Understand Biology Using Single Cell RNA-Sequencing

Hongxu Ding

Submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

ABSTRACT

Understand Biology Using Single Cell RNA-Sequencing

Hongxu Ding

This dissertation summarizes the development of experimental and analytical tools for single cell RNA sequencing (scRNA-Seq), including 1) scPLATE-Seq, a FACS- and plate-based scRNA-Seq platform, which is accurate, robust, fully automated and cost-efficient; 2) metaVIPER, an algorithm for transcriptional regulator activity inference based on scRNA-Seq profiles; and 3) iterClust, a statistical framework for iterative clustering analysis, especially suitable for dissecting hierarchy of heterogeneity among single cells. Further this dissertation summarizes biological questions answered by combining these tools, including 1) understanding inter- and intra-tumor heterogeneity of human glioblastoma; 2) elucidating regulators of β-cell de-differentiation in type-2 diabetes; and 3) developing novel therapeutics targeting cell-state regulators of breast cancer stem cells.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank members of Califano lab for their support for my PhD work, especially Dr. Mariano Alvarez and Dr. Andrea Califano for their valuable guidance of my academic career. I would like to thank all my collaborators, including Dr. Jeremy Worley, Ms. Erin Bush, Dr. Peter Sims for the scPLATE-Seq project; Dr. Eugene Douglass, Dr. Adam Sonabend, Dr. Angeliki Mela, Dr. Sayantan Bose, Dr. Christian Gonzalez, Dr. Peter D. Canoll and Dr. Peter A. Sims for metaVIPER project; Ms. Wanxin Wang for iterClust project; Dr. Jinsook Son and Dr. Domenico Accili for the T2D project; Dr. Jeremy Worley, Ms. Erin Bush, Dr. Beatrice Salvatori, Dr. Prabhjot Mundi, Dr. Daniel Diolaiti, Dr. Siu-Hong Ho, Dr. Andrew Kung, Dr. Peter Sims, and Dr. Piero Dalerba for the BCSC project.

Dedication

In memory of my uncle, Professor Dr. Kewen Liu, who encouraged me pursuing an academic career, and passed away during my PhD study.

Dedicated to my fiancée, Ms. Wanxin Wang, my parents, and the whole family.

Part I

Introduction

The “scientific question” that drives the development of the single cell RNA sequencing (scRNA-Seq) technique is the need to understand biological states at individual cells level, including 1) cell types, e.g. types of neurons in mouse retina [Macosco, 2015], 2) cell cycle stages, e.g. G1, S, G2/M stages of mouse embryonic stem cells [Liu, 2017], and 3) developmental lineage, e.g. myoblast-to-muscle developmental process [Trapnell, 2014].

For the technique per se, scRNA-Seq is supported by two major breakthroughs: 1) the ability of doing reverse transcription from tiny amount of mRNA, and 2) the ability of capturing single cells in a high-throughput manner. The first successful single cell reverse transcription protocol was developed by [Tang, 2009], where poly-A primers are used to capture mRNA. After 1st strand synthesis, the primers are removed, and then a poly-A sequence is aligned to the end of newly synthesized cDNA. Then a poly-T primer is used for the 2nd strand synthesis to get cDNA products. However, this protocol includes 1) multi-steps, and 2) primer cleanup, making it inconvenient to perform. The SMART-Seq technique [Ramsköld, 2011] simplified the protocol by using “template switching” technique. In SMART-Seq protocol, a specific reverse transcriptase is used, which can align a short sequence, know as template switching oligo, directly after the 1st strand synthesis. This template switching oligo can be used as handle for 2nd strand synthesis. Therefore we can add poly-A primer and template switching oligo together in the reaction system, greatly simplified the whole process. Since the starting mRNA is tiny, we need to do way more rounds of PCR compared to bulk sequencing to get a desirable sequencing library. So the next problem they community was facing is PCR bias, which means different genes might have different PCR amplification rate, thus greatly confounds the evaluation of gene expression. So [Islam, 2013] introduced unique molecular identifier (UMI), which barcodes

individual mRNA transcript. In the final sequencing result, reads with same UMIs will be collapsed as 1 transcript, therefore more accurate.

The major challenge for the single cell capturing platform is throughput. Dated back to 2009 original protocol [Tang, 2009], single cells were hand-picked. The later FACS- microfluidic- and microfluidics-based platforms greatly increase the throughput. For instance, FACS-based platforms are compatible with 96 or 384-well plates. Cell suspension is loaded on the FACS machine, which will sort single cells into plates according to the florescent labeling of cells. Most recent microfluidics-based Fluidigm C1 systems can deal with 800 cells/chip. Cell suspension is flowed through microfluidic channels, during which single cells can be captured by individual capture site alongside the channel. The microdroplets-based 10x Genomics Chromium can deal with thousands cells/run. Cell suspension is flowed through microfluidic channels, inside which single cells will be encapsulated by oil droplets, together with magnetic beads coated with sequencing primers as well as reaction buffer.

In this thesis, I will describe the development two analytical tools, metaVIPER and iterClust, for better understand biological states using scRNA-Seq data. Also, I will describe the development scRNA-Seq platform, scPLATE-Seq. Combing these tools, I will answer three specific biological questions, including 1) understanding inter- and intra-tumor heterogeneity of human glioblastoma, 2) elucidating regulators of β-cell de-differentiation in type-2 diabetes, and 3) identifying potential therapeutics targeting cell-state regulators of breast cancer stem cells.

Part II

scPLATE-Seq for Simple, Automated Single-Cell 3’ Transcriptome Profiling Using Early Multiplexing

2.1 Introduction

Since 2012, multiple scRNA-Seq technologies have been introduced [Kolodziejczyk, 2015]. Most involve significant startup cost, an expert operator, or expensive library preparation. Dr. Jeremy Worley and I developed a plate-based library prep that incorporates pre-amplification pooling, unique molecular identifiers (UMIs) [Islam, 2014], and mRNA enrichment using silanol coated magnetic beads. The method uses readily available equipment and is easy to automate. In addition, the method has been optimized to reduce barcode cross contamination that can occur in pooled pre-amplification reactions. This method allows for plate-based 3' scRNA-seq with UMIs at relatively low cost (~5$/cell), using standard lab equipment.

2.2 Methods

2.2.1 scPLATE-Seq protocol

See Figure 2.1 for overview of scPLATE-Seq. See Appendix A for detail.

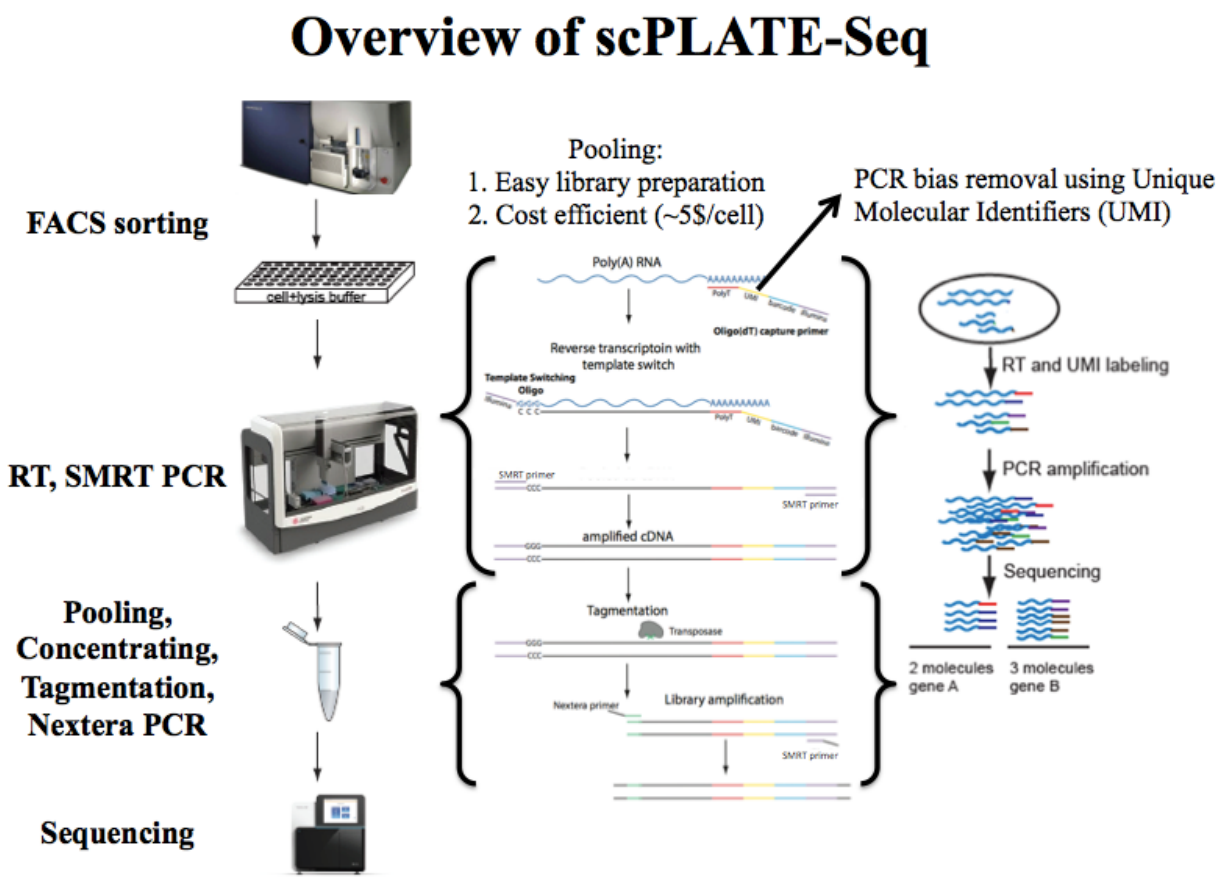

Figure 2.1 Overview of scPLATE-Seq

2.2.2 Expression quantification of HEK293 and U87 single cells.

After demultiplexing, UMIs from the barcode read (8nt) were concatenated onto the 5' part of corresponding mappable reads using a custom Python script (demultiplex.py). Then the concatenated mappable reads were aligned and mapped to UCSC hg38 genome using the STAR 2.5.1a [Dobin, 2013] with option --quantMode TranscriptomeSAM so that output alignments are translated into transcript coordinates for downstream analysis, and --clip5pNbases 8 so that the UMI sequences are not mapped. All other options for STAR are set as default. Then the mapped transcripts per gene and mapped reads per gene were quantified using a custom Python script

(CountUMI.py). When measuring mapped transcripts per gene using hamming distance threshold 1, only if the hamming distance between two UMIs is greater than 1, then they are considered from different transcripts.

2.2.3 Expression quantification of HEK293 bulk sample.

The general procedures are the same as described in 2.2.2, besides only the mapped reads per gene was considered since UMIs were not included in bulk samples. Since the bulk library is also 3' biased (same chemistry with scPLATE-Seq), log2(RPM+1) was used when measuring the correlation between the ensemble single cell and the bulk transcriptome.

2.2.4 Expression quantification of the mix species experiments.

The general procedures are the same as described in 2.2.2, besides the concatenated mappable reads were aligned and mapped to concatenated UCSC hg38 and mm10 genome with option --outFilterMultimapNmax 1 so that only reads that uniquely mapped are considered. When measuring transcripts per gene I used hamming distance threshold 1.

2.2.5 Data and software availability

See Appendix B for detail.

2.3 Results

2.3.1 Measuring robustness of scPLATE-Seq

For robust scRNA-Seq platforms, pooled single cell transcriptomic profiles of relatively homogenous cell lines should recapitulate the corresponding bulk transcriptomic profiles [Wu, 2014]. Therefore we measured the correlation between the ensemble single cell and the bulk

transcriptome were measured. The ensemble single cell was created by computationally pooling all the raw reads from single cell transcriptomes generated using scPLATE-Seq. Results showed bulk transcriptomic profile of HES293 cell line was recapitulated by pooling the corresponding single cells, indicating scPLATE-Seq to be a robust scRNA-Seq platform (Figure 2.2).

