot()` function can be used to create a boxplot of pathway activity scores based on a sample annotation (Figure 4.3, Figure 4.8, left). Finally, the `signatureGeneHeatmap()` function can create a heatmap of an individual signature gene displaying gene expression values, annotation information, and pathway activity scores in a single plot (Figure 4.6). This tool allows researchers to profile a large set of previously described TB signatures automatically. Additionally, users can modify this list to subset the list of pathways to a specific set of interest, or add additional signatures as they become available. By visualizing these pathway activity scores together, users may identify additional heterogeneity that would not be visible using a single pathway activity prediction.

Using the TB signature profiler on our cohort of TB failure samples, we have identified a significant difference between baseline and month two failure samples, where the majority of the month two samples show a decrease in TB pathway activity scores.

This could indicate a response to treatment at two months in these samples, despite the fact that they will eventually go on to fail TB treatment. For samples that showed an elevated level of TB signature activity, these samples tended to be ones that reported missed doses during their treatment, which could indicate the utility of TB pathway signatures to help detect study adherence in TB cohorts, which could serve as a reliable biological check to ensure patients take their medications and that they are showing a response to treatment. Further, treatment adherence can have a confounding effect on pathway activity measurements and controlling for it could increase the predictive ability of TB biomarkers.

Using an existing signature of TB failure on our baseline samples, we failed to distinguish control samples from samples that eventually fail TB treatment. The previously published signature did not show enrichment in genes that show a difference in expression in the cohort, and appeared to be overfit in the dataset that was used to create it. Further, no previously published signature could be used to effectively predict TB treatment failure at baseline.

Differential expression analysis failed to identify a set of genes that reliably separate baseline control and failure samples in our dataset, despite controlling for various other possible sources of variation within the data. Samples appeared to cluster at random and any difference that was identified was minimal and not useful as a biomarker for TB failure. Although none of the available sample annotation information or sequencing batch information for these samples showed a significant association with the clustering, it cannot be ruled out that some unknown source of variation is masking true signal.

To our knowledge, this is the largest set of previously published TB signatures that has been collected, and by using the TB Signature Profiler researchers will be able to leverage this set to profile their own datasets with minimum effort. Standardizing methods of pathway activity measurements will make the results of this analysis more consistent across studies and allow more direct comparisons between cohorts, leading to easier meta-analysis and new insights and better predictive and mechanistic insights into tuberculosis.

Acknowledgments

We thank all of the study participants who made this study possible.

Chapter 5. Conclusion

The work presented in this dissertation represents a set of software tools and data resources that can be used across RNA-Seq technologies and disease areas to analyze and visualize RNA-Seq data. Specifically, novice users can use the SCTK to go from raw count data from scRNA-Seq experiments and perform common analysis and visualization methods interactively without writing any R code. This is the first time that a complete scRNA-Seq analysis workflow has been implemented in an easy to use point-and-click environment. The SCTK software framework is extendable and under active development, which will increase its utility with additional quality control visualizations, analysis techniques, and novel methods as they are developed. In the context of breast cancer, the set of novel growth factor receptor signatures that were created identify additional heterogeneity beyond currently available breast cancer subtyping through immunohistochemistry and showed differences in response to drug therapies. By directly profiling the biologically relevant pathways that can be targeted in breast cancer, the specific drivers of individual tumors can be identified, which could help stratify patients to give them the drugs that will target the specific oncogenic pathways driving their tumor. Finally, a pathway activity approach is also useful in tuberculosis, where existing pathway signatures can help stratify patient samples based on their pathway activity. By building the TB Signature Profiler, users can rapidly profile samples, compare the pathway predictions across multiple TB signatures, or develop new signatures to further stratify their samples, which could lead to a deeper understanding of the underlying pathway activity in latent tuberculosis infection or during active disease and improved monitoring during TB treatment to ensure drug protocol adherence and ensure that drug therapy is working.

By creating software frameworks and data resources for scientists, the approaches developed here can be extended and expanded to address the changing needs of the RNA-Seq analysis environment. In the case of the SCTK, additional analysis modules for cell type prediction, improved data handling, and support for larger and more complex datasets and meta-datasets can be developed within the existing SCTK framework. Additional pathway signatures targeting novel therapeutic targets and cell growth pathways will be developed using the methodologies developed to create the GFRN signatures. Finally, additional pathway targets can be profiled using the TB Signature Profiler framework, extending its utility beyond the 30 collected signatures.

Taken together, these tools represent a novel and widely applicable set of user-friendly software tools and resources. These tools serve as a model of how analysis techniques can be packaged and be made available to users without a deep understanding of the underlying methodologies, but still allow users to perform sophisticated analyses using their own and public data resources, helping leverage these techniques across disease areas and address unmet diagnostic need.

APPENDIX A: Cell lines used in the independent drug assay and the Western blotting experiments.

Cell Line	Dose Response Assay	Western Blots for apoptotic proteins	ER Status	PR Status	HER2 Status	Intrinsic Subtype	Source	Growth Media
AU565	Y	Y	Negative	Negative	Positive	HER2-Luminal	ATCC	A
BT549	Y	Y	Negative	Negative	Negative	Claudin-low	ATCC	B
HCC1143	Y	Y	Negative	Negative	Negative	Basal	ATCC	B
HCC1395	N	Y	Negative	Negative	Negative	Claudin-low	ATCC	B
HCC1419	Y	Y	Negative	Negative	Positive	HER2-Luminal	ATCC	B
HCC1569	Y	N	Negative	Negative	Positive	HER2-Basal	ATCC	B
HCC1806	Y	Y	Negative	Negative	Negative	Basal	ATCC	B
HCC1937	Y	Y	Negative	Negative	Negative	Basal	ATCC	B
HCC1954	Y	Y	Negative	Negative	Positive	HER2-Basal	ATCC	B
HCC2218	Y	Y	Negative	Negative	Positive	HER2-Luminal	ATCC	B
HCC3153	N	Y	Negative	Negative	Negative	Basal	Adi Gazdar (University of Texas-Southwestern Medical Center)	B
HCC38	Y	Y	Negative	Negative	Negative	Claudin-low	ATCC	B
HCC70	Y	Y	Negative	Negative	Negative	Basal	ATCC	B
Hs578T	Y	Y	Negative	Negative	Negative	Claudin-low	ATCC	B
JIMT1	Y	Y	Negative	Negative	Positive	HER2-Basal	ATCC	A
MDAMB 231	N	Y	Negative	Negative	Negative	Claudin-low	ATCC	B
SKBR3	Y	N	Negative	Negative	Positive	HER2-Luminal	ATCC	B
ZR75B	N	Y	Negative	Negative	Negative	Luminal	Mark Lippman (National Cancer Institute)	B
21PT	N	Y	Positive	Unavailable	Unavailable	HER2-Basal	Ruth Sager (Dana–Farber Cancer Institute)	C
BT474	Y	Y	Positive	Positive	Positive	HER2-Luminal	ATCC	A
BT483	Y	Y	Positive	Positive	Negative	Luminal	ATCC	B
CAMA1	Y	Y	Positive	Negative	Negative	Luminal	ATCC	A
HCC1428	Y	Y	Positive	Positive	Negative	Luminal	ATCC	B

LY2	N	Y	Positive	Negative	Negative	Luminal	Mark Lippman (National Cancer Institute)	A
MCF7	Y	Y	Positive	Positive	Negative	Basal	ATCC	A
MDAMB 175	N	Y	Positive	Negative	Negative	Unavailable	ATCC	B
MDAMB 361	Y	N	Positive	Negative	Positive	HER2-Luminal	ATCC	A
SUM52PE	N	Y	Positive	Negative	Positive	Luminal	Asterand Bioscience	D
T47D	Y	Y	Positive	Positive	Negative	Luminal	ATCC	B
ZR751	Y	Y	Positive	Negative	Negative	Luminal	ATCC	B
ZR7530	Y	Y	Positive	Negative	Positive	HER2-Luminal	ATCC	B

Growth Media A: DMEM (Gibco), 10% FBS (Sigma), 1% Anti/Anti (Life Technologies)
Growth Media B: RPMI (Gibco), 10% FBS (Sigma), 1% Anti/Anti (Life Technologies)
Growth Media C: DMEM/F12 (Gibco), 5% FBS (Sigma), 10 ug/mL Insulin, 100 ng/mL Cholera Toxin, 20 ng/mL EGF, 500 ng/mL Hydrocortisone
Growth Media D: F12 (Gibco), 5% FBS (Sigma), 5 μg/ml insulin, and 1 μg/ml hydrocortisone

APPENDIX B: Gene list of optimized gene numbers determined for each GFRN signature.

20 Gene AKT Signature

AKT1, CD248, IGFBP3, SPRR2A, CA9, BEX1, IGFBP5, EPGN, PPP1R3C, GRHL3, TNFAIP2, AKAP12, CTGF, ICAM1, LIF, CXCL3, DKK1, ITGB3, CXCL2, CXCL5

250 Gene BAD Signature

BAD, KLF2, LCE1F, RFC3, C8orf84, BOLA3, DLEU1, MRPS12, PTGES, SLC16A9, PIK3R3, COTL1, LINC00239, NOP16, OPCML, MPV17L2, NEK6, AIMP2, POLR3G, SRM, SPINK6, C19orf48, CKS2, PRMT3, SLC25A15, PAICS, PMM2, CYCS, C14orf1, DCTPP1, C20orf27, CDC20, NETO2, GBP6, LSM2, TFAP4, RBBP8, ISCA1, PRADC1, MYL9, ORC6, PYCRL, PLA2G7, C11orf82, SLC25A10, PPIF, MRPS2, LOC100506895, FAM216A, LOC100506844, TMEM241, CYB5B, NME4, UFSP1, RHOB, TIPIN, LINC00162, CHCHD8, OSR1, EGFLAM, CDK4, FLJ39051, NME1, NEFL, MBLAC2, FLJ42351, CMC2, ZNF593, LIX1L, SORD, RWDD2B, NIP7, RRM2, ALDH1B1, C3orf26, ALDH1L2, POLR3K, SSR3, PRPS1, RASSF6, RAD51AP1, TOMM5, PDK1, RPP40, RRS1, FAM198B, C21orf63, LOC100128881, RRP9, CHCHD3, FAM86EP, MRPL12, C11orf83, ZDHHC14, TMED2, SFRP1, SELRC1, GPATCH4, CT62, CLEC2D, PDSS1, GAPDH, THEM4, MMACHC, MT1G, LOC401397, MKI67IP, NPM1, TUBA1C, SNORD16, LYAR, POLR3H, LYRM4, RUVBL1, NCL, TOMM20, VIM, TUBA1B, CCNB1, CDT1, COQ2, DCLRE1B, PPAT, C11orf24, PEG10, HSPA6, HSPA7, HSPA1A, IL8, DNAJA4, HSPA1B, CCL20, FOXQ1, GDF15, CXCL5, CRYAB, IL17C, TNFAIP2, CFB, KRT23, CXCL2, SAA2, ATF3, DLC1, FOS, NFKBIZ, MYO5C, PRSS22, ERRFI1, CXCL3, DUSP2, ATHL1, AKAP12, FOSB, LIF, INHBA, GABRE, CDRT1, CXCL6, EGR3, DUSP1, HSP90AA1, ZFAND2A, HMOX1, BMF, RRAD, GSDMB, BIRC3, GRB7, HSPH1, SLC34A2, LTF, FERMT3, SGPP2, GAB2, BAG3, KRT80, HSPB8, DNAJB4, LCN2, DEDD2, CXCL1, TNFRSF11B, DUSP6, MUM1L1, TIAM2, KLHL24, OLFM4, BCORL1, DFNB31, IFRD1, DAPK1, STARD13, ETS1, NFKBID, TLR2, PRDM1, LOC146880, IER5, IER3, DNER, SAA1, PNLDC1, GPRC5A, STON1, ZC3H12A, GSDMC, GM2A, PDZD2, MAFF, GDF6, SBSN, SEMA6C, DNAJC6, PPP1R15A, DUSP5, LIMCH1, CLDN4, RB1CC1, MGAT4A, NYNRIN, DNAJB1, PLEKHA6, FNIP2, ABCA1, PLA2G4C, ULK1, IL7R, ENGASE, C17orf103, SLC24A6, CNNM3, AGAP11, GLCCI1, CCL2, IL17RB, ABCG1, DDIT3, CACHD1, ABTB2, SATB1, INSR, TMEM2, TNFSF14, GCNT2, ARHGAP19, ZNF217, BRD3, CYLD, IL34

50 Gene EGFR Signature

EGR1, IFI6, MT2A, MMP3, MT1X, EGR3, DUSP6, CCNA1, GJB2, IFI27, S100A9, MT1G, LINC00525, IL7R, IFITM1, IL6R, OGFRL1, MT1E, DLL1, SYT12, LTF, WDHD1, SOCS3, MCM5, ODZ2, KRT4, KRT81, SPINK6, SAMD11, KRT86, WISP2, KRT85, ATOH8, HSPB3, MMP7, ALPP, IFITM10, CD24, PGF, DIO2, ID2, CRYAB, HSPB8, IGSF23, CLDN7, DLX3, MAOA, WNT7B, BCO2, EGFR

10 Gene HER2 Signature

ERBB2, PNMA2, PDGFB, EEF1A2, MIR3944, HSPA6, HSPA7, IFIT1, DNAJA4, CCL2

100 Gene IGF1R Signature

IGF1R, BHLHA15, DDIT3, CHAC1, ZSCAN12P1, RND1, CRELD2, PDIA4, C12orf39, HSPA5, ZNF165, ATF3, HERPUD1, STC2, PTX3, FICD, CDC6, SLC7A11, C17orf28, IRF1, SDF2L1, DNAJB9, ANXA6, KLHDC7B, DDR2, ERO1LB, HYOU1, AGR2, CNTD2, SEL1L, HSP90B3P, ADM2, ASNS, HSP90B1, DERL3, CCL2, DNAJC3, PSAT1, MSTO2P, SH3BGR, ALDH1L2, TMEM50B, NUCB2, ICAM1, GDF15, SOCS3, PCK2, KIAA0226L, WARS, FBXO16, DNAJA4, HSPA6, HSPA1A, HSPA7, ACTBL2, CRYAB, HSPA1B, OXTR, CXCL6, HSP90AA1, ATHL1, HMOX1, DKK1, LCE1C, CDSN, ALDH1A3, OLFM4, PDK4, CLDN4, HSPB8, HSPB2, RAD23A, GM2A, HSPH1, C4orf26, HSPA8, DNAJA1, BMP4, BAMBI, SLIT2, HSD17B11, FAM101B, FKBP4, BID, GDF6, BCAS4, CACYBP, TTYH2, RASA3, C10orf10, MAP2K3, FLNC, FAM25A, BAG3, CCNE1, PCGF3, SRRM3, ADCK3, PLSCR4, ANKRD1

200 Gene KRAS (G12V) Signature

MAL, KRAS, LCE3D, DHRS9, LCE3E, NPTX1, IL1RL1, PRSS22, PRR9, DCLK1, AKAP12, S100A7, HAS2, FAM25A, PAPL, LOC100131726, DIO3, KLK6, AGPAT9, ARC, LY6D, NKD2, PAEP, DIRAS3, ANPEP, SPRR2D, CYB5R2, LCE1F, CEACAM1, STC1, HYAL1, SERPINB1, BMP6, AQP5, SPRR1A, FERMT1, TAGLN3, CA6, SCNN1D, LCE1C, TMEM45B, CALB2, SOX8, ANGPTL4, ASPRV1, SLC5A1, CEACAM6, TNFRSF11B, WNT9A, S100P, EEF1A2, ISG20, TRPV3, PLA2G4E, SRMS, PADI1, SH2D2A, GJB4, ADAM8, FAM83A, SULT2B1, CXCL3, CALB1, CNFN, EGR3, G0S2, HBEGF, SERPINB2, FOS, LCE1E, ANO1, APOBEC3A, KCNN4, LOC100505839, EGR1, RHCG, ODC1, RPSAP52, CYP4F22, EMP1, TGM2, PNMA2, TMEM121, AGR2, SCNN1G, PAQR5, SSTR1, LOXL4, DUSP6, SYTL5, S100A1, ZBED2, WNT7B, ROBO4, NGEF, CCNA1, IVL, SOCS1, LIF, KRT18, HSPA1A, HSPA1B, HSPA7, DNAJA4, CCL26, CRYAB, BAG3, HSPB8, HSP90AA1, HSP90AA4P, DNAJB1, ATF3, OXTR, HSPH1, SH3BGR, DNAJB4, CCL2, ACTBL2, HMOX1, ZFAND2A, IL7R, CHAC1, ULBP1, DNAJA1, UBB, GLYATL2, UBC, CDRT1, EPSTI1, FAM49A, BST2, LOC100130238, HSPD1, HSPA8, ID4, TNFAIP2, MGC16121, DUSP8, MB21D1, DLC1, FILIP1L, SESN2, LAMP3, BEX1, CHORDC1, ZNF323, LOC285629, HSPE1, HSP90AA6P, LOC727896, GBP1, CACYBP, IFRD1, C21orf7, FERMT3, MORC4, TMEM27, METTL7A, ABHD3, GREM1, CFB, CCDC117, LIMCH1, ENGASE, LGR5, DFNB31, LCN10, SLC16A14, DIO2, CYFIP2, CLU, ALOXE3, ADM2, IFI44L, NECAB2, ASAP3, COL1A1, ARHGAP24, SLC34A2, MARVELD3, ABCB1, LHFPL2, RGS2, CSRP2, HERC5, ZNF761, MICB, FAM26E, GDF5, ANGPTL7, FKBP4, C4orf49, SOD2, SLC2A12, STIP1, MITF, TRIM22, GSR, BBOX1, DDIT3