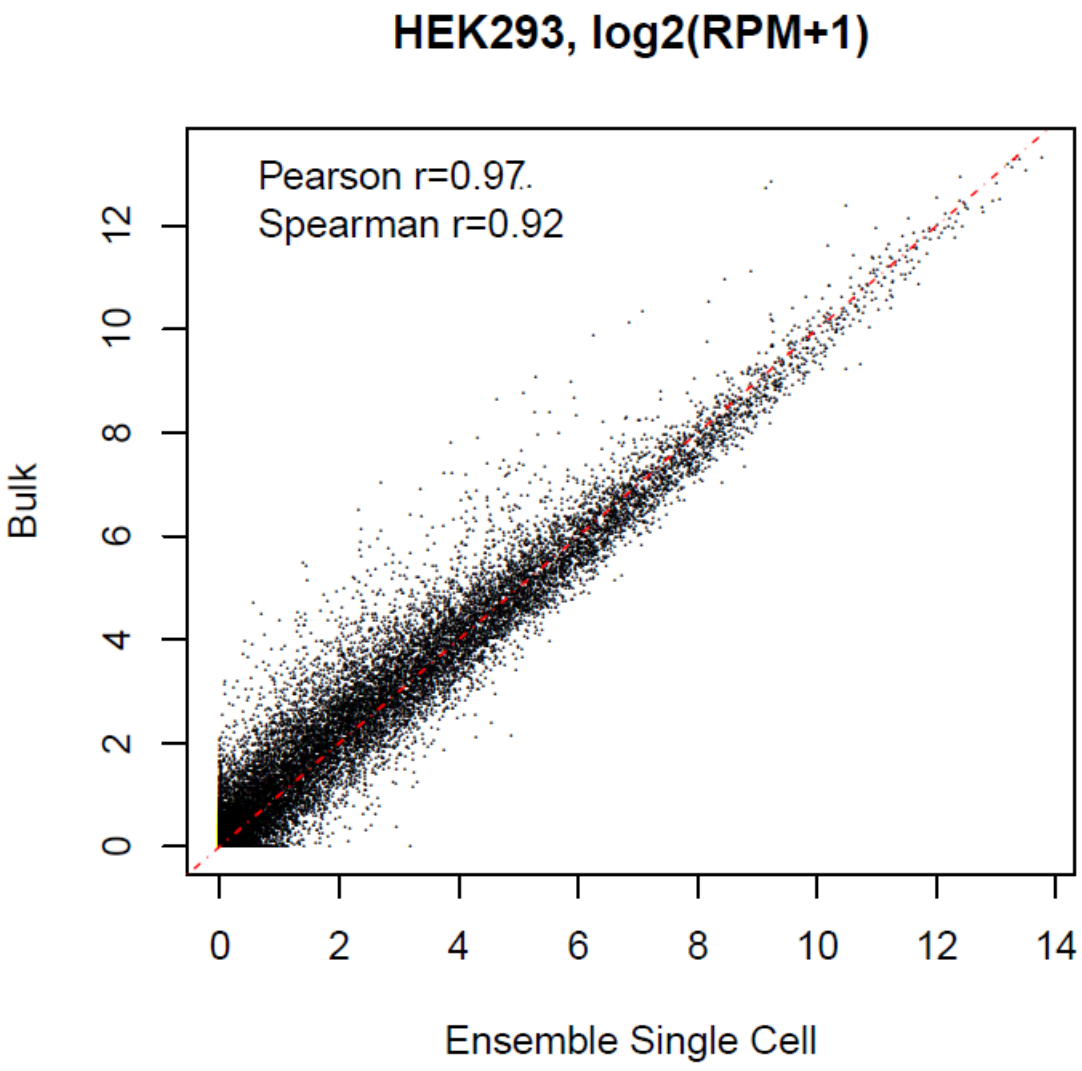

Figure 2.2 Measuring robustness of scPLATE-Seq. For each gene, the log2(RPM+1) values from the ensemble and bulk were plotted against each other.

2.3.2 Measuring non-singlet rate and sample barcode cross-talk rate of scPLATE-Seq

Mixture of human (U87) and mouse (3T3) cells were analyzed by scPLATE-Seq. Comparable human and mouse mapped reads/transcripts among all samples is caused by sample barcode cross-talk, while that among few samples is caused by non-singlet. Results showed non-singlet events are rare, while in general sample barcode cross-talk events are significantly reduced by reducing rounds of Nextera PCR and purifying SMRT PCR products using SPRI or silanol coated magnetic beads. Sample barcode cross-talk usually happens during Nextera PCR amplification, when SMRT PCR products from each individual well are pooled. Excessive free-floating RT (Reverse Transcription) primers, which contain sample-barcode, can be attached to SMRT PCR products from different wells during Nextera PCR amplification, causing cross-contamination between wells. SMRT PCR products purification will remove free-floating RT primers, therefore significantly reduce sample barcode cross-talk events (SMRT PCR, initial amplification of RT products within each well; Nextera PCR, final amplification and adding Illumina primers after SMRT PCR products are pooled). Noticeably, sample barcode cross-talk events are more visible when using UMI to quantify transcriptomic profiles, compared to using mapped reads. This is because transcriptomic profile quantification using UMI is more sensitive, while cross-talk events happen at later PCR cycles will not significantly influence percentage of sample-specific mapped reads (Figure 2.3).

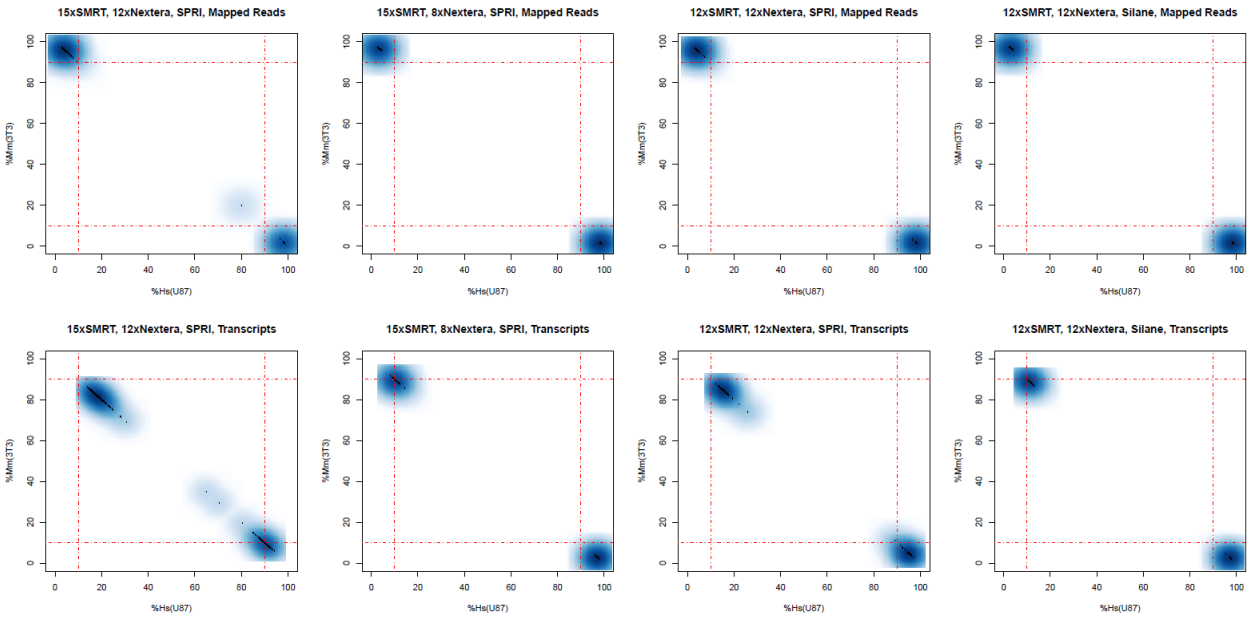

Figure 2.3 Measuring non-singlet rate and sample barcode cross-talk rate of scPLATE-Seq. Percentage of reads (A-D) and transcripts (E-H) uniquely mapped to human/mouse under 15 cycles SMRT, 12 cycles Nextera, SPRI cleanup (A and E), 15 cycles SMRT, 8 cycles Nextera, SPRI cleanup (B and F), 12 cycles SMRT, 12 cycles Nextera, SPRI cleanup (C and G) and 12 cycles SMRT, 12 cycles Nextera, Silane cleanup (D and H) were presented. See Methods for details.

2.3.3 Measuring effectiveness of hamming correction

UMIs were introduced to reduce PCR amplification rate bias in scRNA-Seq. During reverse transcription (RT), each individual transcript will be labeled with a unique UMI sequence. When quantifying transcriptomic profiles, sequencing reads mapped to the same gene with same UMI sequence will be collapsed and considered as from the same transcript. Hamming distance correction is used to correct PCR or sequencing errors on UMI sequences (Figure 2.4).

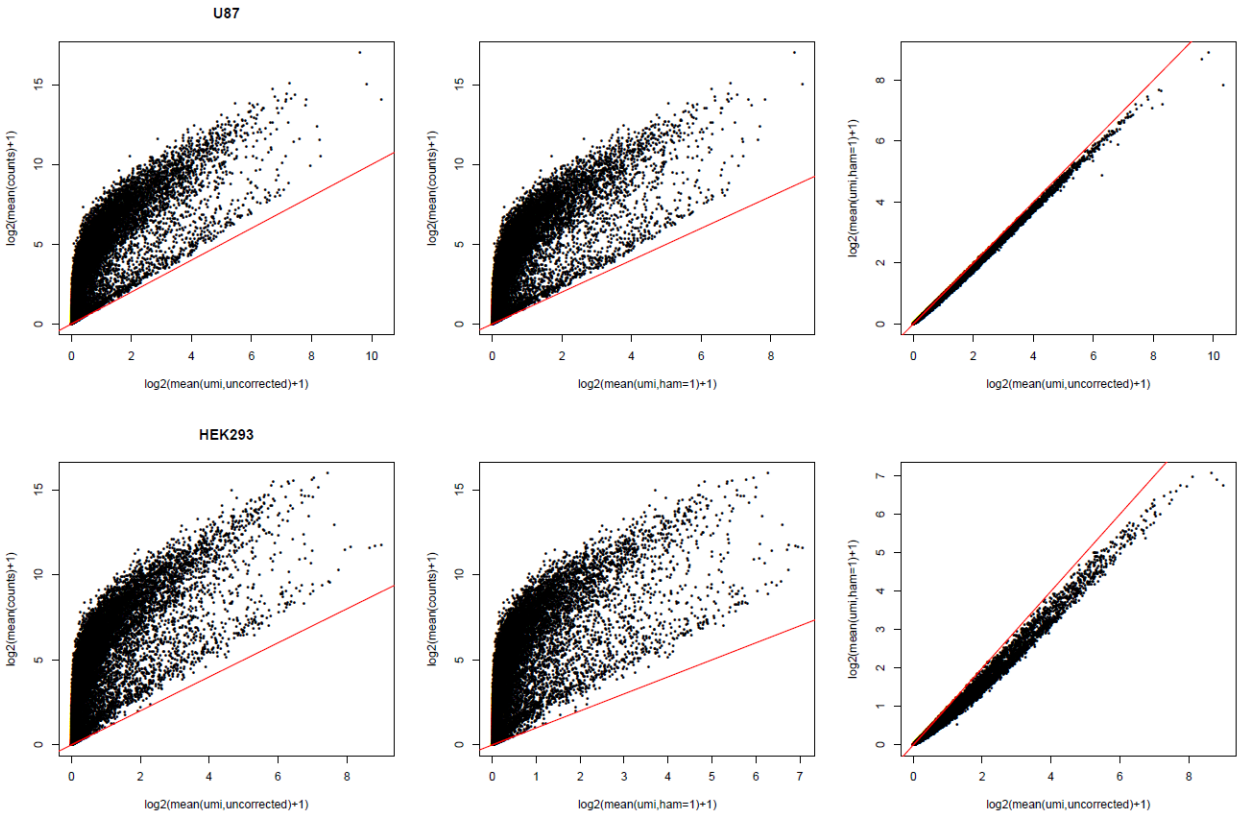

Figure 2.4 Measuring effectiveness of hamming correction. (A, D) Comparison between mapped reads and detected transcripts without hamming correction. (B, E) Comparison between mapped reads and detected transcripts with hamming correction (distance 1). (C, F) Comparison between detected transcripts with and without hamming correction.

2.3.4 Measuring sensitivity and technical noise level of scPLATE-Seq

As measure of sensitivity, I checked the distribution of number of mapped reads, mapped transcripts and detected genes per cell. As measure of technical noise level, I checked average expression and corresponding CV for each gene (Figure 2.5).

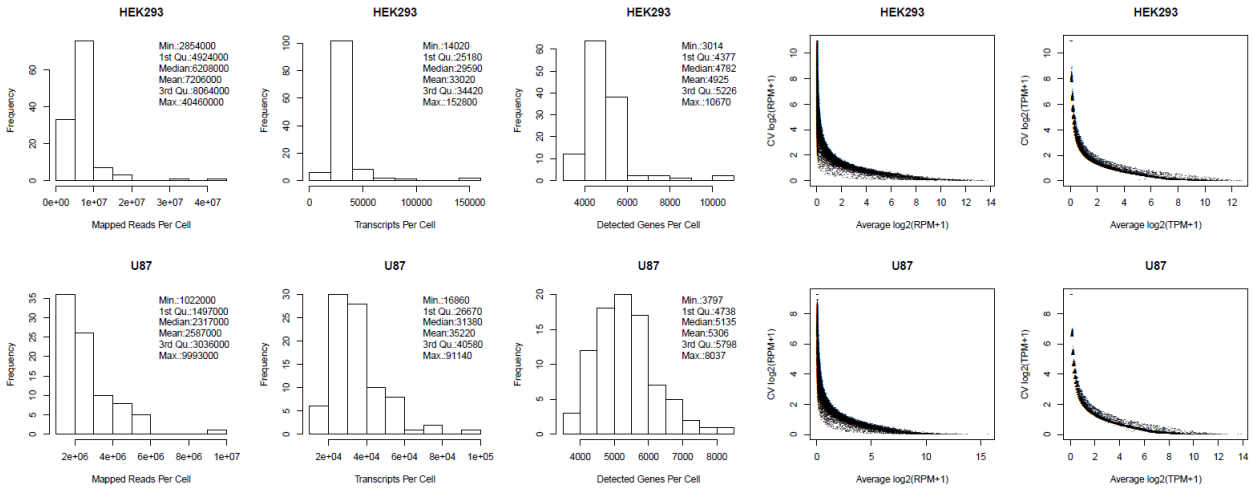

Figure 2.5 Measuring sensitivity and technical noise level of scPLATE-Seq. (A-C and F-H) Sensitivity: distribution of number of mapped reads, mapped transcripts and detected genes (mapped reads/transcripts >0) of HEK293 and U87 cells. (D, E, I and J) Technical noise level: average expression measured by log2(RPM+1) or log2(TPM+1) and corresponding CV for each gene in HEK293 and U87 cells.

2.3.5 Measuring biological differences over batch effect of scPLATE-Seq

Results showed different cell lines (HEK293 and U87) were separated on 2D tSNE plot [Maaten**,** 2008]. Slight batch effect was observed, that HEK293 cells profiled in the same plate tend to cluster together. However such slight batch effect is not influencing the interpretation of biology (Figure 2.6).

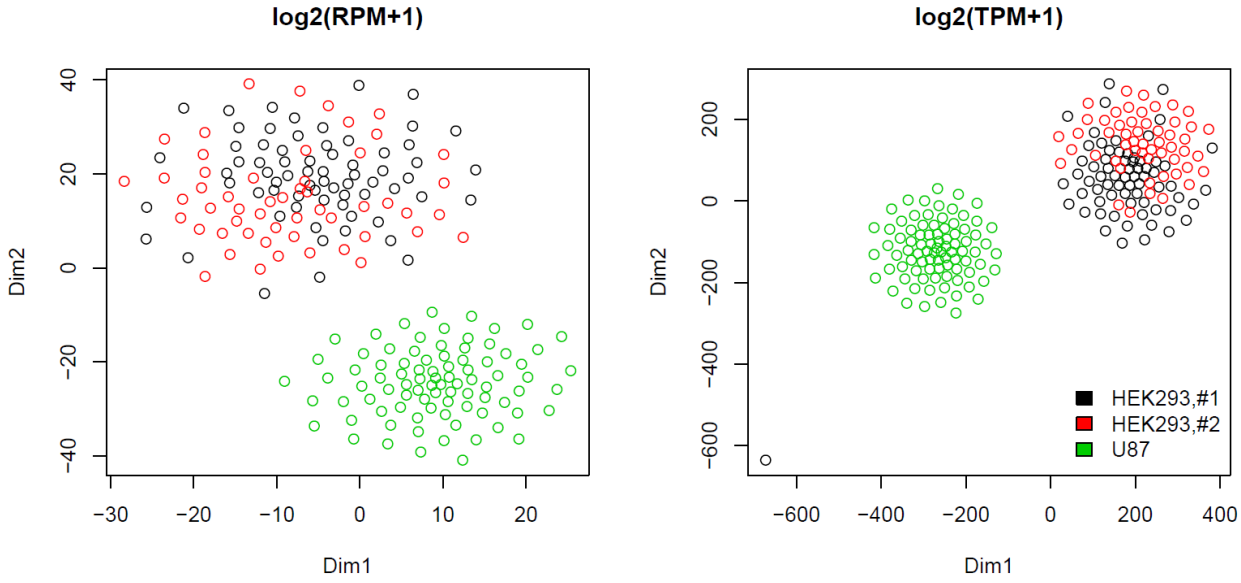

Figure 2.6 Measuring biological differences over batch effect of scPLATE-Seq. (A and B) tSNE [Maaten, 2008] plot based on log2(RPM+1) (A) and log2(TPM+1) measurements of U87 cells, and two batches of HEK293 cells profiled by scPLATE-Seq.

Part III

Quantitative Assessment of Protein Activity in Orphan Tissues and Single Cells Using the metaVIPER Algorithm [Ding, 2018a]

3.1 Introduction

Most biological events are characterized by the transition between two cellular states representing either two stable physiologic conditions, such as during lineage specification [Clevers, 2006; Thiery, 2009] or a physiological and a pathological one, such as during tumorigenesis [Hanahan & Weinberg, 2011; Thiery, 2002]. In either case, cell state transitions are initiated by a coordinated change in the activity of key regulatory proteins, typically organized into highly interconnected and auto-regulated modules, which are ultimately responsible for the maintenance of a stable endpoint state. Dr. Califano have used the term "master regulator" (MR) to refer to the specific proteins, whose concerted activity is necessary and sufficient to implement a given cell state transition [Califano & Alvarez, 2017]. Critically, individual MR proteins can be systematically elucidated by computational analysis of regulatory models (interactomes) using MARINa (Master Regulator Inference algorithm) [Lefebvre, 2010] and its most recent implementation supporting individual sample analysis, VIPER (Virtual Inference of Protein activity by Enriched Regulon) [Alvarez, 2016]. These algorithms prioritize the proteins representing the most direct mechanistic regulators of a cell state transition, by assessing the enrichment of their transcriptional targets in genes that are differentially expressed. For instance, a protein would be considered significantly activated in a cell-state transition if its positively regulated and repressed targets were significantly enriched in overexpressed and underexpressed genes, respectively. The opposite would, of course, be the case for an inactivated protein. As proposed in [Alvarez, 2016], this enrichment can be effectively quantitated as Normalized Enrichment Score (NES) using the Kolmogorov-Smirnov statistics [Subramanian, 2005]. Dr. Alvarez have shown that the NES can then be effectively used as a proxy for the differential activity of a specific protein [Alvarez, 2016]. Critically, such an approach requires

accurate and comprehensive assessment of protein transcriptional targets. This can be accomplished using reverse-engineering algorithms, such as ARACNe (Accurate Reverse Engineering of Cellular Networks) [Basso, 2005] and others (reviewed in [Hecker, 2009]), as also discussed in [Basso, 2005].

MARINa and VIPER have helped elucidate MR proteins for a variety of tumor related [Aytes, 2014; Bisikirska, 2016; Carro, 2010; Chen, 2014; Chudnovsky, 2014; Della Gatta, 2012; Rodriguez-Barrueco, 2015], neurodegenerative [Aubry, 2015; Brichta, 2015; Ikiz, 2015], stem cell [Kushwaha, 2015; Talos, 2017], developmental [Lefebvre, 2010] and neurobehavioral [Repunte-Canonigo, 2015] phenotypes that have been experimentally validated. The dependency of this algorithm on availability of tissue specific models, however, constitutes a significant limitation because use of non-tissue-matched interactomes severely compromises algorithm performance [Aytes, 2014]. Since ARACNe requires $N \geq 100$ tissue-specific gene expression profiles, representing statistically independent samples, some tissue contexts may lack adequate data for accurate interactome inference. These "orphan tissues" include, for instance, rare or poorly characterized cancers as well as progenitor states during lineage differentiation. In addition, the specific tissue lineage of a sample may be poorly defined, thus preventing selection of appropriate interactome models. Consider, for instance, a single cell isolated from a heterogeneous sample, such as whole brain or stroma-infiltrated tumor, where many highly distinct and often uncharacterized cell lineages are inextricably commingled.

To address this challenge, Dr. Califano, Dr. Alvarez and I reasoned that while regulatory models are clearly lineage specific, due to the distinct epigenetic state of the cells, the transcriptional targets of a specific protein (i.e., its regulon) may be at least partially conserved across a small subset of distinct lineages. Thus, once a sufficient number of tissue specific interactomes is

available, the likelihood that one or more of them may represent a good model for the regulon of a specific protein increases, even though one may not know a priori which model may represent the best match for each protein. Indeed, the regulons of different proteins may be optimally represented within different interactomes. This is further helped by the fact that VIPER analysis is robust if at least 40% of a protein's regulon is accurately inferred [Alvarez, 2016]. Conversely, as shown in Figure 3.1 when a protein regulon is incorrectly assessed for a specific tissue, it is not consistent with the tissue-specific gene expression signature, thus producing no significant enrichment. Taken together, these observations constitute the basis for the implementation of a context-independent algorithm for protein activity assessment (metaVIPER).

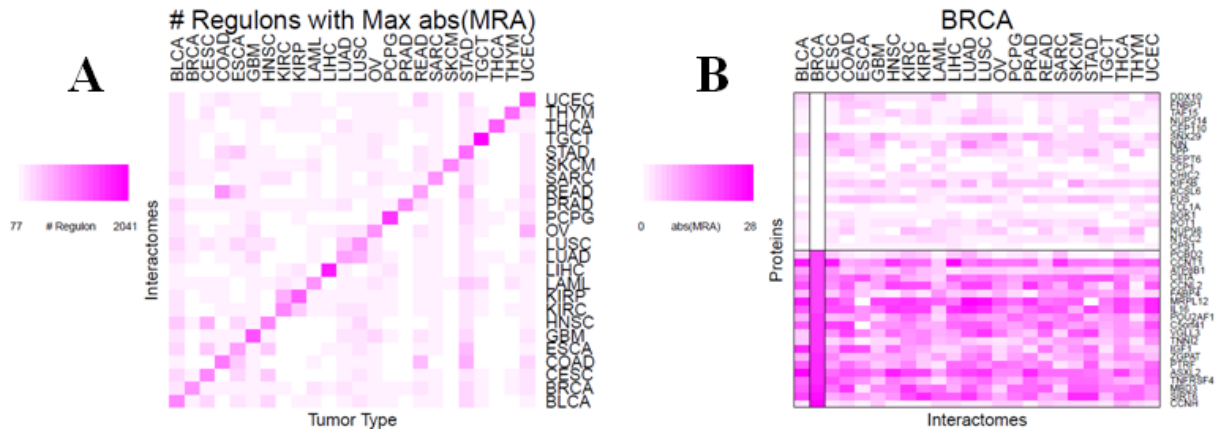

Figure 3.1 Highest absolute NES values are obtained from tissue-matched interactomes. A) Since tissue-matching regulons are in general best models for proteins in that specific tissue, I conclude that correctly assessed regulons usually give high absolute activity. This is demonstrated by that across all tissue types, tissue-matching interactome harbors the most regulons with highest absolute activity. B) In some particular cases, tissue-matching regulons may not constitute the best model. For instance, as shown, in breast invasive carcinoma (BRCA), proteins in the upper panel are best modeled by regulons from other tissues. Also, in some cases, tissue-matching regulons may not be the only appropriate model. For instance, besides BRCA regulons, proteins in the lower panel can also be appropriately modeled by regulons from other tissue types.

MetaVIPER implements a statistical framework for evidence integration across a large repertoire of context-specific interactomes, see Methods for details. The algorithm is based on the assumption that only regulons that accurately represent the transcriptional targets of specific proteins in the tissue of interest will produce statistically significant enrichment in genes that are differentially expressed in that tissue (Figure 3.2A).

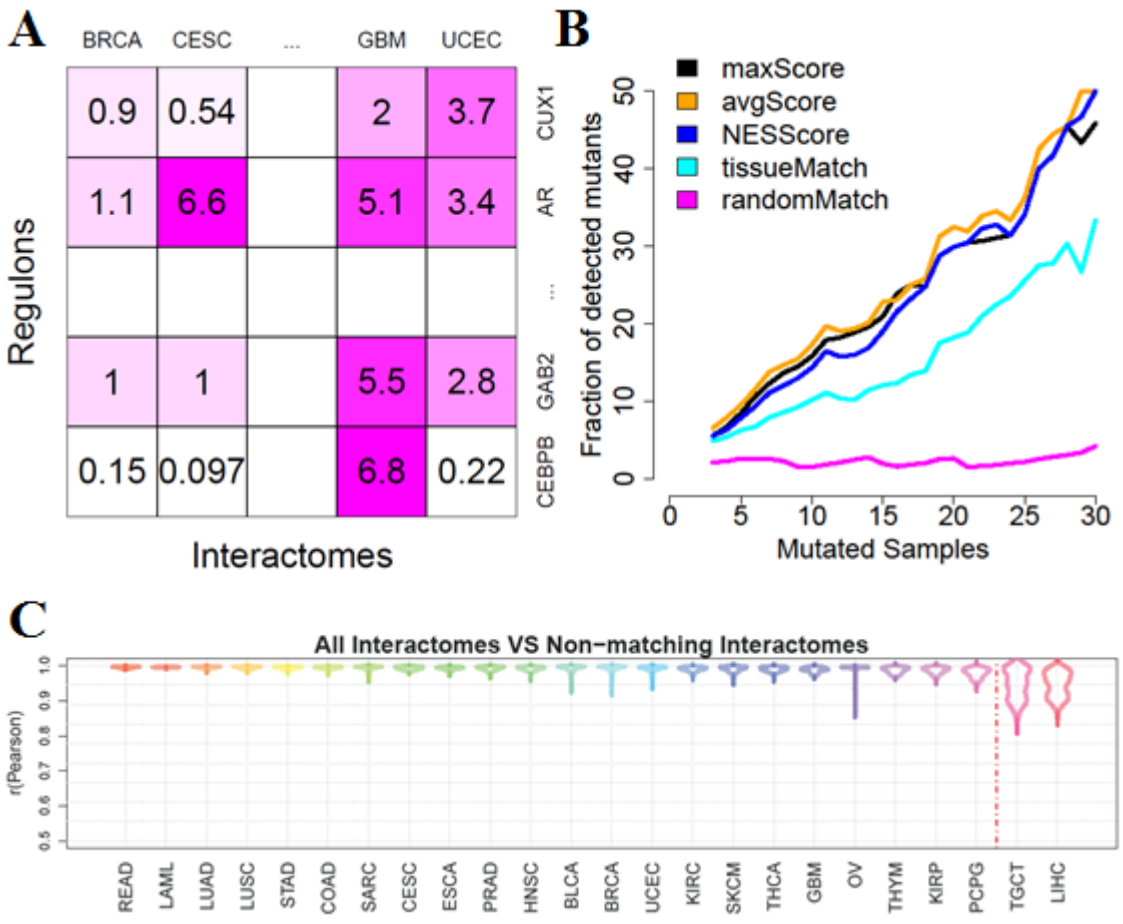

Figure 3.2: Inferring protein activity with metaVIPER. (A) Overview of metaVIPER. The set of transcriptional targets for each regulatory protein (its regulon) constitutes the fundamental building blocks of an interactome, which reflect its overall, context-specific regulatory control structure. MetaVIPER identifies the regulon that best recapitulates the regulatory targets of a protein by assessing its enrichment in the tissue specific differential expression signature. In the example shown here, for instance, the regulon for protein CUX1 in an unknown or orphan tissue is better recapitulated by the uterine corpus endometrial carcinoma (UCEC)-based regulon, while the transcriptional program for the androgen receptor protein (AR) is better recapitulated by the Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) and glioblastoma (GBM)-based regulons. The numbers indicate –log10(p-value) for enrichment of the regulons on the gene expression signature, as computed by VIPER. (B) Impact of recurrent coding somatic mutations on metaVIPER-inferred protein activity. Fraction of proteins showing significant association between metaVIPER-inferred protein activity and somatic mutations ($p < 0.01$) is presented. VIPER analysis was performed using the tissue-matched network (tissueMatch), metaVIPER was performed by integrating the results from individual interactomes using maxScore, avgScore and NESScore methods; the baseline control was computed by using intercatomes selected at random (randomMatch). The X-axis represents the minimum number of TCGA samples presenting the specific gene mutation required for incluion of the encoded protein in the analysis. (C) Inference of protein activity for orphan tissues. MetaVIPER can effectively reproduce differential protein activity in TCGA tissues, even when the corresponding matched interactome is removed from the analysis. The only partial exception is represented by for two tissue lineages – liver hepatocellular carcinoma (LIHC) and testicular germ cell tumors (TGCT) – which are defined by highly specific regulatory programs. The probability density distribution for the correlation between protein activities (NES) inferred by metaVIPER using all available interactomes vs. metaVIPER using all but the tissue-matched interactome (Pearson's correlation) across all samples is shown by the violin plots.