200 Gene RAF Signature

RAF1, DHRS9, CA6, SPRR2D, PRSS22, S100A7, STC1, IL1RL1, PAEP, BMP6, LCE3D, HAS2, CEACAM1, FGFBP2, AGPAT9, SPP1, DIO3, DIRAS3, ISG20, TNFRSF11B, LOC100131726, DCLK1, SERPINB1, CRTAM, AQP5, ATP12A, LY6D, FERMT1, ASPRV1, SRMS, CEACAM6, CYB5R2, FAM83A, SLC5A1, SERPINB2, TMEM45B, KLK6, CALB2, SYTL5, CRHR1, GJB4, CCL24, LY6H, SERPINB3, LCE1F, SSTR1, KIAA1199, ENDOU, NTSR1, SCNN1D, PNMA2, EEF1A2, CXCL17, EMP1, TMPRSS4, CXCR1, RLBP1, WFDC3, LCE1E, TMCC3, SPRR3, SMOX, WNT9A, ADAM8, SHC4, HMGA2, GUCY1B3, CEACAM3, HPSE, RPSAP52, NCF2, SNTB1, TAGLN3, PI3, NAV3, SOCS1, PADI1, PKIB, CD55, GPR110, NOX5, NGEF, LBH, FGF1, GAL, S100A4, PLAU, PAPL, SNX9, EDNRA, BPGM, SHF, PLLP, IL23A, FIBCD1, PPBP, B3GNT3, C15orf62, TMEM163, RORB, ANPEP, CHST6, KCNJ15, GLRX, MALL, RASSF8, APOA1, CCNA1, PITPNC1, IRAK2, SLC26A9, TMEM158, CLEC2B, RTKN2, ITGA2, ANO1, ETV5, CLDN10, KCNN4, PLAUR, SDR16C5, GABRA2, PGF, TGFA, LOC100505839, PMP22, RAPH1, RASA3, LRRC8C, FAM176A, ATG16L1, MCTP1, AKAP12, GDNF, CHRNA9, PI15, HBEGF, B3GNT2, MAP1B, ELK3, PTPN22, PTAFR, SPRY4, SH2D2A, STRA6, BMP2, KRT18, CARD11, ETV1, ITGB7, WNT7A, TTC9, SLCO4A1, ODC1, CSGALNACT2, SLC9A2, LY6K, SREK1IP1, GRB7, ROBO4, ARHGAP25, ZPLD1, FAM100B, DAB2, PAQR5, METTL7B, LRAT, SPRY2, SLC1A1, LYPD5, SLC10A6, C14orf49, PRDM8, RAC2, PTPRE, HSPA6, HSPA7, DNAJA4, HSPA1A, HSPA1B, TNFAIP2, ACTBL2, CCL2, STEAP4, ATF3, MGC16121, CRYAB, RASD2, CD248, PIK3C2B, SLC34A2, FILIP1L, EPHA4, ELF3, FAM46B, EPGN, HSPB8, USP2, SLC47A2, CXCR7, ETV7, CCL28, WNT4, CFB, C10orf81, IGFBP5, LOC285629, ANGPTL7, GPR1, EPSTI1, EDN1, EVPLL, SAA2, EPHA3, LIMCH1, CA2, BBOX1, USH1G, SERPINB13, GRAMD2, CXCL12, RARB, PAQR7, CYP1B1, DAPK1, GABRE, APCDD1, ATHL1, CXCL2, SLC27A2, KIT, ZDHHC8P1, KANK4, OXTR, KMO, KCNJ5, NEFM, AMOT, FERMT3, IFI44L, TRIM22, RECK, SYNM, C10orf67, FBXO32, NOTCH1, SEMA5B, DNAJC6, PROM1, CD180, MTUS1, SLC30A10, DNAJB4, SYBU, MYO18B, PLD6, SPINK1, ADM, PCDH19, GBP6, TRIM6, FBXW10, ST6GALNAC5, EFNA5, TMCC2, SYTL2, MTSS1L, FOSL2, METTL7A, TNS3, ENGASE, RASD1, SOSTDC1, ZNF488, FSTL4, CDRT1, ASAP3, SLC2A12, EGFL6, INPPL1, FIGN, TCF4, HS6ST1, PDZK1IP1, PARP9, LRRN1, CORO6, SAA1, ZNF711, CSRP2, DACT1, NAV2, ARRDC4, GDF6, CCRN4L, SSBP2, NEFL, LZTS1, SESN2, FBXW7, LGR5, ESR1, TLR1, ABHD4, SMO, FAM198B, SCD5, MAP3K14, PPP1R3C, NAP1L2, PLK2, COBLL1, KLHDC7B, DLC1, BST2, SOX6, TRIM16L, SOWAHB, BBC3, VAV3, GDF15, TNNI2, ZNF323, TP73, BMP5, CITED2, TRAFD1, FDXR, PNLDC1, TSPYL2, NTN1, PCYOX1L, SOD2, LRRC56, CTH, LXN, PER3, HSPD1, RAB30, CES3, ZNF608, SNHG4, DNAJA1, VGLL3, GLYATL2, OTUD1, ACSL1, LOC283547, PER1, EGLN3

LIST OF JOURNAL ABBREVIATIONS

Adv. Exp. Med. Biol.	Advances in Experimental Medicine and Biology
Adv. Med.	Advances in Medicine
Am. J. Respir. Crit. Care Med.	American Journal of Respiratory and Critical Care Medicine
Arch. Pathol. Lab. Med.	Archives of Pathology and Laboratory Medicine
Biochem. Biophys. Res. Commun.	Biochemical and Biophysical Research Communications
Biochim. Biophys. Acta.	Biochimica et Biophysica Acta
Biomed Res. Int.	BioMed Research International
Breast Cancer Res.	Breast Cancer Research
Breast Cancer Res. Treat.	Breast Cancer Research and Treatment
CA Cancer J. Clin.	CA: A Cancer Journal for Clinicians
Cancer Biol. Ther.	Cancer Biology and Therapy
Cancer Discov.	Cancer Discovery
Cancer Inform.	Cancer Informatics
Cancer Res.	Cancer Research
Cancer Treat. Rev.	Cancer Treatment Reviews
Cell Res.	Cell Research
Clin. Genet.	Clinical Genetics
Clin. Proteomics	Clinical Proteomics
Clin. Transl. Oncol.	Clinical and Translational Oncology:
Core Evid.	Core Evidence

EMBO Mol. Med.	EMBO Molecular Medicine
Endocr. Relat. Cancer	Endocrine-related Cancer
Eur. Respir. J.	The European Respiratory Journal
Front. Endocrinol.	Frontiers in Endocrinology
Genes Cancer	Genes and Cancer
Genome Biol.	Genome Biology
Genome Med.	Genome Medicine
Int. J. Cancer	International Journal of Cancer.
J. Biol. Chem.	The Journal of Biological Chemistry
J. Carcinog. Mutagen	Journal of Carcinogenesis and Mutagenesis
J. Cell. Biochem.	Journal of Cellular Biochemistry
J. Clin. Microbiol.	Journal of Clinical Microbiology
J. Clin. Oncol.	Journal of Clinical Oncology
J. Educ. Psychol.	Journal of Educational Psychology
J. Mach. Learn. Res.	Journal of Machine Learning Research
J. Mol. Med.	Journal of Molecular Medicine
J. Natl. Cancer Inst.	Journal of the National Cancer Institute
J. R. Stat. Soc. Ser. C Appl. Stat.	Journal of the Royal Statistical Society. Series C, Applied Statistics
J. Steroid Biochem. Mol. Biol.	The Journal of Steroid Biochemistry and Molecular Biology
J. Transl. Med.	Journal of Translational Medicine
Lancet Respir. Med.	The Lancet. Respiratory Medicine
Mod. Pathol.	Modern Pathology

Mol. Biol. Int.	Molecular Biology International
Mol. Cancer Ther.	Molecular Cancer Therapeutics
Mol. Oncol.	Molecular Oncology
Mol. Syst. Biol.	Molecular Systems Biology
N. Engl. J. Med.	The New England Journal of Medicine
Nat. Biotechnol.	Nature Biotechnology
Nat. Commun.	Nature Communications
Nat. Med.	Nature Medicine
Nat. Methods	Nature Methods
Nat. Protoc.	Nature Protocols
Nat. Rev. Cancer	Nature Reviews. Cancer
Nat. Rev. Genet.	Nature Reviews. Genetics
Nat. Rev. Mol. Cell Biol.	Nature Reviews. Molecular Cell Biology
Nucleic Acids Res.	Nucleic Acids Research
Pharmacogenomics J.	The Pharmacogenomics Journal
PLoS Comput. Biol.	PLoS Computational Biology
PLoS Med.	PLoS Medicine
PLoS Pathog.	PLoS Pathogens
Proc. Natl. Acad. Sci. U.S.A.	Proceedings of the National Academy of Sciences of the United States of America
Sci. Rep.	Scientific Reports
Sci. Transl. Med.	Science Translational Medicine
Semin. Cancer Biol.	Seminars in Cancer Biology

Ther. Adv. Med. Oncol.	Therapeutic Advances in Medical Oncology
Trends Biochem. Sci.	Trends in Biochemical Sciences

BIBLIOGRAPHY

Akiyama, T., Dass, C.R. & Choong, P.F.M. 2009, "Bim-targeted cancer therapy: a link between drug action and underlying molecular changes", *Mol. Cancer Ther.,* vol. 8, no. 12, pp. 3173-3180.

Anderson, S.T., Kaforou, M., Brent, A.J., Wright, V.J., Banwell, C.M., Chagaluka, G., Crampin, A.C., Dockrell, H.M., French, N., Hamilton, M.S., Hibberd, M.L., Kern, F., Langford, P.R., Ling, L., Mlotha, R., Ottenhoff, T.H.M., Pienaar, S., Pillay, V., Scott, J.A., Twahir, H., Wilkinson, R.J., Coin, L.J., Heyderman, R.S., Levin, M. & Eley, B. 2014, "Diagnosis of childhood tuberculosis and host RNA expression in Africa", *N. Engl. J. Med.,* vol. 370, no. 18, pp. 1712-1723.

Arteaga, C.L. & Engelman, J.A. 2014, "ERBB receptors: from oncogene discovery to basic science to mechanism-based cancer therapeutics", *Cancer Cell,* vol. 25, no. 3, pp. 282-303.

Babraham Bioinformatics 2011, , *FastQC: a quality control tool for high throughput sequence data*. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Badve, S., Dabbs, D.J., Schnitt, S.J., Baehner, F.L., Decker, T., Eusebi, V., Fox, S.B., Ichihara, S., Jacquemier, J., Lakhani, S.R., Palacios, J., Rakha, E.A., Richardson, A.L., Schmitt, F.C., Tan, P., Tse, G.M., Weigelt, B., Ellis, I.O. & Reis-Filho, J. 2011, "Basal-like and triple-negative breast cancers: a critical review with an emphasis on the implications for pathologists and oncologists", *Mod. Pathol.,* vol. 24, no. 2, pp. 157-167.

Barbie, D.A., Tamayo, P., Boehm, J.S., Kim, S.Y., Moody, S.E., Dunn, I.F., Schinzel, A.C., Sandy, P., Meylan, E., Scholl, C., Fröhling, S., Chan, E.M., Sos, M.L., Michel, K., Mermel, C., Silver, S.J., Weir, B.A., Reiling, J.H., Sheng, Q., Gupta, P.B., Wadlow, R.C., Le, H., Hoersch, S., Wittner, B.S., Ramaswamy, S., Livingston, D.M., Sabatini, D.M., Meyerson, M., Thomas, R.K., Lander, E.S., Mesirov, J.P., Root, D.E., Gilliland, D.G., Jacks, T. & Hahn, W.C. 2009, "Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1", *Nature,* vol. 462, no. 7269, pp. 108-112.

Baselga, J. 2011, "Targeting the phosphoinositide-3 (PI3) kinase pathway in breast cancer", *Oncologist,* vol. 16 Suppl 1, pp. 12-19.

Berry, M.P.R., Graham, C.M., McNab, F.W., Xu, Z., Bloch, S.A.A., Oni, T., Wilkinson, K.A., Banchereau, R., Skinner, J., Wilkinson, R.J., Quinn, C., Blankenship, D., Dhawan, R., Cush, J.J., Mejias, A., Ramilo, O., Kon, O.M., Pascual, V., Banchereau, J., Chaussabel, D. & O'Garra, A. 2010, "An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis", *Nature,* vol. 466, no. 7309, pp. 973-977.

Best, D.J. & Roberts, D.E. 1975, "Algorithm AS 89: The Upper Tail Probabilities of Spearman's Rho", *J. R. Stat. Soc. Ser. C Appl. Stat.,* vol. 24, no. 3, pp. 377-379.

Bild, A.H., Yao, G., Chang, J.T., Wang, Q., Potti, A., Chasse, D., Joshi, M., Harpole, D., Lancaster, J.M., Berchuck, A., Olson,John A.,,Jr, Marks, J.R., Dressman, H.K., West, M. & Nevins, J.R. 2006, "Oncogenic pathway signatures in human cancers as a guide to targeted therapies", *Nature,* vol. 439, no. 7074, pp. 353-357.

Blankley, S., Graham, C.M., Levin, J., Turner, J., Berry, M.P.R., Bloom, C.I., Xu, Z., Pascual, V., Banchereau, J., Chaussabel, D., Breen, R., Santis, G., Blankenship, D.M., Lipman, M. & O'Garra, A. 2016, "A 380-gene meta-signature of active tuberculosis compared with healthy controls", *Eur. Respir. J.,* vol. 47, no. 6, pp. 1873-1876.

Bloom, C.I., Graham, C.M., Berry, M.P.R., Rozakeas, F., Redford, P.S., Wang, Y., Xu, Z., Wilkinson, K.A., Wilkinson, R.J., Kendrick, Y., Devouassoux, G., Ferry, T., Miyara, M., Bouvry, D., Valeyre, D., Gorochov, G., Blankenship, D., Saadatian, M., Vanhems, P., Beynon, H., Vancheeswaran, R., Wickremasinghe, M., Chaussabel, D., Banchereau, J., Pascual, V., Ho, L., Lipman, M. & O'Garra, A. 2013, "Transcriptional blood signatures distinguish pulmonary tuberculosis, pulmonary sarcoidosis, pneumonias and lung cancers", *PLoS One,* vol. 8, no. 8.

Booy, E.P., Henson, E.S. & Gibson, S.B. 2011, "Epidermal growth factor regulates Mcl-1 expression through the MAPK-Elk-1 signalling pathway contributing to cell survival in breast cancer", *Oncogene,* vol. 30, no. 20, pp. 2367-2378.

Boucher, M.J., Morisset, J., Vachon, P.H., Reed, J.C., Lainé, J. & Rivard, N. 2000, "MEK/ERK signaling pathway regulates the expression of Bcl-2, Bcl-X(L), and Mcl-1 and promotes survival of human pancreatic cancer cells", *J. Cell. Biochem.,* vol. 79, no. 3, pp. 355-369.

Brady, S.W., McQuerry, J.A., Qiao, Y., Piccolo, S.R., Shrestha, G., Jenkins, D.F., Layer, R.M., Pedersen, B.S., Miller, R.H., Esch, A., Selitsky, S.R., Parker, J.S., Anderson, L.A., Dalley, B.K., Factor, R.E., Reddy, C.B., Boltax, J.P., Li, D.Y., Moos, P.J., Gray, J.W., Heiser, L.M., Buys, S.S., Cohen, A.L., Johnson, W.E., Quinlan, A.R., Marth, G., Werner, T.L. & Bild, A.H. 2017, "Combating subclonal evolution of resistant cancer phenotypes", *Nat. Commun.,* vol. 8, no. 1, pp. 883.

Brennecke, P., Anders, S., Kim, J.K., Kolodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C. & Heisler, M.G. 2013, "Accounting for technical noise in single-cell RNA-seq experiments", *Nat. Methods,* vol. 10, no. 11, pp. 1093-1095.

Busby, M.A., Stewart, C., Miller, C.A., Grzeda, K.R. & Marth, G.T. 2013, "Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression", *Bioinformatics,* vol. 29, no. 5, pp. 656-657.

Cancer Genome, A.N. 2012, "Comprehensive molecular portraits of human breast tumours", *Nature,* vol. 490, no. 7418, pp. 61-70.

Carpenter, R.L. & Lo, H. 2013, "Regulation of Apoptosis by HER2 in Breast Cancer", *J. Carcinog. Mutagen,* vol. 2013.

Chang, W., Cheng, J., Allaire, J.J., Xie, Y. & McPherson, J. 2017, *, shiny: Web Application Framework for R* [Homepage of Comprehensive R Archive Network (CRAN)], [Online]. Available: 1.0.5.

Cheang, M.C.U., Martin, M., Nielsen, T.O., Prat, A., Voduc, D., Rodriguez-Lescure, A., Ruiz, A., Chia, S., Shepherd, L., Ruiz-Borrego, M., Calvo, L., Alba, E., Carrasco, E., Caballero, R., Tu, D., Pritchard, K.I., Levine, M.N., Bramwell, V.H., Parker, J., Bernard, P.S., Ellis, M.J., Perou, C.M., Di Leo, A. & Carey, L.A. 2015, "Defining breast cancer intrinsic subtypes by quantitative receptor expression", *Oncologist,* vol. 20, no. 5, pp. 474-482.

Cohen, A.L., Soldi, R., Zhang, H., Gustafson, A.M., Wilcox, R., Welm, B.E., Chang, J.T., Johnson, E., Spira, A., Jeffrey, S.S. & Bild, A.H. 2011, "A pharmacogenomic method for individualized prediction of drug sensitivity", *Mol. Syst. Biol.,* vol. 7, pp. 513.

Corbit, K.C., Trakul, N., Eves, E.M., Diaz, B., Marshall, M. & Rosner, M.R. 2003, "Activation of Raf-1 signaling by protein kinase C through a mechanism involving Raf kinase inhibitory protein", *J. Biol. Chem.,* vol. 278, no. 15, pp. 13061-13068.

Czabotar, P.E., Lessene, G., Strasser, A. & Adams, J.M. 2014, "Control of apoptosis by the BCL-2 protein family: implications for physiology and therapy", *Nat. Rev. Mol. Cell Biol.,* vol. 15, no. 1, pp. 49-63.

Daemen, A., Griffith, O.L., Heiser, L.M., Wang, N.J., Enache, O.M., Sanborn, Z., Pepin, F., Durinck, S., Korkola, J.E., Griffith, M., Hur, J.S., Huh, N., Chung, J., Cope, L., Fackler, M.J., Umbricht, C., Sukumar, S., Seth, P., Sukhatme, V.P., Jakkula, L.R., Lu, Y., Mills, G.B., Cho, R.J., Collisson, E.A., van't Veer, L.,J., Spellman, P.T. & Gray, J.W. 2013, "Modeling precision treatment of breast cancer", *Genome Biol.,* vol. 14, no. 10, pp. R110.

Datta, S.R., Dudek, H., Tao, X., Masters, S., Fu, H., Gotoh, Y. & Greenberg, M.E. 1997, "Akt phosphorylation of BAD couples survival signals to the cell-intrinsic death machinery", *Cell,* vol. 91, no. 2, pp. 231-241.

Davis, N.M., Sokolosky, M., Stadelman, K., Abrams, S.L., Libra, M., Candido, S., Nicoletti, F., Polesel, J., Maestro, R., D'Assoro, A., Drobot, L., Rakus, D., Gizak, A., Laidler, P., Dulińska-Litewka, J., Basecke, J., Mijatovic, S., Maksimovic-Ivanic, D., Montalto, G., Cervello, M., Fitzgerald, T.L., Demidenko, Z., Martelli, A.M., Cocco, L., Steelman, L.S. & McCubrey, J.A. 2014, "Deregulation of the EGFR/PI3K/PTEN/Akt/mTORC1 pathway in breast cancer: possibilities for therapeutic intervention", *Oncotarget,* vol. 5, no. 13, pp. 4603-4650.