To assess whether metaVIPER can effectively assess protein activity in context-independent fashion I perform a number of distinct benchmarks. First, I assessed whether results produced by analysis of context-specific interactomes (e.g., inferred from breast cancer samples) could be effectively reproduced when only interactomes from other tissues are used in integrative fashion. I also test whether the ability to assess dysregulation of proteins whose encoding gene harbored a recurrent somatic alteration was improved by metaVIPER. Finally, I assess the algorithm's ability to transform low-depth single cell RNA-Seq profiles into highly reproducible protein activity profiles that accurately reflect cell state, while removing technical artifacts and batch

effects, compared to state of the art gene expression based methods. These improvements significantly increase the ability to analyze the biological function and relevance of gene products whose mRNAs are undetectable in low-depth, single cell RNA-Seq data (dropout effect), without any a priori knowledge of the single cell's lineage. In particular, it allows more stringent analysis of critical lineage markers, for which no mRNA reads may be detectable in individual cells, either individually or as a set, supporting a "virtual FACS" analysis.

3.2 Methods

3.2.1 Regulatory networks

All regulatory networks were reverse engineered by ARACNe [Basso, 2005] and summarized in Table. 3.1. Core TCGA RNA-Seq derived interactomes are available in R-package aracne.networks from Bioconductor [Giorgi, 2017]. The TCGA human SKCM network was assembled from RNA-Seq profiles. TCGA RNA-Seq level 3 data (counts per gene) was obtained from the TCGA data portal, and normalized by Variance Stabilization Transformation (VST), as implemented in the DESeq package from Bioconductor [Anders & Huber, 2010]. The human B lymphocyte interactome were reported by [Basso, 2005]. The human T lymphocyte interactome were reported by [Piovan, 2013]. The human brain tumor regulatory networks were assembled from four more gene expression datasets besides the TCGA glioblastoma RNA-Seq dataset. For the Rembrandt, [Phillips, 2006], TCGA-Agilent and TCGA-Affymetrix, informative probe clusters were assembled with the cleaner algorithm [Alvarez, 2009] and the expression data was summarized and normalized with the MAS5 algorithm as implemented in the affy R-package from Bioconductor [Gautier, 2004]. Differences in sample distributions were removed with the robust spline normalization procedure implemented in the lumi R-package from Bioconductor

[Du, 2008]. In a similar way, differences in sample distribution for the TCGA-Agilent dataset were removed by the robust spline normalization method. ARACNe was run with 100 bootstrap iterations using 1,813 transcription factors (genes annotated in Gene Ontology molecular function database, as GO:0003700, 'transcription factor activity', or as GO:0003677, 'DNA binding', and GO:0030528, 'transcription regulator activity', or as GO:00034677 and GO: 0045449, 'regulation of transcription'), 969 transcriptional cofactors (a manually curated list, not overlapping with the transcription factor list, built upon genes annotated as GO:0003712, 'transcription cofactor activity', or GO:0030528 or GO:0045449) and 3,370 signaling pathway related genes (annotated in GO Biological Process database as GO:0007165 'signal transduction' and in GO cellular component database as GO:0005622, 'intracellular', or GO:0005886, 'plasma membrane'). Parameters were set to 0 DPI (Data Processing Inequality) tolerance and MI (Mutual Information) p-value (using MI computed by permuting the original dataset as null model) threshold of 10^{-8}.

Table 3.1 Interactomes used in this work and datasets used to reverse engineer them.

Tissue Type	Expression source	Acronym	# Samples	# Regulators	# Targets	# Interactions
Bladder urothelial carcinoma	TCGA RNA-Seq	BLCA	427	6054	19785	489101
Breast invasive carcinoma	TCGA RNA-Seq	BRCA	1212	6054	19359	331919
Cervical squamous cell carcinoma and endocervical adenocarcinoma	TCGA RNA-Seq	CESC	309	6056	19839	583961
Colon adenocarcinoma	TCGA RNA-Seq	COAD	500	6056	19820	413789
Esophageal carcinoma	TCGA RNA-Seq	ESCA	198	5961	18679	529286
Glioblastoma multiforme	TCGA RNA-Seq	GBM	166	6056	19858	563850
Head and neck squamous cell carcinoma	TCGA RNA-Seq	HNSC	566	6055	19772	423104
Kidney renal clear cell carcinoma	TCGA RNA-Seq	KIRC	606	6054	19843	350478
Kidney renal papillary cell carcinoma	TCGA RNA-Seq	KIRP	323	6055	19858	452653
Acute myeloid leukemia	TCGA RNA-Seq	LAML	179	6007	19269	531535
Liver hepatocellular carcinoma	TCGA RNA-Seq	LIHC	423	6056	19829	469922
Lung adenocarcinoma	TCGA RNA-Seq	LUAD	576	6055	19742	399513
Lung squamous cell carcinoma	TCGA RNA-Seq	LUSC	552	6054	19741	455032
Ovarian serous cystadenocarcinoma	TCGA RNA-Seq	OV	299	6007	19140	647358
Pheochromocytoma and paraganglioma	TCGA RNA-Seq	PCPG	187	6506	19861	603617
Prostate adenocarcinoma	TCGA RNA-Seq	PRAD	550	6053	19820	330922
Rectum adenocarcinoma	TCGA RNA-Seq	READ	177	6056	19856	557911
Sarcoma	TCGA RNA-Seq	SARC	265	6112	20479	526591
Skin cutaneous melanoma	TCGA RNA-Seq	SKCM	472	6053	19840	425361
Stomach adenocarcinoma	TCGA RNA-Seq	STAD	307	6056	21663	561858
Testicular germ cell tumors	TCGA RNA-Seq	TGCT	156	6056	19860	432621
Thyroid carcinoma	TCGA RNA-Seq	THCA	568	6053	19861	317582
Thymoma	TCGA RNA-Seq	THYM	122	6056	19862	387923
Uterine corpus endometrial carcinoma	TCGA RNA-Seq	UCEC	581	6055	19716	469845
T lymphocyte	[Della Gatta, 2012]	T	233	5086	13834	324963
B lymphocyte	[Lefebvre, 2010]	B	201	3651	8699	207336
Skin cutaneous melanoma	TCGA RNA-Seq	SKCM	472	6201	19840	432922
Glioblastoma multiforme	REMBRANDT	NA	804	3921	12683	482879
Glioblastoma multiforme	[Phillips, 2006]	NA	176	3099	8964	382144
Glioblastoma multiforme	TCGA affymetrix	NA	202	3433	9812	421108
Glioblastoma multiforme	TCGA agilent	NA	202	5560	17355	988514

2.2.2 Association somatic mutations with metaVIPER inference

I consider somatic mutations that happen in the same amino acid of a protein within at least 3 patients as recurrent somatic mutations. Then for each protein, I did enrichment analysis with

activity profile for each patient as signature, and patient harboring recurrent somatic mutation for that specific protein as enriching set. I consider proteins with significant enrichment score ($p < 0.01$) as showing significant association between inferred protein activity and recurrent somatic mutations. Then I check the fraction of proteins that can be associated with recurrent somatic mutations, and use that as criteria in evaluating the performance between VIPER and metaVIPER. In order to get enough mutated patient for each protein, this analysis is done in a tumor type non-specific manner.

3.2.3 Preparation of glioblastoma mouse model

PDGFB-IRES- CRE expressing retrovirus was injected into the rostral subcortical white matter of adult *Pten* $^{lox/lox}$ */p53* $^{lox/lox}$ */luciferase-* $^{stop\text{-}lox}$ transgenic mice [Lei, 2011; Sonabend, 2013]. Mice developed brain tumors with the histopathological features of glioblastoma by 28 days post injection with retrovirus.

3.2.4 Generate scRNA-Seq profiles for glioblastoma mouse model

In Dr. Canoll's lab, following IACUC guidelines, animals were sacrificed at the first sign of morbidity. Ex-vivo gross total resection of the tumor was performed and tumor cells were isolated using enzymatic digestion [Gensert, 2001]. The isolated cells were cultured in a 2:1 ratio of basal media (DMEM, N2, T3, 0.5% FBS, and penicillin/streptomycin/amphotericin) in B104 conditioned media [Canoll, 1996]. This media was further supplemented with PDGF-AA (Sigma-Aldrich; St. Louis, MO) and FGFb (Gibco; Grand Island, NY) to a concentration of 10ng/ml. Dr. Sims' lab loaded dissociated cells into a Fluidigm Integrated Fluidic Circuit with capture sites designed for 10-17 μm diameter cells after staining the single cell suspension with Calcein AM (Life Technologies). They then imaged the cells that had been captured on-chip

with both bright field and fluorescence microscopy using an inverted Nikon Eclipse Ti-U epifluorescence microscope with a 20x, 0.75 NA air objective (Plan Apo λ, Nikon), a 473 nm diode laser (Dragon Lasers), and an electron multiplying charge coupled device (EMCCD) camera (iXON3, Andor Technologies). This allowed them to identify capture sites with zero, one, and more than one cell and also to identify capture sites containing living cells, based on the Calcein AM fluorescence. They then lysed the cells, reverse transcribed mRNA into cDNA, and pre-amplified full-length cDNA by PCR automatically using the Fluidigm C1 Autoprep instrument according to the manufacturer's instructions. Finally, they harvested individual cDNA libraries from the microfluidic device and converted them into indexed, Illumina sequencing libraries by in vitro transposition and PCR using the Nextera system (Illumina). The pooled libraries were sequenced on a single lane of an Illumina HiSeq 2000 with single-end 100-bp reads. After demultiplexing, the resulting raw reads were aligned to the murine genome and transcriptome annotation (mm10, UCSC annotation from Illumina iGenomes) with Tophat 2. Uniquely aligned, exonic reads were then quantified for each gene using HTSeq. In total they obtained 85 single cells.

3.2.5 Data and software availability

See Appendix C for detail.

3.3 Results

3.3.1 Overview of metaVIPER

Let's assume a tissue context T for which a matched tissue specific interactome were not available. Furthermore, without loss of generality, let's focus on a specific protein of interest P

and on its T-specific regulon R_T. Given a sufficient number of additional tissues T_1 … T_N for which accurate, context-specific interactomes are available, I hypothesize that R_T will be at least partially recapitulated in one or more of them. Based on previous results[7], VIPER can accurately infer differential protein activity, as long as 40% or more of its transcriptional targets are correctly identified. As a result, even partial regulon overlap may suffice. Indeed, paradoxically, there are cases where a protein's regulon may be more accurately represented in a non-tissue matched interactome than in the tissue-specific one. This may occur, for instance, when expression of the gene encoding for the protein of interest has little variability in the tissue of interest and greater variability in a distinct tissue context where the targets are relatively well conserved. A key challenge, however, is that one does not know *a priori* which of the tissue-specific interactomes may provide reasonable vs. poor models for R_T.

To address this challenge, I leverage previous studies showing that if an interactome-specific regulon provides poor R_T representation, approaching random selection in the limit, then it will also not be statistically significantly enriched in genes that are differentially expressed in a tissue-specific signature S_T. Thus, if one were to compute the enrichment of all available regulons for the protein P in the signature S_T, only those providing a good representation will produce statistically significant enrichment, if P is differentially active in the tissue of interest. Conversely, if the protein is not differentially active in T, then no regulon R_{T1} … R_{TN} should produce statistically significant enrichment. If these assumptions were correct, given a sufficient number of tissue-specific interactomes, this would provide an efficient way to integrate across them to compute the differential activity of arbitrary proteins in tissue contexts for which a suitable interactome model may be missing.