De Abreu, F.B., Schwartz, G.N., Wells, W.A. & Tsongalis, G.J. 2014, "Personalized therapy for breast cancer", *Clin. Genet.,* vol. 86, no. 1, pp. 62-67.

Deng, J., Shimamura, T., Perera, S., Carlson, N.E., Cai, D., Shapiro, G.I., Wong, K. & Letai, A. 2007, "Proapoptotic BH3-only BCL-2 family protein BIM connects death signaling from epidermal growth factor receptor inhibition to the mitochondrion", *Cancer Res.,* vol. 67, no. 24, pp. 11867-11875.

DeSantis, C.E., Lin, C.C., Mariotto, A.B., Siegel, R.L., Stein, K.D., Kramer, J.L., Alteri, R., Robbins, A.S. & Jemal, A. 2014, "Cancer treatment and survivorship statistics, 2014", *CA Cancer J. Clin.,* vol. 64, no. 4, pp. 252-271.

Dorman, S.E. & Chaisson, R.E. 2007, "From magic bullets back to the magic mountain: the rise of extensively drug-resistant tuberculosis", *Nat. Med.,* vol. 13, no. 3, pp. 295-298.

Dvorkin-Gheva, A. & Hassell, J.A. 2014, "Identification of a novel luminal molecular subtype of breast cancer", *PLoS One,* vol. 9, no. 7.

El-Chaar, N., Piccolo, S.R., Boucher, K.M., Cohen, A.L., Chang, J.T., Moos, P.J. & Bild, A.H. 2014, "Genomic classification of the RAS network identifies a personalized treatment strategy for lung cancer", *Mol. Oncol.,* vol. 8, no. 7, pp. 1339-1354.

Esmail, H., Lai, R.P., Lesosky, M., Wilkinson, K.A., Graham, C.M., Horswell, S., Coussens, A.K., Barry,Clifton E.,,3rd, O'Garra, A. & Wilkinson, R.J. 2018, "Complement pathway gene activation and rising circulating immune complexes characterize early disease in HIV-associated tuberculosis", *Proc. Natl. Acad. Sci. U.S.A.,* vol. 115, no. 5.

Ewels, P., Magnusson, M., Lundin, S. & Käller, M. 2016, "MultiQC: summarize analysis results for multiple tools and samples in a single report", *Bioinformatics,* vol. 32, no. 19, pp. 3047-3048.

Faber, A.C., Corcoran, R.B., Ebi, H., Sequist, L.V., Waltman, B.A., Chung, E., Incio, J., Digumarthy, S.R., Pollack, S.F., Song, Y., Muzikansky, A., Lifshits, E., Roberge, S., Coffman, E.J., Benes, C.H., Gómez, H.,L., Baselga, J., Arteaga, C.L., Rivera, M.N., Dias-Santagata, D., Jain, R.K. & Engelman, J.A. 2011, "BIM expression in treatment-naive cancers predicts responsiveness to kinase inhibitors", *Cancer Discov.,* vol. 1, no. 4, pp. 352-365.

Faber, A.C., Li, D., Song, Y., Liang, M., Yeap, B.Y., Bronson, R.T., Lifshits, E., Chen, Z., Maira, S., García-Echeverría, C., Wong, K. & Engelman, J.A. 2009, "Differential induction of apoptosis in HER2 and EGFR addicted cancers following PI3K inhibition", *Proc. Natl. Acad. Sci. U.S.A.,* vol. 106, no. 46, pp. 19503-19508.

Fan, J., Salathia, N., Liu, R., Kaeser, G.E., Yung, Y.C., Herman, J.L., Kaper, F., Fan, J., Zhang, K., Chun, J. & Kharchenko, P.V. 2016, "Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis", *Nat. Methods,* vol. 13, no. 3, pp. 241-244.

Farabaugh, S.M., Boone, D.N. & Lee, A.V. 2015, "Role of IGF1R in Breast Cancer Subtypes, Stemness, and Lineage Differentiation", *Front. Endocrinol.,* vol. 6, pp. 59.

Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., Macgrogan, G., Bergh, J., Cameron, D., Goldstein, D., Duss, S., Nicoulaz, A., Brisken, C., Fiche, M., Delorenzi, M. & Iggo, R. 2005, "Identification of molecular apocrine breast tumours by microarray analysis", *Oncogene,* vol. 24, no. 29, pp. 4660-4671.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., Linsley, P.S. & Gottardo, R. 2015, "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data", *Genome Biol.,* vol. 16, pp. 278.

Franke, T.F., Hornik, C.P., Segev, L., Shostak, G.A. & Sugimoto, C. 2003, "PI3K/Akt and apoptosis: size matters", *Oncogene,* vol. 22, no. 56, pp. 8983-8998.

Fulda, S. & Debatin, K. 2006, "Extrinsic versus intrinsic apoptosis pathways in anticancer chemotherapy", *Oncogene,* vol. 25, no. 34, pp. 4798-4811.

Goard, C.A. & Schimmer, A.D. 2013, "An evidence-based review of obatoclax mesylate in the treatment of hematological malignancies", *Core Evid.,* vol. 8, pp. 15-26.

Groenendijk, F.H. & Bernards, R. 2014, "Drug resistance to targeted therapies: déjà vu all over again", *Mol. Oncol.,* vol. 8, no. 6, pp. 1067-1083.

Gu, Z., Eils, R. & Schlesner, M. 2016, "Complex heatmaps reveal patterns and correlations in multidimensional genomic data", *Bioinformatics,* vol. 32, no. 18, pp. 2847-2849.

Guedj, M., Marisa, L., de Reynies, A., Orsetti, B., Schiappa, R., Bibeau, F., MacGrogan, G., Lerebours, F., Finetti, P., Longy, M., Bertheau, P., Bertrand, F., Bonnet, F., Martin, A.L., Feugeas, J.P., Bièche, I., Lehmann-Che, J., Lidereau, R., Birnbaum, D., Bertucci, F., de Thé, H. & Theillet, C. 2012, "A refined molecular taxonomy of breast cancer", *Oncogene,* vol. 31, no. 9, pp. 1196-1206.

Guo, Y., Zhao, S., Li, C., Sheng, Q. & Shyr, Y. 2014, "RNAseqPS: A Web Tool for Estimating Sample Size and Power for RNAseq Experiment", *Cancer Inform.,* vol. 13, pp. 1-5.

Gustafson, A.M., Soldi, R., Anderlind, C., Scholand, M.B., Qian, J., Zhang, X., Cooper, K., Walker, D., McWilliams, A., Liu, G., Szabo, E., Brody, J., Massion, P.P., Lenburg, M.E., Lam, S., Bild, A.H. & Spira, A. 2010, "Airway PI3K pathway activation is an early and reversible event in lung cancer development", *Sci. Transl. Med.,* vol. 2, no. 26.

Hammond, M.E., Hayes, D.F., Dowsett, M., Allred, D.C., Hagerty, K.L., Badve, S., Fitzgibbons, P.L., Francis, G., Goldstein, N.S., Hayes, M., Hicks, D.G., Lester, S., Love, R., Mangu, P.B., McShane, L., Miller, K., Osborne, C.K., Paik, S., Perlmutter, J., Rhodes, A., Sasano, H., Schwartz, J.N., Sweep, F.C.G., Taube, S., Torlakovic, E.E., Valenstein, P., Viale, G., Visscher, D., Wheeler, T., Williams, R.B., Wittliff, J.L. & Wolff, A.C. 2010, "American Society of Clinical Oncology/College Of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer", *J. Clin. Oncol.,* vol. 28, no. 16, pp. 2784-2795.

Hanahan, D. & Weinberg, R.A. 2011, "Hallmarks of cancer: the next generation", *Cell,* vol. 144, no. 5, pp. 646-674.

Hänzelmann, S., Castelo, R. & Guinney, J. 2013, "GSVA: gene set variation analysis for microarray and RNA-seq data", *BMC Bioinformatics,* vol. 14, pp. 7.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K.J., Rozenblatt-Rosen, O., Dor, Y., Regev, A. & Yanai, I. 2016, "CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq", *Genome Biol.,* vol. 17, no. 1, pp. 77.

Hassan, M., Watari, H., AbuAlmaaty, A., Ohba, Y. & Sakuragi, N. 2014, "Apoptosis and molecular targeting therapy in cancer", *Biomed Res. Int.,* vol. 2014, pp. 150845.

Hennessy, B.T., Lu, Y., Gonzalez-Angulo, A., Carey, M.S., Myhre, S., Ju, Z., Davies, M.A., Liu, W., Coombes, K., Meric-Bernstam, F., Bedrosian, I., McGahren, M., Agarwal, R., Zhang, F., Overgaard, J., Alsner, J., Neve, R.M., Kuo, W., Gray, J.W., Borresen-Dale, A. & Mills, G.B. 2010, "A Technical Assessment of the Utility of

Reverse Phase Protein Arrays for the Study of the Functional Proteome in Non-microdissected Human Breast Cancers", *Clin. Proteomics,* vol. 6, no. 4, pp. 129-151.

Herschkowitz, J.I., Simin, K., Weigman, V.J., Mikaelian, I., Usary, J., Hu, Z., Rasmussen, K.E., Jones, L.P., Assefnia, S., Chandrasekharan, S., Backlund, M.G., Yin, Y., Khramtsov, A.I., Bastein, R., Quackenbush, J., Glazer, R.I., Brown, P.H., Green, J.E., Kopelovich, L., Furth, P.A., Palazzo, J.P., Olopade, O.I., Bernard, P.S., Churchill, G.A., Van Dyke, T. & Perou, C.M. 2007, "Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors", *Genome Biol.,* vol. 8, no. 5, pp. R76.