To determine the best strategy for integrating the statistics of the enrichment across multiple interactomes, I compared several approaches. Specifically, for each protein, I first computed enrichment using a tissue-matched interactome (*tissueMatch*). This corresponds to the original implementation of the VIPER algorithm. I then compared these results to those obtained using different metrics to integrate across the regulons of all non-tissue-matched interactomes, including: (a) the NES with the most statistically significant absolute value (*maxScore*); (b) the average of all NES scores (*avgScore*) and (c) the weighted-average of all NES scores, weighed by the NES absolute value (*NESScore*). For these tests, I used a total of 24 interactomes generated from TCGA cohorts, see Table 3.1 [Giorgi, 2017].

To objectively evaluate the performance of these alternative integrative methods, I considered a comprehensive set of proteins, whose genes harbor recurrent somatic mutations, as reported by both TCGA and COSMIC (see Methods). These mutations drive tumorigenesis by altering the activity of key oncogenes and tumor suppressors and have been used to identify proteins for targeted inhibitors, based on the oncogene additional paradigm [Weinstein, 2002]. I thus assessed method performance by assessing the statistical significance of the correlation between metaVIPER-inferred protein activity and the presence of a recurrent genetic alterations in the corresponding gene locus ($p < 0.01$), under the assumption that better methods would yield higher significance. To produce an optimal metric across all recurrent mutational events, I assessed correlation as a function of recurrence (Figure 3.2B). Indeed, the more recurrent a mutation is, the more likely it is to be functionally relevant and thus affect the corresponding protein's activity. Recurrence is reported as the number of samples in TCGA and COSMIC where a specific gene locus was mutated, see Methods. As shown in Figure 3.2B, there is a clear trend showing that the more recurrently mutated a gene locus is, the larger the fraction of

proteins showing statistically significant correlation between metaVIPER-inferred protein activity and mutational state. For instance, about 50% of the genes harboring locus-specific mutations in at least 30 TCGA samples could be detected as producing differentially active proteins by metaVIPER analysis ($p < 0.01$).

Surprisingly, based on this metric, all four strategies for cross-tissue integration (metaVIPER) significantly outperformed the use of tissue-specific interactomes, i.e. the original VIPER algorithm (*tissueMatch*). This suggests that integrating the structure of regulatory networks across a large number of representative tissue types provides a more informative regulon representation on an individual protein basis. The *randomMatch* method serves as a baseline negative control, in which for each sample, protein activity was computed using VIPER with an interactome selected at random. As discussed in the following sections, I performed several additional benchmarks to comprehensively and systematically assess the method's performance in orphan tissues, as well in single cells.

3.3.2 MetaVIPER-based protein activity inference in orphan tissues

Small sample size severely undermines the performance of ARACNe, which typically requires at least 100 independent samples, representative of the same tissue lineage [Margolin, 2006] to perform accurate regulon inference for VIPER analysis. This significantly limits the ability to accurately measure protein activity in orphan tissues, defined as rare or poorly characterized tissue types, for which the number of available gene expression profiles is not sufficient to produce an accurate interactome model. For instance, considering tumor cohorts in the TCGA repository, I identified Cholangiocarcinoma (N = 36) and Uterine Carcinosarcoma (N = 57) could be considered orphan tissues for which an accurate ARACNe network could not be

generated. Orphan tissues also include a variety of normal or non-cancer, disease-related cell states that lack appropriate gene expression profile characterization, including many of the intermediate states of differentiation representing multipotent or progenitor population.

Since metaVIPER is designed to infer protein activity without requiring a tissue-specific regulatory model, I designed an objective benchmark to assess metaVIPER's ability to accurately measure protein activity in orphan tissues. I first assembled a gold-standard set using metaVIPER to assess the activity of all proteins for which an ARACNe regulon was generated (see Methods), in each sample of each TCGA cohort, using all available TCGA interactomes including the tissue-matched one. This is preferred to using only the tissue-matched interactome because from the objective benchmark using mutational data this methodology has emerged as being more accurate than the original VIPER analysis. However, for completeness, I also report results of this analysis using the tissue-matched interactomes as gold-standard, see Figure 3.3. I then performed the same analysis using metaVIPER with all available TCGA interactomes, except for the tissue-matched one. For instance, consider Rectum Adenocarcinoma (READ) as a tumor for which an ARACNe interactome could not be accurately inferred. I would then compute the VIPER-inferred activity of all proteins in each TCGA READ sample using either all available TCGA interactomes (gold-standard reference) or all interactomes except for the READ interactome, exactly as if it were not available. I then measure overall protein activity correlation between the two analyses as a quality metric for metaVIPER ability to correctly infer protein activity in the absence of a tissue-matched interactome. This benchmark was performed for each of the all 24 tissue types in TCGA, see Table 3.1 [Giorgi, 2017]. Results show extremely strong average correlation ($\rho > 0.97$) between the two analyses for 22 out of 24 tissues (excluding LIHC and TGCT). This suggests that, even in the absence of a tissue-matched model, most tissues may

be studied virtually without loss of resolution using metaVIPER (Figure 3.2C, Figure 3.3). Thus most orphan tissues can be studied using metaVIPER with virtually no notable result quality degradation. Not surprisingly, the two outlier tissues have a rather unique nature. Indeed, liver hepatocellular carcinoma (LIHC) is originated from hepatocytes, which are unique endoderm derived secretory cells [Thorgeirsson, 2002]. Similarly, testicular germ cell tumors (TGCT) originate from testicular germ cells, which are specialized pluripotent cells that give rise to gametes [Bosl & Motzer, 1997]. Hepatocytes and testicular germ cells are thus highly specialized issues with no other related tissues among the 24 in TCGA. However, as the number of interactomes in my repertoire grows the probability of having true outlier tissues will decrease. Note, however that, despite their specialized nature even the two outlier tissues presented high average correlation with the results of the tissue-matched analysis ($\rho > 0.95$).

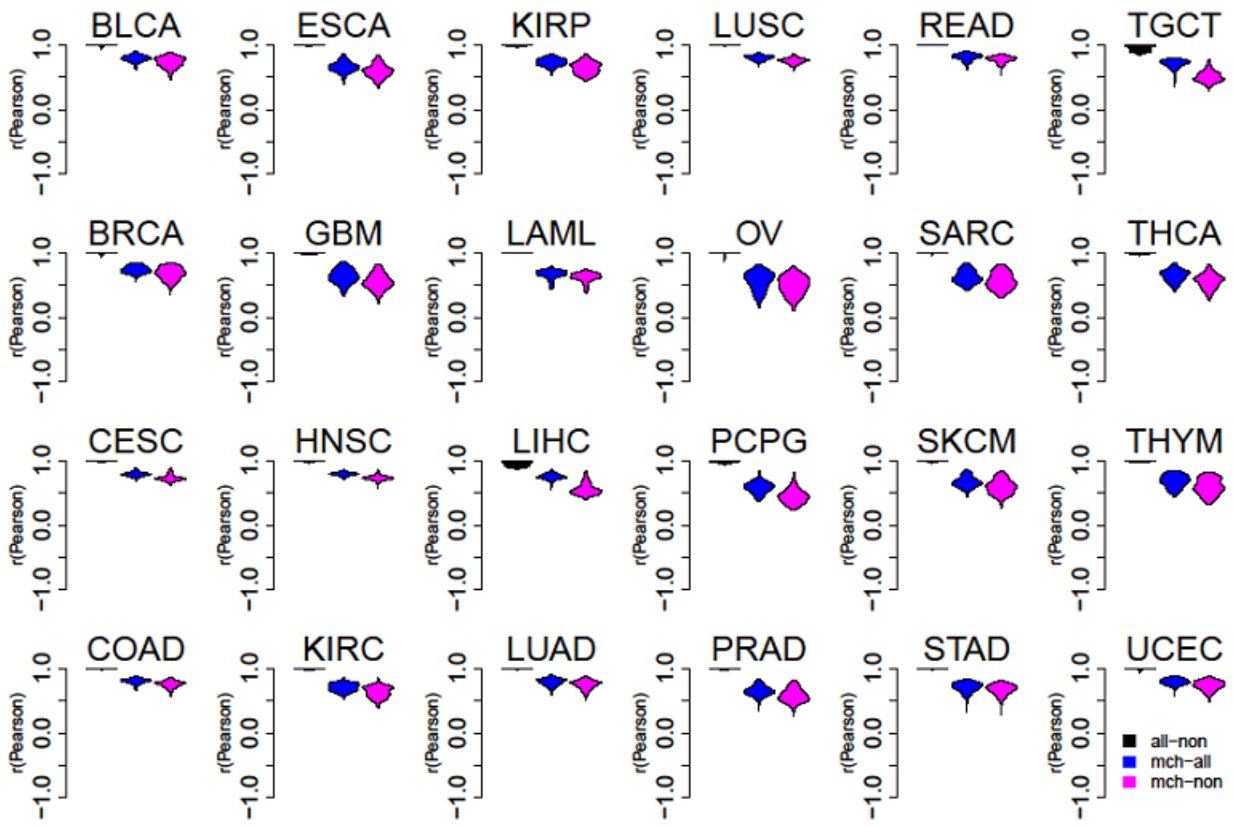

Figure 3.3 Inference protein activity for orphan tissues. Correlation of protein activity inferred from 1) metaVIPER, with all available interactomes (all), 2) metaVIPER, with all non-matching interactomes (non) and 3) VIPER with matching interactome (mch). Then violin plots show the probability density distribution for the Pearson's correlation coefficient for each of the evaluated tissue types.

This raises the important issue of an objective metric to assess whether metaVIPER – when used with a specific repertoire of tissue specific interactomes – is adequate for inferring protein activity in tissues lacking a matched interactome (i.e., orphan tissues). To achieve this goal, as proposed in [Basso, 2005], I will use the *Empirical Cumulative Distribution Function of the absolute value of the VIPER Normalized Enrichment Score* ($ECDF_{|NES|}$) of all proteins in an orphan tissue sample or samples [Aytes, 2014]. In Figure 3.4, I show violin plots for the $ECDF_{|NES|}$ of each TCGA cohort, using the corresponding tissue-matched interactome. The

rightmost plot (TCGA) shows the average of all cohort-specific probability densities. This provides a useful reference to assess whether a specific interactome repertoire is adequate for the metaVIPER-based analysis of an orphan tissue. For instance, I analyzed LAML samples using only a GBM interactome, which would be clearly inappropriate since LAML and GBM cells belong to epigenetically distinct lineages. The result is shown in the first-to-last violin plot (Neg.Ctrl.). As shown this ECDF is clearly an outlier with respect to All-TCGA. Thus, by comparing the ECDF for a tissue of interest against the All TCGA reference, one can effectively assess the quality of the analysis.

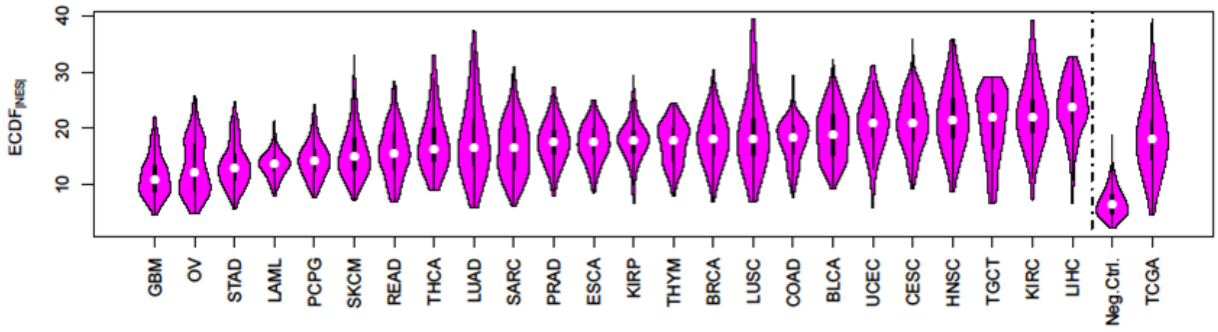

Figure 3.4 Quality control for protein activity analysis. Since properly assigned regulons give high absolute normalized enrichment score (Figure 3.1), therefore I use the *Empirical Cumulative Distribution Function* of the absolute value of the VIPER *Normalized Enrichment Score* ($ECDF_{|NES|}$) of all proteins with significant predicted activity to estimate whether the protein activity analysis is satisfactory. I provided the distribution of the proposed score within each tumor type (GBM, OV etc.) as well as among all TCGA samples (TCGA) using tissue-matching interactome as references for trustworthy protein activity analysis. I also analyzed LAML samples with GBM interactome, which is completely misassigned (LAML and GBM have distinct lineage origination) as the negative control (Neg.Ctrl.). If metaVIPER analysis gives similar result, probably the included interactomes don't have a satisfactory coverage on regulatory information of the analyzed samples.

3.3.3 Single cell analysis

The last few years have seen tremendous development of single-cell profiling methodologies and in particular of single cell RNA-Seq (scRNA-Seq). The advent of these technologies provides

new insight in understanding transition, maintenance, and cell-cell communication processes, across cell states and at an individual cell resolution [Kolodziejczyk, 2015]. However, a major challenge of these approaches is related to the very low depth of sequencing ranging between 10K and 200K reads per cell. While this is sufficient to perform coarse analyses, such as multi-dimensional clustering to identify molecularly distinct sub-populations, it is extremely ineffective in precisely quantitating the expression of individual genes. Indeed, the vast majority of genes lack even one mRNA read in individual cells (dropouts) and a large number have a single read. Due to these significant dropout effects, elucidating biological mechanisms at the single cell level remains challenging. In contrast, as shown in [Alvarez, 2016], VIPER analysis is largely unaffected by sequencing depth because differential protein activity is assessed based on the differential expression of hundreds of transcriptional targets. Thus, measurement and biological noise sources are effectively averaged out, resulting in highly reproducible measurements. Indeed, I have shown that VIPER-inferred protein activity profiles from FFPE samples were extremely well correlated to those from fresh-frozen samples, despite dramatic loss of correlation at the gene expression level [Alvarez, 2016], leading to NYS CLIA approval of two VIPER-based tests. As a result, one would expect VIPER to be well suited to performing analysis of single cell populations in a way that is amenable to quantitative protein activity assessment.