Hicks, S.C., Townes, F.W., Teng, M. & Irizarry, R.A. 2017, "Missing data and technical variability in single-cell RNA-sequencing experiments", *Biostatistics,* vol. 19, no. 4.

Hollander, M., Wolfe, D.A. & Chicken, E. 2013, *Nonparametric Statistical Methods,* John Wiley & Sons.

Hotelling, H. 1933, "Analysis of a complex of statistical variables into principal components", *J. Educ. Psychol.,* vol. 24, no. 6, pp. 417-441.

Huang, C., Tu, S., Lien, H., Jeng, J., Liu, J., Huang, C., Wu, Y., Liu, C., Lai, L. & Chuang, E.Y. 2012, "Prediction consistency and clinical presentations of breast cancer molecular subtypes for Han Chinese population", *J. Transl. Med.,* vol. 10 Suppl 1, pp. S10.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B.S., Bravo, H.C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K.D., Irizarry, R.A., Lawrence, M., Love, M.I., MacDonald, J., Obenchain, V., Oleś, A.,K., Pagès, H., Reyes, A., Shannon, P., Smyth, G.K., Tenenbaum, D., Waldron, L. & Morgan, M. 2015, "Orchestrating high-throughput genomic analysis with Bioconductor", *Nat. Methods,* vol. 12, no. 2, pp. 115-121.

Hynes, N.E. 2000, "Tyrosine kinase signalling in breast cancer", *Breast Cancer Res.,* vol. 2, no. 3, pp. 154-157.

Ian Freshney, R. & Freshney, M.G. 2004, *Culture of Epithelial Cells,* John Wiley & Sons.

Iqbal, N. & Iqbal, N. 2014, "Human Epidermal Growth Factor Receptor 2 (HER2) in Cancers: Overexpression and Therapeutic Implications", *Mol. Biol. Int.,* vol. 2014, pp. 852748.

Jacobsen, M., Repsilber, D., Gutschmidt, A., Neher, A., Feldmann, K., Mollenkopf, H.J., Ziegler, A. & Kaufmann, S.H.E. 2007, "Candidate biomarkers for discrimination between infection and disease caused by Mycobacterium tuberculosis", *J. Mol. Med.,* vol. 85, no. 6, pp. 613-621.

Johnson, W.E., Li, C. & Rabinovic, A. 2007, "Adjusting batch effects in microarray expression data using empirical Bayes methods", *Biostatistics,* vol. 8, no. 1, pp. 118-127.

Kaforou, M., Wright, V.J., Oni, T., French, N., Anderson, S.T., Bangani, N., Banwell, C.M., Brent, A.J., Crampin, A.C., Dockrell, H.M., Eley, B., Heyderman, R.S., Hibberd, M.L., Kern, F., Langford, P.R., Ling, L., Mendelson, M., Ottenhoff, T.H., Zgambo, F., Wilkinson, R.J., Coin, L.J. & Levin, M. 2013, "Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: a case-control study", *PLoS Med.,* vol. 10, no. 10.

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D. & Kent, W.J. 2004, "The UCSC Table Browser data retrieval tool", *Nucleic Acids Res.,* vol. 32, pp. 493.

Kharchenko, P.V., Silberstein, L. & Scadden, D.T. 2014, "Bayesian approach to single-cell differential expression analysis", *Nat. Methods,* vol. 11, no. 7, pp. 740-742.

Kolch, W., Heidecker, G., Kochs, G., Hummel, R., Vahidi, H., Mischak, H., Finkenzeller, G., Marmé, D. & Rapp, U.R. 1993, "Protein kinase C alpha activates RAF-1 by direct phosphorylation", *Nature,* vol. 364, no. 6434, pp. 249-252.

Lee, S., Wu, L.S., Huang, G., Huang, K., Lee, T. & Weng, J.T. 2016, "Gene expression profiling identifies candidate biomarkers for active and latent tuberculosis", *BMC Bioinformatics,* vol. 17 Suppl 1, pp. 3.

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E. & Storey, J.D. 2012, "The sva package for removing batch effects and other unwanted variation in high-throughput experiments", *Bioinformatics,* vol. 28, no. 6, pp. 882-883.

Lemmon, M.A. & Schlessinger, J. 2010, "Cell signaling by receptor tyrosine kinases", *Cell,* vol. 141, no. 7, pp. 1117-1134.

Leong, S., Zhao, Y., Joseph, N.M., Hochberg, N.S., Sarkar, S., Pleskunas, J., Hom, D., Lakshminarayanan, S., Horsburgh, C.R., Roy, G., Ellner, J.J., Johnson, W.E. & Salgame, P. 2018, "Existing blood transcriptional classifiers accurately discriminate active tuberculosis from latent infection in individuals from south India", *Tuberculosis,* vol. 109, pp. 41-51.

Letai, A.G. 2008, "Diagnosing and exploiting cancer's addiction to blocks in apoptosis", *Nat. Rev. Cancer,* vol. 8, no. 2, pp. 121-132.

Ley, R., Balmanno, K., Hadfield, K., Weston, C. & Cook, S.J. 2003, "Activation of the ERK1/2 signaling pathway promotes phosphorylation and proteasome-dependent degradation of the BH3-only protein, Bim", *J. Biol. Chem.,* vol. 278, no. 21, pp. 18811-18816.

Liao, Y., Smyth, G.K. & Shi, W. 2014, "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features", *Bioinformatics,* vol. 30, no. 7, pp. 923-930.

Liao, Y., Smyth, G.K. & Shi, W. 2013, "The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote", *Nucleic Acids Res.,* vol. 41, no. 10.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P. & Mesirov, J.P. 2011, "Molecular signatures database (MSigDB) 3.0", *Bioinformatics,* vol. 27, no. 12, pp. 1739-1740.

Love, M.I., Huber, W. & Anders, S. 2014, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2", *Genome Biol.,* vol. 15, no. 12, pp. 550.

Lun, A.T.L., Pagès, H. & Smith, M.L. 2018, "beachmat: A Bioconductor C++ API for accessing high-throughput biological data from a variety of R matrix types", *PLoS Comput. Biol.,* vol. 14, no. 5.

Luo, J., Deng, Z., Luo, X., Tang, N., Song, W., Chen, J., Sharff, K.A., Luu, H.H., Haydon, R.C., Kinzler, K.W., Vogelstein, B. & He, T. 2007, "A protocol for rapid generation of recombinant adenoviruses using the AdEasy system", *Nat. Protoc.,* vol. 2, no. 5, pp. 1236-1247.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., Trombetta, J.J., Weitz, D.A., Sanes, J.R., Shalek, A.K., Regev, A. & McCarroll, S.A. 2015, "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets", *Cell,* vol. 161, no. 5, pp. 1202-1214.

Maertzdorf, J., McEwen, G., Weiner, J.,3rd, Tian, S., Lader, E., Schriek, U., Mayanja-Kizza, H., Ota, M., Kenneth, J. & Kaufmann, S.H. 2016, "Concise gene signature for point-of-care classification of tuberculosis", *EMBO Mol. Med.,* vol. 8, no. 2, pp. 86-95.

Maertzdorf, J., Weiner, J.,3rd, Mollenkopf, H., Network, T., Bauer, T., Prasse, A., Müller-Quernheim, J. & Kaufmann, S.H.E. 2012, "Common patterns and disease-related signatures in tuberculosis and sarcoidosis", *Proc. Natl. Acad. Sci. U.S.A.,* vol. 109, no. 20, pp. 7853-7858.

Marusyk, A. & Polyak, K. 2010, "Tumor heterogeneity: causes and consequences", *Biochim. Biophys. Acta,* vol. 1805, no. 1, pp. 105-117.

Masuda, H., Zhang, D., Bartholomeusz, C., Doihara, H., Hortobagyi, G.N. & Ueno, N.T. 2012, "Role of epidermal growth factor receptor in breast cancer", *Breast Cancer Res. Treat.,* vol. 136, no. 2, pp. 331-345.

Matallanas, D., Birtwistle, M., Romano, D., Zebisch, A., Rauch, J., von Kriegsheim, A. & Kolch, W. 2011, "Raf family kinases: old dogs have learned new tricks", *Genes Cancer,* vol. 2, no. 3, pp. 232-260.

McCarthy, D.J., Campbell, K.R., Lun, A.T.L. & Wills, Q.F. 2017, "Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R", *Bioinformatics,* vol. 33, no. 8, pp. 1179-1186.

McCubrey, J.A., Steelman, L.S., Chappell, W.H., Abrams, S.L., Franklin, R.A., Montalto, G., Cervello, M., Libra, M., Candido, S., Malaponte, G., Mazzarino, M.C., Fagone, P., Nicoletti, F., Bäsecke, J., Mijatovic, S., Maksimovic-Ivanic, D., Milella, M., Tafuri, A., Chiarini, F., Evangelisti, C., Cocco, L. & Martelli, A.M. 2012, "Ras/Raf/MEK/ERK and PI3K/PTEN/Akt/mTOR cascade inhibitors: how mutations can result in therapy resistance and how to overcome resistance", *Oncotarget,* vol. 3, no. 10, pp. 1068-1111.