Unfortunately, however, when dealing with heterogeneous samples, the specific tissue context of each individual cell cannot be determined *a priori*. Even if this were possible, it is unlikely that context specific interactomes would be available for rare lineages and progenitor states that are captured by single cell profiling methodologies. MetaVIPER represents a useful alternative in these cases, because, while preserving the robustness of VIPER, it is agnostic to tissue type and

should thus be well-suited to analysis of single cell gene expression profiles from heterogeneous tissues.

To illustrate metaVIPER applicability to single cell expression profile data, Drs. Canoll and Sims groups specifically profiled 85 single cells (see Methods) from a mouse glioblastoma (GBM) model [Lei, 2011; Sonabend, 2013]. Previous studies have demonstrated that GBM comprises two major subtypes, mesenchymal (MES) and Proneural (PN), which may present different proliferation capability (Prolif) [Carro, 2010; Phillips, 2006; Verhaak, 2010; Ceccarelli, 2016]. I inferred protein activity at the single cell level by metaVIPER analysis across 5 brain tumor interactomes, and 24 TCGA human cancer tissue interactomes (see Methods and Table 3.1). Contrary to gene expression profile analysis, the inferred protein activity signatures clearly captured single cells representing MES and PN subtypes. Indeed, unsupervised metaVIPER analysis recapitulated previously reported subtype-specific master regulator proteins [Carro, 2010], which were identified among the most dysregulated on a single-cell basis (Figure 3.5A). Such level of resolution could not be recapitulated by differential gene expression analysis, largely due to transcript-level noise in individual cells (Figure 3.5B). Unsupervised clustering analysis of metaVIPER-inferred protein activity efficiently separated single cells in two major groups, with ~40% of the cells recapitulating the activity pattern of previously described MR proteins of MES (*FOSL1, FOSL2, RUNX1, CEBPB, CEBPD, MYCN, ELF4*), and the remaining ~60% recapitulating those of the PN such as *OLIG2* and *ZNF217*. In sharp contrast, unsupervised, gene-expression based cluster analysis could not effectively separate individual cells in distinct clusters (Figure 3.5A, B). Indeed, ~40% of the critical subtype-related proteins were undetectable at the gene expression level in any of the single cells (black horizontal bars in Figure 3.5A). Expression profiles from single cells are very noisy, due to low sequencing depth,

thus reducing the ability to study their biology. Indeed, low depth of sequencing represents a major confounding factor that can be effectively remedied by metaVIPER analysis.

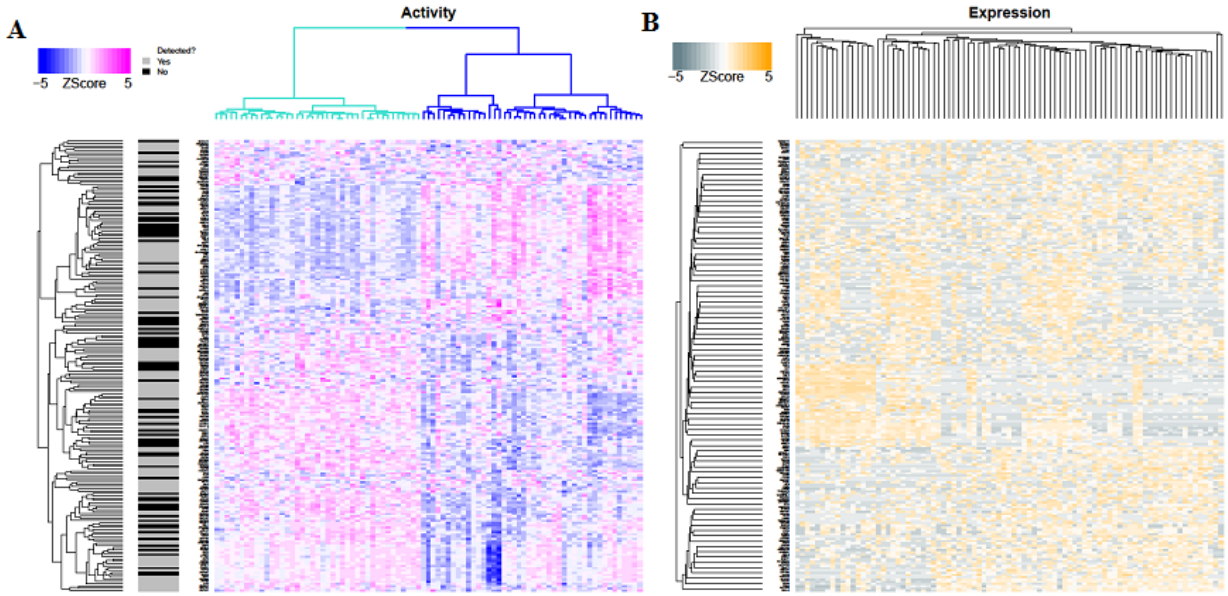

Figure 3.5 Inference of protein activity for single cells from GBM mouse model. (A) MetaVIPER-based protein activity analysis of single cells from a mouse GBM model [Lei, 2011; Sonabend, 2013] by unsupervised clustering using all annotated transcriptional factors, co-transcriptional factors and signaling proteins. Two major clusters were identified, corresponding to established Mesenchymal (MES, blue) and Proneural (PN, turquoise) subtypes, with varying proliferative (Prolif) potential. Indeed, among the top 200 transcriptional factors (i.e., with the highest inter-cluster activity variability), I found established master regulatory transcriptional factors of the MES (*FOSL1, FOSL2*, *RUNX1*, *CEBPB*, *CEBPD*, *MYCN, ELF4*), PN (*OLIG2, ZNF217*) and Prolif (*HMGB2*, *SMAD4*, *PTTG1*, *E2F1*, *E2F8*, *FOXM1*) subtypes [Carro, 2010]. (B) Subtype representation is lost when clustering is performed based on gene expression profiles.

Quality of single cell gene expression profiles is generally reflected by the number of detected genes [Kolodziejczyk, 2015]. Higher quality gene expression profiles, as identified by higher transcriptome complexity, tend to result in higher correlation between the profiles of single cells in the same sub-population clusters (Figure 3.6A, B). Once processed with metaVIPER, however, not only intra-population correlation between individual cells increases significantly but it also becomes virtually independent of transcriptome complexity (Figure 3.6C). This is because

protein activity inference is based on the expression of many target genes and is thus much more robust than estimating gene expression from a single measurement, thus improving resilience to low-quality data.

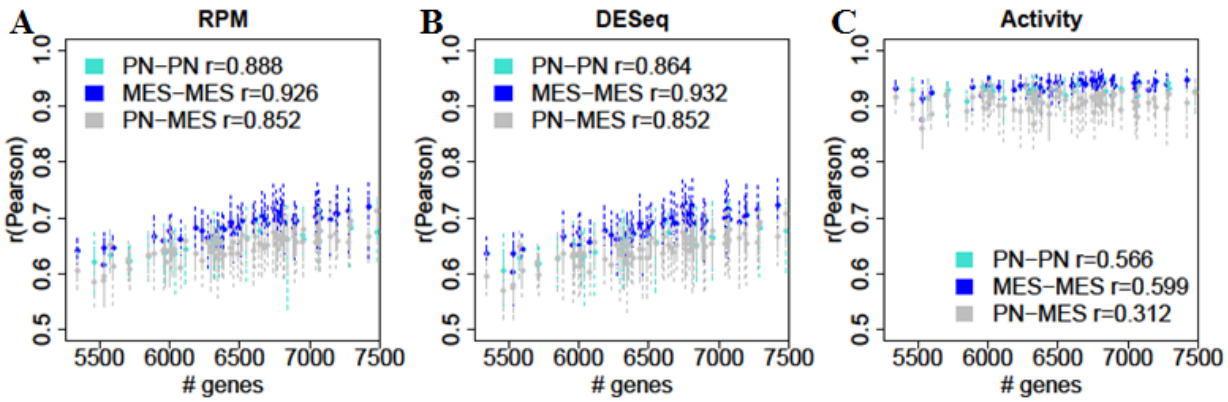

Figure 3.6 Single cell quality as confounding factor in understanding heterogeneity and regulatory properties. With expression measured by both log2(rpm+1) (A) and variance stabilizing transformation (VST) in DESeq R package [Anders & Huber, 2010] (B), cells with higher quality (higher number of genes detected) tend to have higher correlation with other cells. In some cases, inter-population correlation between high quality cells even exceeds intra-population correlation between low quality cells. This is no longer seen with metaVIPER predicted protein activity (C), indicating it to be a more robust measurement of heterogeneity and regulatory properties from single cells

I further tested my methodology on single cell data from tissue representing a complex mixture of melanoma cells and infiltrating B and T lymphocytes [Tirosh, 2016]. By integrating interactomes representative of skin cutaneous melanoma (SKCM, see Methods), B [Basso, 2005] and T [Piovan, 2013] lymphocytes, as well as 24 TCGA human cancer tissue [Giorgi, 2017] (Table 3.1), metaVIPER was able to infer protein activity profiles that effectively discriminate between these different cell types (Figure 3.7A). Furthermore, it revealed differential activity of established lineage markers that could not be detected at the gene expression level (Figure 3.7B-J). This represents a critical value of this approach, as many important lineage markers and other transcriptional regulators may yield no scRNA-Seq reads, due to their relatively low transcript

abundance combined with low sequencing depth. Based on a metric assessing the dynamic range of protein activity in different sub-clusters, metaVIPER significantly outperformed single-regulon-based VIPER analysis on this dataset (Figure 3.8). Most importantly, metaVIPER correctly inferred the differential, tissue-specific activity of established lineage determinants at the single cell level (Figure 3.7B-J). For instance, *PAX5* [Nutt, 1999], *EBF1* [Lin, 2010] and *E2A* [Bain, 1994] showed significantly higher activity in B lymphocytes (one-tail, $p < 10^{-10}$); *MITF* [Levy, 2006], *CTNNB1* [Rubinfeld, 1997] and *HMGB1* [Lotze, 2005] showed significantly higher activity in melanoma cells (one-tail $p < 10^{-10}$); finally, *BCL11B* [Li, 2010], *FOXP3* [Hori, 2003] and *TBET* [Szabo, 2000] showed significantly higher activity in T lymphocytes (one-tail $p < 10^{-10}$). Conversely, I could not detect significant gene expression differences for most of these genes (e.g. $p_{HMGB1} > 0.9$) in melanoma cells, or expression was barely detected at all (average transcripts per million < 1), see E2A in B lymphocytes or FOXP3 and TBET in T lymphocytes, for instance (Figure 3.7B-J).

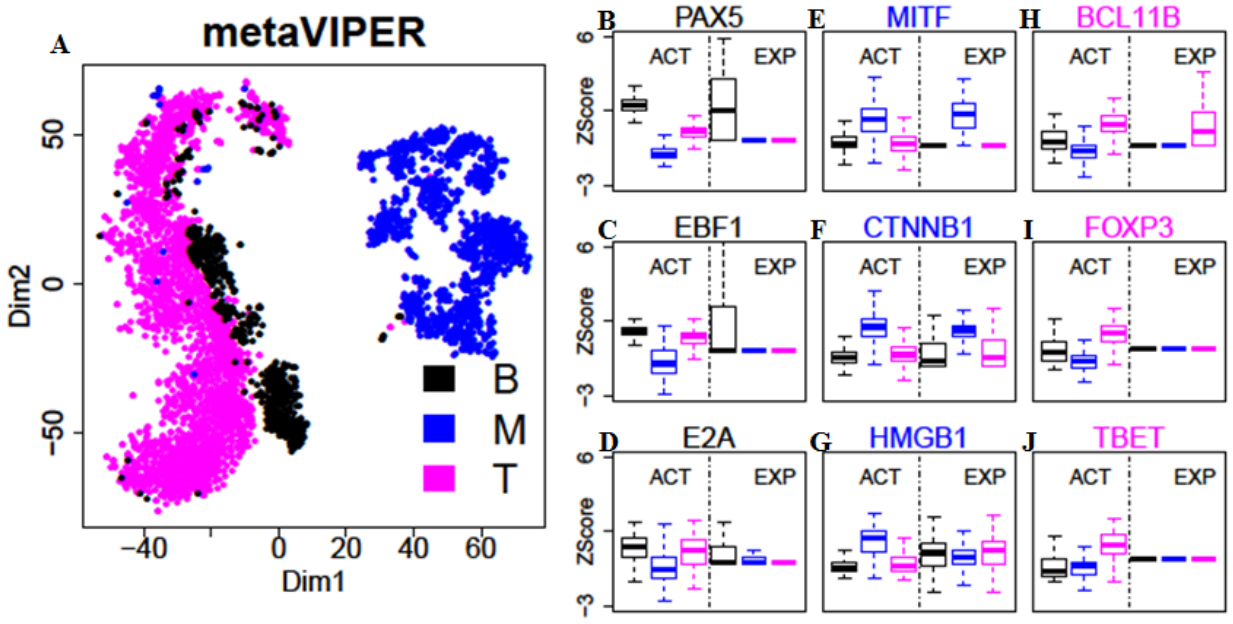

Figure 3.7 Inference of protein activity for single cells profiled by [Tirosh, 2016]. (A) Annotated cell types (B: B lymphocyte T: T lymphocyte, M: melanoma cell) were separated by t-SNE analysis, using metaVIPER-inferred activity for all annotated transcriptional factors, co-transcriptional factors and signaling proteins. Boxplots show metaVIPER-inferred activity, as well as gene expression for tissue- specific lineage markers, including *PAX5* [Nutt, 1999], *EBF1* [Lin, 2010] and *E2A* [Bain, 1994] for B lymphocyte (B-D), *MITF* [Levy, 2006], *CTNNB1* [Rubinfeld, 1997] and *HMGB1* [Lotze, 2005] for melanocyte (E-G), *BCL11B* [Li, 2010], *FOXP3* [Hori, 2003] and *TBET* [Szabo, 2000] for T lymphocyte (H-J). While these markers are significantly differentially active in these tissues, they could not be effectively assessed at the single cell level, either because no mRNA reads were detected or because markers were not statistically significant in terms of differential gene expression. Boxplots showed the median, lower/upper whiskers and hinges of z-scores.