McCubrey, J.A., Steelman, L.S., Chappell, W.H., Abrams, S.L., Wong, E.W.T., Chang, F., Lehmann, B., Terrian, D.M., Milella, M., Tafuri, A., Stivala, F., Libra, M., Basecke, J., Evangelisti, C., Martelli, A.M. & Franklin, R.A. 2007, "Roles of the Raf/MEK/ERK pathway in cell growth, malignant transformation and drug resistance", *Biochim. Biophys. Acta,* vol. 1773, no. 8, pp. 1263-1284.

Mendoza, M.C., Er, E.E. & Blenis, J. 2011, "The Ras-ERK and PI3K-mTOR pathways: cross-talk and compensation", *Trends Biochem. Sci.,* vol. 36, no. 6, pp. 320-328.

Montero, J., Sarosiek, K.A., DeAngelo, J.D., Maertens, O., Ryan, J., Ercan, D., Piao, H., Horowitz, N.S., Berkowitz, R.S., Matulonis, U., Jänne, P.,A., Amrein, P.C., Cichowski, K., Drapkin, R. & Letai, A. 2015, "Drug-induced death signaling strategy rapidly predicts cancer response to chemotherapy", *Cell,* vol. 160, no. 5, pp. 977-989.

Mosesson, Y. & Yarden, Y. 2004, "Oncogenic growth factor receptors: implications for signal transduction therapy", *Semin. Cancer Biol.,* vol. 14, no. 4, pp. 262-270.

Nahta, R., Hortobágyi, G.,N. & Esteva, F.J. 2003, "Growth factor receptors in breast cancer: potential for therapeutic intervention", *Oncologist,* vol. 8, no. 1, pp. 5-17.

Nakamura, T., Yabuta, Y., Okamoto, I., Aramaki, S., Yokobayashi, S., Kurimoto, K., Sekiguchi, K., Nakagawa, M., Yamamoto, T. & Saitou, M. 2015, "SC3-seq: a method for highly parallel and quantitative measurement of single-cell gene expression", *Nucleic Acids Res.,* vol. 43, no. 9.

Nalluri, S., Peirce, S.K., Tanos, R., Abdella, H.A., Karmali, D., Hogarty, M.D. & Goldsmith, K.C. 2015, "EGFR signaling defines Mcl^{-}1 survival dependency in neuroblastoma", *Cancer Biol. Ther.,* vol. 16, no. 2, pp. 276-286.

Paplomata, E. & O'Regan, R. 2014, "The PI3K/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers", *Ther. Adv. Med. Oncol.,* vol. 6, no. 4, pp. 154-166.

Parker, J.S., Mullins, M., Cheang, M.C.U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J.F., Stijleman, I.J., Palazzo, J., Marron, J.S., Nobel, A.B., Mardis, E., Nielsen, T.O., Ellis, M.J., Perou, C.M. & Bernard, P.S. 2009, "Supervised risk predictor of breast cancer based on intrinsic subtypes", *J. Clin. Oncol.,* vol. 27, no. 8, pp. 1160-1167.

Patani, N., Martin, L. & Dowsett, M. 2013, "Biomarkers for the clinical management of breast cancer: international perspective", *Int. J. Cancer,* vol. 133, no. 1, pp. 1-13.

Paweletz, C.P., Charboneau, L., Bichsel, V.E., Simone, N.L., Chen, T., Gillespie, J.W., Emmert-Buck, M., Roth, M.J., Petricoin,E F,,III & Liotta, L.A. 2001, "Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front", *Oncogene,* vol. 20, no. 16, pp. 1981-1989.

Pearson, K. 1901, "LIII. On lines and planes of closest fit to systems of points in space", *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science,* vol. 2, no. 11, pp. 559-572.

Perona, R. 2006, "Cell signalling: growth factors and tyrosine kinase receptors", *Clin. Transl. Oncol.,* vol. 8, no. 2, pp. 77-82.

Perou, C.M., Sørlie, T., Eisen, M.B., van, d.R., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S.X., Lønning, P.E., Børresen-Dale, A.L., Brown, P.O. & Botstein, D. 2000, "Molecular portraits of human breast tumours", *Nature,* vol. 406, no. 6797, pp. 747-752.

Perou, C.M. 2010, "Molecular stratification of triple-negative breast cancers", *Oncologist,* vol. 15 Suppl 5, pp. 39-48.

Picelli, S., Björklund, Å,K., Faridani, O.R., Sagasser, S., Winberg, G. & Sandberg, R. 2013, "Smart-seq2 for sensitive full-length transcriptome profiling in single cells", *Nat. Methods,* vol. 10, no. 11, pp. 1096-1098.

Prat, A., Parker, J.S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J.I., He, X. & Perou, C.M. 2010, "Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer", *Breast Cancer Res.,* vol. 12, no. 5, pp. R68.

Rahman, M., MacNeil, S.M., Jenkins, D.F., Shrestha, G., Wyatt, S.R., McQuerry, J.A., Piccolo, S.R., Heiser, L.M., Gray, J.W., Johnson, W.E. & Bild, A.H. 2017, "Activity of distinct growth factor receptor network components in breast tumors uncovers two biologically relevant subtypes", *Genome Med.,* vol. 9, no. 1, pp. 40.

Ricci, M.S. & Zong, W. 2006, "Chemotherapeutic approaches for targeting cell death pathways", *Oncologist,* vol. 11, no. 4, pp. 342-357.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. & Smyth, G.K. 2015, "limma powers differential expression analyses for RNA-sequencing and microarray studies", *Nucleic Acids Res.,* vol. 43, no. 7.

Roberts, P.J. & Der, C.J. 2007, "Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer", *Oncogene,* vol. 26, no. 22, pp. 3291-3310.

Sachs, M. 2017, "plotROC: A Tool for Plotting ROC Curves", *Journal of Statistical Software, Code Snippets,* vol. 79, no. 2, pp. 1-19.

Saini, K.S., Loi, S., de Azambuja, E., Metzger-Filho, O., Saini, M.L., Ignatiadis, M., Dancey, J.E. & Piccart-Gebhart, M. 2013, "Targeting the PI3K/AKT/mTOR and Raf/MEK/ERK pathways in the treatment of breast cancer", *Cancer Treat. Rev.,* vol. 39, no. 8, pp. 935-946.

Sambarey, A., Devaprasad, A., Mohan, A., Ahmed, A., Nayak, S., Swaminathan, S., D'Souza, G., Jesuraj, A., Dhar, C., Babu, S., Vyakarnam, A. & Chandra, N. 2017, "Unbiased Identification of Blood-based Biomarkers for Pulmonary Tuberculosis by Modeling and Mining Molecular Interaction Networks", *EBioMedicine,* vol. 15, pp. 112-126.

Santen, R.J., Song, R.X., McPherson, R., Kumar, R., Adam, L., Jeng, M. & Yue, W. 2002, "The role of mitogen-activated protein (MAP) kinase in breast cancer", *J. Steroid Biochem. Mol. Biol.,* vol. 80, no. 2, pp. 239-256.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. 2015, "Spatial reconstruction of single-cell gene expression data", *Nat. Biotechnol.,* vol. 33, no. 5, pp. 495-502.

Schlessinger, J. 2000, "Cell signaling by receptor tyrosine kinases", *Cell,* vol. 103, no. 2, pp. 211-225.

Scriba, T.J., Penn-Nicholson, A., Shankar, S., Hraha, T., Thompson, E.G., Sterling, D., Nemes, E., Darboe, F., Suliman, S., Amon, L.M., Mahomed, H., Erasmus, M., Whatney, W., Johnson, J.L., Boom, W.H., Hatherill, M., Valvo, J., De Groote, M.A., Ochsner, U.A., Aderem, A., Hanekom, W.A., Zak, D.E. & other members of the ACS cohort, study team 2017, "Sequential inflammatory processes define human progression from M. tuberculosis infection to tuberculosis disease", *PLoS Pathog.,* vol. 13, no. 11.

Shen, Y., Rahman, M., Piccolo, S.R., Gusenleitner, D., El-Chaar, N., Cheng, L., Monti, S., Bild, A.H. & Johnson, W.E. 2015, "ASSIGN: context-specific genomic profiling of multiple heterogeneous biological pathways", *Bioinformatics,* vol. 31, no. 11, pp. 1745-1753.

Singhania, A., Verma, R., Graham, C.M., Lee, J., Tran, T., Richardson, M., Lecine, P., Leissner, P., Berry, M.P.R., Wilkinson, R.J., Kaiser, K., Rodrigue, M., Woltmann, G., Haldar, P. & O'Garra, A. 2018, "A modular transcriptional signature identifies phenotypic heterogeneity of human tuberculosis infection", *Nat. Commun.,* vol. 9, no. 1, pp. 2308.

Sloot, R., Schim van der Loeff, Maarten,F., van Zwet, E.,W., Haks, M.C., Keizer, S.T., Scholing, M., Ottenhoff, T.H.M., Borgdorff, M.W. & Joosten, S.A. 2015, "Biomarkers Can Identify Pulmonary Tuberculosis in HIV-infected Drug Users Months Prior to Clinical Diagnosis", *EBioMedicine,* vol. 2, no. 2, pp. 172-179.

Soldi, R., Cohen, A.L., Cheng, L., Sun, Y., Moos, P.J. & Bild, A.H. 2013, "A genomic approach to predict synergistic combinations for breast cancer treatment", *Pharmacogenomics J.,* vol. 13, no. 1, pp. 94-104.

Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van, d.R., Jeffrey, S.S., Thorsen, T., Quist, H., Matese, J.C., Brown, P.O., Botstein, D., Lønning, P.E. & Børresen-Dale, A.L. 2001, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications", *Proc. Natl. Acad. Sci. U.S.A.,* vol. 98, no. 19, pp. 10869-10874.

Sotiriou, C., Neo, S., McShane, L.M., Korn, E.L., Long, P.M., Jazaeri, A., Martiat, P., Fox, S.B., Harris, A.L. & Liu, E.T. 2003, "Breast cancer classification and prognosis based on gene expression profiles from a population-based study", *Proc. Natl. Acad. Sci. U.S.A.,* vol. 100, no. 18, pp. 10393-10398.

Spitz, F. & Furlong, E.E.M. 2012, "Transcription factors: from enhancer binding to developmental control", *Nat. Rev. Genet.,* vol. 13, no. 9, pp. 613-626.

Stegle, O., Teichmann, S.A. & Marioni, J.C. 2015, "Computational and analytical challenges in single-cell transcriptomics", *Nat. Rev. Genet.,* vol. 16, no. 3, pp. 133-145.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. & Mesirov, J.P. 2005, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles", *Proc. Natl. Acad. Sci. U.S.A.,* vol. 102, no. 43, pp. 15545-15550.

Suliman, S., Thompson, E., Sutherland, J., Weiner Rd, J., Ota, M.O.C., Shankar, S., Penn-Nicholson, A., Thiel, B., Erasmus, M., Maertzdorf, J., Duffy, F.J., Hill, P.C., Hughes, E.J., Stanley, K., Downing, K., Fisher, M.L., Valvo, J., Parida, S.K., van, d.S., Tromp, G., Adetifa, I.M.O., Donkor, S., Howe, R., Mayanja-Kizza, H., Boom, W.H., Dockrell, H., Ottenhoff, T.H.M., Hatherill, M., Aderem, A., Hanekom, W.A., Scriba, T.J., Kaufmann, S.H., Zak, D.E., Walzl, G. & and the GC6-74 and ACS cohort, study groups 2018, "Four-gene Pan-African Blood Signature Predicts Progression to Tuberculosis", *Am. J. Respir. Crit. Care Med.,* vol. 197, no. 9, pp. 1198-1208.

Sweeney, T.E., Braviak, L., Tato, C.M. & Khatri, P. 2016, "Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis", *Lancet Respir. Med.,* vol. 4, no. 3, pp. 213-224.

Tang, P. & Tse, G.M. 2016, "Immunohistochemical Surrogates for Molecular Classification of Breast Carcinoma: A 2015 Update", *Arch. Pathol. Lab. Med.,* vol. 140, no. 8, pp. 806-814.

The International Agency for Research on Cancer 2018, *Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018*, WHO.

The R Core Team 2014, *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013*.

Thompson, E.G., Du, Y., Malherbe, S.T., Shankar, S., Braun, J., Valvo, J., Ronacher, K., Tromp, G., Tabb, D.L., Alland, D., Shenai, S., Via, L.E., Warwick, J., Aderem, A., Scriba, T.J., Winter, J., Walzl, G., Zak, D.E. & Catalysis TB–Biomarker Consortium 2017, "Host blood RNA signatures predict the outcome of tuberculosis treatment", *Tuberculosis,* vol. 107, pp. 48-58.

Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth,Marc H.,,2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J., Cohen, O., Shah, P., Lu, D., Genshaft, A.S., Hughes, T.K., Ziegler, C.G.K., Kazer, S.W., Gaillard, A., Kolb, K.E., Villani, A., Johannessen, C.M., Andreev, A.Y., Van Allen, E.,M., Bertagnolli, M., Sorger, P.K., Sullivan, R.J., Flaherty, K.T., Frederick, D.T., Jané-Valbuena, J., Yoon, C.H., Rozenblatt-Rosen, O., Shalek, A.K., Regev, A. & Garraway, L.A. 2016, "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq", *Science,* vol. 352, no. 6282, pp. 189-196.

Townsend, K.J., Trusty, J.L., Traupman, M.A., Eastman, A. & Craig, R.W. 1998, "Expression of the antiapoptotic MCL1 gene product is regulated by a mitogen activated protein kinase-mediated pathway triggered through microtubule disruption and protein kinase C", *Oncogene,* vol. 17, no. 10, pp. 1223-1234.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S. & Rinn, J.L. 2014, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells", *Nat. Biotechnol.,* vol. 32, no. 4, pp. 381-386.

Tung, P., Blischak, J.D., Hsiao, C.J., Knowles, D.A., Burnett, J.E., Pritchard, J.K. & Gilad, Y. 2017, "Batch effects and the effective design of single-cell gene expression studies", *Sci. Rep.,* vol. 7, pp. 39921.

Van, d.M. & Hinton, G. 2008, "Visualizing data using t-SNE", *J. Mach. Learn. Res.,* vol. 9, pp. 1532-4435.

Vera-Badillo, F., Templeton, A.J., de Gouveia, P., Diaz-Padilla, I., Bedard, P.L., Al-Mubarak, M., Seruga, B., Tannock, I.F., Ocana, A. & Amir, E. 2014, "Androgen receptor expression and outcomes in early breast cancer: a systematic review and meta-analysis", *J. Natl. Cancer Inst.,* vol. 106, no. 1.

Vo, T. & Letai, A. 2010, "BH3-only proteins and their effects on cancer", *Adv. Exp. Med. Biol.,* vol. 687, pp. 49-63.

Vogler, M. 2014, "Targeting BCL2-Proteins for the Treatment of Solid Tumours", *Adv. Med.,* vol. 2014, pp. 943648.

Walter, N.D., Miller, M.A., Vasquez, J., Weiner, M., Chapman, A., Engle, M., Higgins, M., Quinones, A.M., Rosselli, V., Canono, E., Yoon, C., Cattamanchi, A., Davis, J.L., Phang, T., Stearman, R.S., Datta, G., Garcia, B.J., Daley, C.L., Strong, M., Kechris, K., Fingerlin, T.E., Reves, R. & Geraci, M.W. 2016, "Blood Transcriptional Biomarkers for Active Tuberculosis among Patients in the United States: a Case-Control Study with Systematic Cross-Classifier Evaluation", *J. Clin. Microbiol.,* vol. 54, no. 2, pp. 274-282.

Watters, J.W. & Roberts, C.J. 2006, "Developing gene expression signatures of pathway deregulation in tumors", *Mol. Cancer Ther.,* vol. 5, no. 10, pp. 2444-2449.

Weigel, M.T. & Dowsett, M. 2010, "Current and emerging biomarkers in breast cancer: prognosis and prediction", *Endocr. Relat. Cancer,* vol. 17, no. 4, pp. 245.

Weston, C.R., Balmanno, K., Chalmers, C., Hadfield, K., Molton, S.A., Ley, R., Wagner, E.F. & Cook, S.J. 2003, "Activation of ERK1/2 by deltaRaf-1:ER* represses Bim expression independently of the JNK or PI3K pathways", *Oncogene,* vol. 22, no. 9, pp. 1281-1293.

Wickham, H. 2010, *ggplot2: elegant graphics for data analysis (Use R!),* Springer New York, NY.

Williams, M.M. & Cook, R.S. 2015, "Bcl-2 family proteins in breast development and cancer: could Mcl-1 targeting overcome therapeutic resistance?", *Oncotarget,* vol. 6, no. 6, pp. 3519-3530.

Wolff, A.C., Hammond, M.E., Hicks, D.G., Dowsett, M., McShane, L.M., Allison, K.H., Allred, D.C., Bartlett, J.M.S., Bilous, M., Fitzgibbons, P., Hanna, W., Jenkins, R.B., Mangu, P.B., Paik, S., Perez, E.A., Press, M.F., Spears, P.A., Vance, G.H., Viale, G., Hayes, D.F., American Society of, Clinical Oncology & College of, A.P. 2013, "Recommendations for human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists clinical practice guideline update", *J. Clin. Oncol.,* vol. 31, no. 31, pp. 3997-4013.

World Health Organization 2016, *Global tuberculosis report 2016*, WHO.

Wuillème-Toumi, S., Trichet, V., Gomez-Bougie, P., Gratas, C., Bataille, R. & Amiot, M. 2007, "Reciprocal protection of Mcl-1 and Bim from ubiquitin-proteasome degradation", *Biochem. Biophys. Res. Commun.,* vol. 361, no. 4, pp. 865-869.

Yarden, Y. & Sliwkowski, M.X. 2001, "Untangling the ErbB signalling network", *Nat. Rev. Mol. Cell Biol.,* vol. 2, no. 2, pp. 127-137.

Zak, D.E., Penn-Nicholson, A., Scriba, T.J., Thompson, E., Suliman, S., Amon, L.M., Mahomed, H., Erasmus, M., Whatney, W., Hussey, G.D., Abrahams, D., Kafaar, F., Hawkridge, T., Verver, S., Hughes, E.J., Ota, M., Sutherland, J., Howe, R., Dockrell, H.M., Boom, W.H., Thiel, B., Ottenhoff, T.H.M., Mayanja-Kizza, H., Crampin, A.C., Downing, K., Hatherill, M., Valvo, J., Shankar, S., Parida, S.K., Kaufmann, S.H.E., Walzl, G., Aderem, A., Hanekom, W.A. & ACS and GC6-74 cohort, study groups 2016, "A blood RNA signature for tuberculosis disease risk: a prospective cohort study", *Lancet,* vol. 387, no. 10035, pp. 2312-2322.

Zhang, W. & Liu, H.T. 2002, "MAPK signal pathways in the regulation of cell proliferation in mammalian cells", *Cell Res.,* vol. 12, no. 1, pp. 9-18.

CURRICULUM VITAE

[redacted]

[redacted]

[redacted]

[redacted]