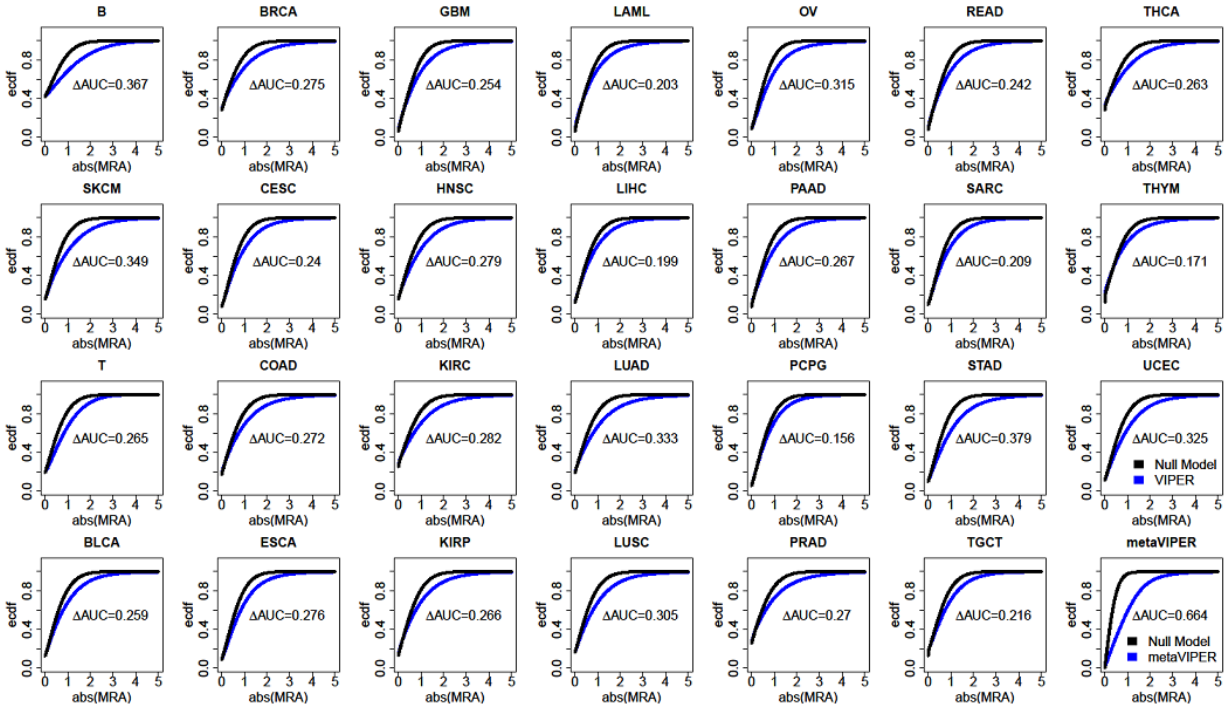

Figure 3.8 metaVIPER integration outperforms analysis with single interactome. As shown in Figure 3.1, for a single cell type, the tissue-matched regulatory model usually gives the highest absolute protein activity inferences. That is to say, in a scenario where different cell types exist, the best regulatory model will give high variance of protein activity across the dataset. I compared the variance of protein activity inferences based on VIPER analysis across 27 distinct tissue-lineage contexts with that of metaVIPER integration. In all cases, metaVIPER outperformed the VIPER analysis. The figures show the ΔAUC between ecdf curve for a null model, built by uniformly shuffling the expression profile sample-wise (shown in black), and that of VIPER/metaVIPER analysis (shown in red), among which metaVIPER gives the highest value.

To provide a more systematic comparison of the improvements offered by metaVIPER analysis of single cells against approaches based on state-of-the art gene expression analysis algorithms, using the same mixture of T, B, and melanoma cells described in the previous section. Most methods designed to address the gene dropout issue in scRNA-Seq profiles are not intended to perform differential expression analysis of two individual cells but rather only of single cell subsets representing molecularly distinct clusters/subtypes [Kharchenko, 2014; Finak, 2015; Vu, 2016]. To perform this analysis, I thus quantified single cell gene expression using RSEM [Li & Dewey, 2011], which pre-assembles sequencing reads into transcripts, thus providing more accurate single cell gene expression quantification [Vallejos, 2017]. I then assessed the fraction of single cell pairs from two distinct clusters (e.g., B and T cell related) that could recapitulate differentially expressed genes and differentially active proteins, as originally detected from their corresponding bulk cell populations. For each cluster, I generated "synthetic bulk" expression profiles by averaging 100 randomly selected single cells, based on which I generated "synthetic bulk" protein activity profiles. As shown in the corresponding t-SNE plots, synthetic bulk profiles from metaVIPER-inferred protein activity analysis (Figure 3.9B) were much tighter than those produced by gene expression analysis (Figure 3.9A), suggesting that VIPER-inferred protein activity is more reproducible across samples than mRNA expression. Finally, I assessed the fraction of the 100 most differentially expressed genes and differentially active proteins (as assessed from bulk sample analysis) that could be recapitulated in a given fraction of single cells when compared to the bulk expression of a different cluster (e.g., a single T-cell vs. all cells in the melanoma cluster). As shown in Figure 3.9C, differential activity (turquoise curve) significantly outperformed RSEM-based differential gene expression analysis (yellow curve). This becomes even more evident when considering pairs of differentially expressed genes or

active proteins (e.g., gene X and Y being both differentially expressed in a single cell if they are both differentially expressed in the bulk) (Figure 3.9D). The latter is important as it supports use of metaVIPER to generate analyses similar to what is normally accomplished by FACS, using two or more markers, using any of the ~6,000 proteins assessed by the algorithm not limited by antibody availability. This is shown in Figure 3.9E - J, where virtual FACS plots are shown for critical lineage markers of these populations using gene expression (top plots) or protein activity (bottom plots). As shown, it is virtually impossible to identify cell clusters based on selected marker-pairs at the gene expression level. Indeed, most of the cells are found either on the x-axis (no detectable expression of the Y-marker) or on the y-axis (no detectable expression of the X-marker) or at the intersection of the two axes (no detectable expression of either marker). In contrast, metaVIPER analysis generates virtual FACS plots that are consistent with what would be produced by an actual FACS assay. For instance, consider CD19 and CD3, which are classic B and T cells markers, respectively. From metaVIPER analysis (Figure 3.9J), one can clearly identify a CD19+/CD3- cluster corresponding to B cells, a CD19-/CD3+ cluster corresponding to T cells, and a CD19-/CD3- cluster corresponding to melanoma cells. Yet, this is not possible when considering single cell gene expression (Figure 3.9I).

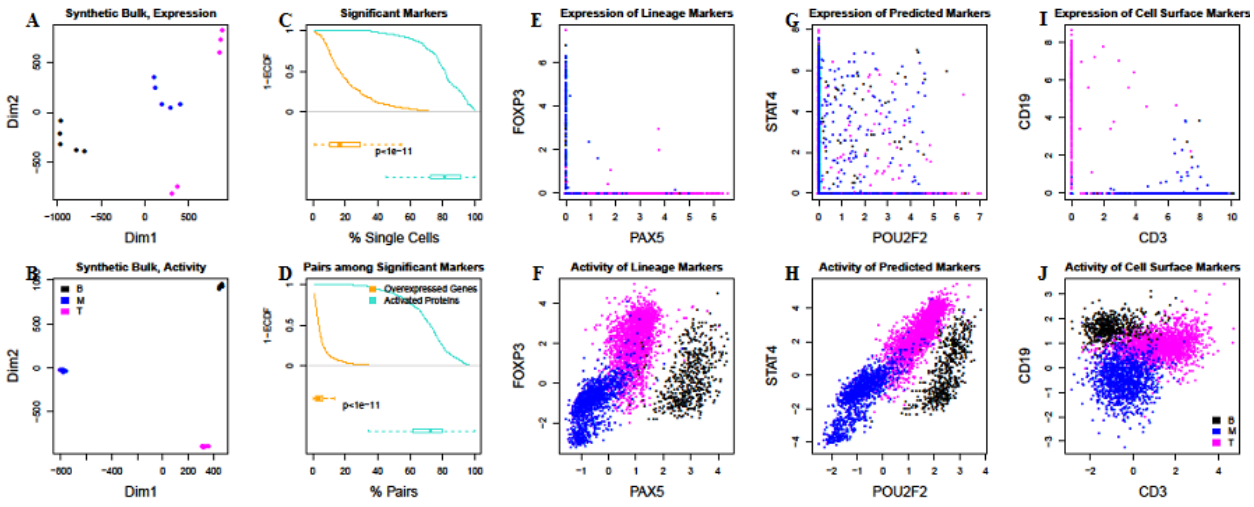

Figure 3.9 Comparative analysis of single cell metaVIPER performance compared to gene expression based methods. I identified the 100 most differentially expressed genes and differentially active proteins based on the analysis of 5 synthetic bulk samples created by averaging the expression of 100 randomly selected single cells from the melanoma, B cell, and T cell population clusters, respectively. (A, B) Based on t-SNE analysis, synthetic bulk samples clustered more tightly when analyzed based on VIPER-inferred protein activity than based on gene expression. (C). This panel shows the percent of the top 100 most differentially expressed genes/active proteins recapitulated as significantly differentially expressed/active in a given fraction of individual cells against the average expression/activity in a distinct cluster (e.g., a T cell vs the average of all B cells). The yellow and turquoise curves (1-ECDF) and boxplots (median, lower/upper whiskers and hinges) summarized the results of RSEM and metaVIPER-based analyses, respectively. (D). The same analyses were repeated to assess reproducible differential expression/activity of a gene/protein pair, as relevant for virtual FACS analyses. (F-H) Virtual FACS analyses using expression and activity of established lineage marker TFs by RSEM and metaVIPER-based analysis (see main text and Figure 2.7 for details). (I-K) Virtual FACS analysis using expression and activity of *STAT4* and *POU2F* – both identified as differentially expressed and active candidate biomarkers from bulk sample analyses, – using the same methods. (L-N) Virtual FACS analysis based on expression and activity of CD3 and CD19 cell surface markers, as used in standard FACS analyses, using the same methods.

Finally an additional value of the algorithm is that processes that are not consistent with the transcriptional regulatory architecture of the cells of interest are effectively filtered out by the interactome analysis. This is useful, for instance, in eliminating bias due to different chemistry of single-cell profiling or batch effects due to use of different gene expression quantification methodologies (Figure 3.10 and 3.11). This is helpful as these biases and batch effects represent a major obstacle to the integrative analysis of gene expression data generated in different labs or using slightly different reagent batches.

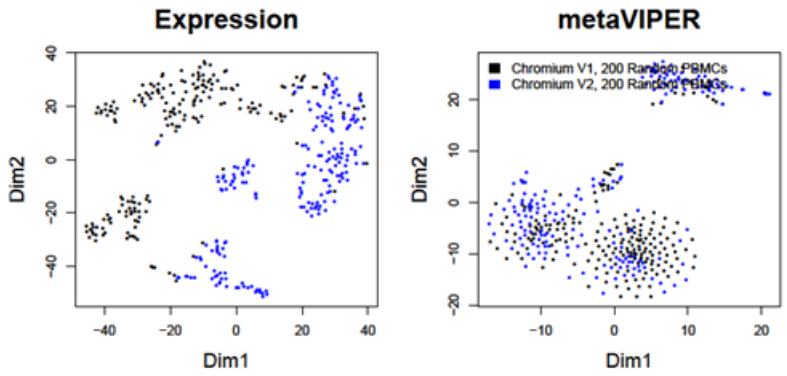

Figure 3.10 metaVIPER reduces discrepancies between different scRNA-Seq data sources. I analyzed filtered PBMC scRNA-Seq data generated using 10x Genomics V1 (Black) and V2 (Blue) chemistry. To make them comparable, I randomly selected 200 single cells for each data sources. Discrepancies between different scRNA-Seq data sources were observed using expression, while no longer seen using metaVIPER prediction.

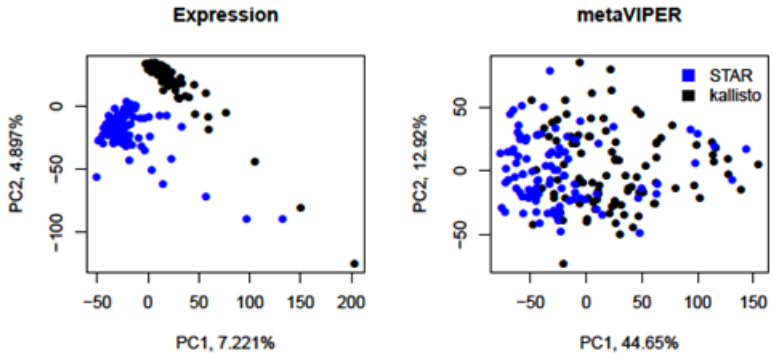

Figure 3.11 metaVIPER reduces discrepancies between different expression quantification tools. I analyzed scRNA-Seq data reported by [Wu, 2014]. Transcriptomic profiles were quantified using STAR [Dobin, 2013] (Blue) and kallisto [Bray, 2016] (Black). Discrepancies between different expression quantification tools were no longer seen using metaVIPER predicted protein activity.

Taken together, these data show that metaVIPER represents a useful methodology for the analysis of single cell data and, in particular, for the identification of lineage specific regulatory programs and lineage markers in samples comprising a heterogeneous mixture of single cells.

3.4 Discussion

I have shown that integration of multiple interactomes using an evidence integration platform (metaVIPER) can provide accurate assessment of protein activity independent of tissue lineage. By systematic, I mean that activity of 6,000 proteins can be reproducibly assessed from any tissue, independent of their gene expression; this is especially valuable in single-cell analyses. MetaVIPER can thus help infer activity of key regulators in tissues lacking a matched interactome – either due to low sample availability (orphan tissues) or to lack of tissue lineage information – as well as in highly heterogeneous single cell populations isolated from bulk tissue. I propose a specific metric ($ECDF_{|NES|}$) to assess whether a specific repertoire of interactomes is adequate for the metaVIPER analysis of an unknown or orphan tissue.

MetaVIPER is especially useful for the study of single cell biology, as its results are largely independent of sequencing depth and allow quantitative inference of protein activity even when the corresponding mRNA is undetectable. Indeed, differential activity of established lineage markers of T, B, and melanoma cells could be clearly assessed in single cells from a complex mixture, even though most of these markers were either not detected or could not be identified as statistically significantly differentially expressed at the mRNA level. The reduction in bias and batch effects is an additional advantage, allowing integration of datasets from multiple labs or generated at different times, thus addressing the important issue of single-cell data reproducibility.

Among the most obvious limitations of the method, metaVIPER cannot accurately measure activity of proteins whose regulons are not adequately represented in at least one of the available interactomes. This includes proteins whose targets are exceedingly tissue-specific within rare

tissue types and single cell sub-populations, for instance in liver hepatocellular carcinoma and testicular germ cell tumors. As more interactomes are assembled, including by ARACNe analysis of single cell data from homogeneous sub-populations, this limitation will be increasingly mitigated. This suggests that a concerted effort toward the generation of regulatory models representing distinct cellular compartments should be undertaken.

It should be noted that, while I used ARACNe as a methodology for interactome generation, there are many alternative/complementary methods to accomplish the same goal, ranging from DNA binding-site analysis [Lachmann, 2010, Lachmann, 2010], to correlation-based [Butte, 2000] and graphical-model-based [Friedman, 2004], to literature-based approaches [Kramer, 2014]. Comparison of VIPER performance using several of these methods was already discussed in [Alvarez, 2016] and is thus not repeated here. In terms of the VIPER algorithm, as also discussed in [Alvarez, 2016], alternative algorithms to transform a gene expression profile into a protein activity profile are still lacking but a thorough performance comparison can be easily performed once they become available. In general, the metaVIPER approach is independent of the specific algorithms used for either interactome reverse engineering or analysis and should thus be still fully applicable once VIPER alternatives will emerge.

I have shown that VIPER-based elucidation of MR proteins using tissue lineage-specific interactomes can effectively identify reprogramming and pluripotency factors [Carro, 2010; Kushwaha, 2015; Talos, 2017; Dutta, 2016] as well as determinants of tumor states [Aytes, 2014; Bisikirska, 2016; Carro, 2010] and resistance to targeted therapy [Rodriguez-Barrueco, 2015; Piovan, 2013]. As a result, application of metaVIPER to single-cell populations identified by cluster analysis could help identify critical determinants of lineage development as well as distinct dependencies within molecularly heterogeneous sub-population in cancer tissues. For

instance, it may help identify critical dependencies in chemoresistant cell niches, including rare tumor-initiating and tumor stem cell niches that have been shown to have poor sensitivity to standard chemotherapy and targeted therapy. Similarly, it could help identify drivers leading to aberrant reprogramming of physiologic cell states, such as recently reported in type II diabetes [Talchai, 2012].

Part IV

iterClust: a Statistical Framework for Iterative Clustering Analysis [Ding, 2018b]

4.1 Introduction

In a scenario where two clusters may exist (A and B), with B further divided into two sub-clusters (B1 and B2), the more pronounced differences between A and B may prevent subtle differences between B1 and B2 from being revealed. To solve this problem and to better describe the sub-cluster hierarchy, I propose to perform cluster analysis iteratively, such that individual clusters may be subdivided into smaller ones until further subdivisions are no longer statistically significant. Thus, for example, differences between A and B would lead to identification of two clusters in the first iteration, while B1 and B2 would be further identified in iteration 2. Previous effort in iterative clustering analysis [Usoskin, 2015] lacks systematic criteria in determining key clustering parameters, e.g. optimal number of clusters among iterations. The iterClust Bioconductor R package provides an unsupervised statistical framework for iterative clustering analysis, which can be used to, for instance discover biological heterogeneity, especially in single cell analyses of heterogeneous tissues, where cell lineages impose a relatively strong hierarchical structure, or solve general clustering problems.

4.2 Methods

4.2.1 iterClust framework

R function iterClust() performs iterative clustering analysis by organizing user-defined functions in the following workflow:

(1) **ith iteration start**
(2) **featureSelect()**, select clustering features in this iteration.
(3) **clustHetero()**, confirm observation sets to be splitted in this iteration are heterogeneous.

(4) **coreClust()**, for heterogeneous observation sets confirmed by clustHetero(), generate several clustering schemes.

(5) **clustEval()**, choose the optimal scheme given by coreClust().

(6) **obsEval()**, evaluate how each observation is clustered.

(7) **obsOutlier()**, poorly clustered observations are removed.

(8) **ith iteration end**

iterClust takes diverse feature selection methods [Saeys, 2007]; clustering algorithms, e.g. partition-based, hierarchy-based [Kaufman & Rousseeuw, 2009], density-based [Ester, 1996] and graph-based [Newman & Girvan, 2004]; and cluster/observation evaluation methods, e.g. sampling-based consensus score [Monti, 2003] or regular silhouettes score [Rousseeuw, 1987]. In addition, parameters for all user-defined functions can be set up as a function of the iteration, for instance, clustHetero() can be set up such that looser threshold parameters may be used as the iteration depth increases to deal with more and more subtle heterogeneity. In addition, featureSelect() can be used to select clustering features based on previous iterations. For instance this can help exclude features used to identify coarser clusters in prior iterations to unveil novel, more subtle heterogeneity at the current iteration. Taken together, these two functions make iterClust a highly flexible statistical framework for iterative cluster analysis. The results of iterClust are organized by iteration. Within a specific iteration, for each cluster, the corresponding observation names and clustering features are recorded, providing a comprehensive clustering trajectory.

4.2.2 Data and software availability

See Appendix D for detail.

4.3 Results

4.3.1 Running time and influencing factors of iterClust

As a statistical framework, the running time, as well as influencing factors of iterClust is majorly dependent on the clustering algorithm in coreClust() function that is specified by the user. As an example, I benchmarked iterClust on a public human PBMC (Peripheral Blood Mononuclear Cell) scRNA-Seq dataset. The original dataset was sub-sampled into different sizes, and pam() function (Partition Around Medoids, in R package cluster) was used in coreClust() function. In this case, running time increases exponentially and linearly as number of cells and genes increases, respectively, agreeing with the property of pam() function (Figure 4.1).

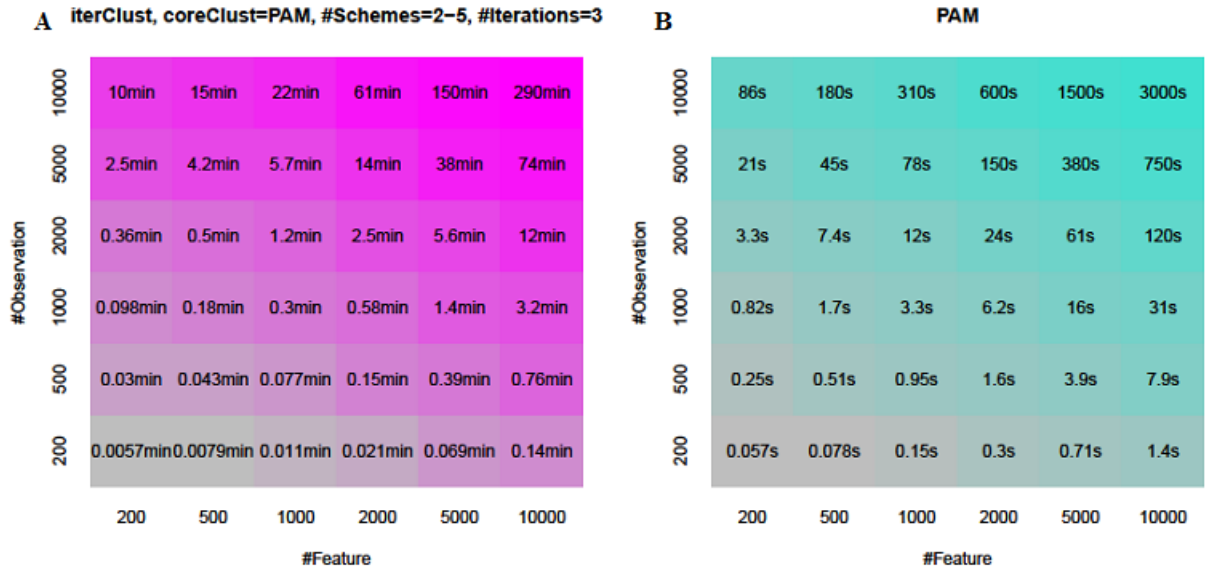

Figure 4.1 Running time and influencing factors of iterClust. Parameters used: pam() function (Partition Around Medoids, in R package cluster) in coreClust() function; k (number of clustering schemes generated within each iteration) was set as 2-5; number of iterations per-formed was set as 3. Running time increases exponentially and linearly as number of cells and genes increases, respectively (A), which agrees with the property of pam() function (B). Testing datasets with different sizes were generated by randomly subsampling cells and genes in the PBMC dataset.

4.3.2 Performance of iterClust in heterogeneity detection

As shown in Figure 4.2, within the PBMC dataset, in the 1st iteration, iterClust identified T-cell and APC (Antigen Presenting Cell) clusters. In the 2nd iteration, the algorithm further separated the original 2 clusters into additional sub-clusters, including monocyte and B-cells in the APC cluster (monocyte and B-cell are two major types of APC), as well as effector T-cell and naïve/memory T-cells in the T-cell cluster. Critically, all clusters identified by the analysis were characterized by well-established cell-type-specific gene expression (Figure 4.3). The finer grain sub-division was not the optimal solution using single pass analysis (Figure 4.4). Taken together, iterClust can correctly elucidate complex hierarchical substructures that contribute to tissue heterogeneity in PBMC single cell dataset, with more pronounced differences in starting iterations, followed by relatively subtle differences, providing a comprehensive clustering trajectory. I further confirmed these conclusions on independent scRNA-Seq datasets (Figure 4.5 and 4.6), as well as general benchmarking datasets for clustering analysis (Figure 4.7 and 4.8).

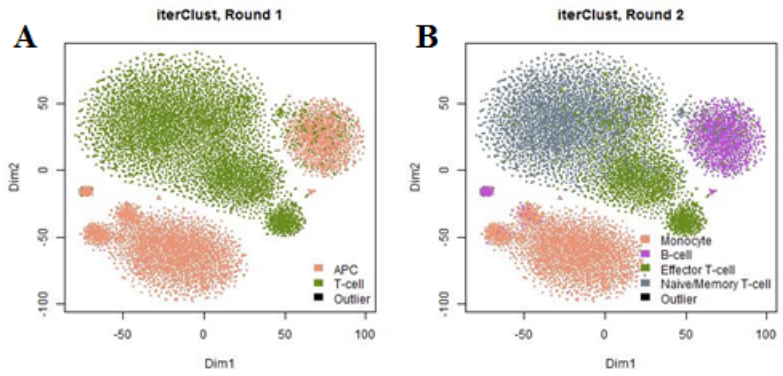

Figure 4.2 Revealing cell types within human PBMC using iterClust. For illustration purpose, the data was projected on 2D-space with t-SNE plots [Maaten & Hinton, 2008], on which iterClust discovered clusters were colored. iterClust in 1st round (A) separated two major cell types, T-cell and APC (Antigen Presenting Cell) and 2nd round (B) further dissect these clusters, separating monocyte and B-cell in APC cluster, as well as Effector T-cell and Naïve/Memory T-cell among T-cell cluster.

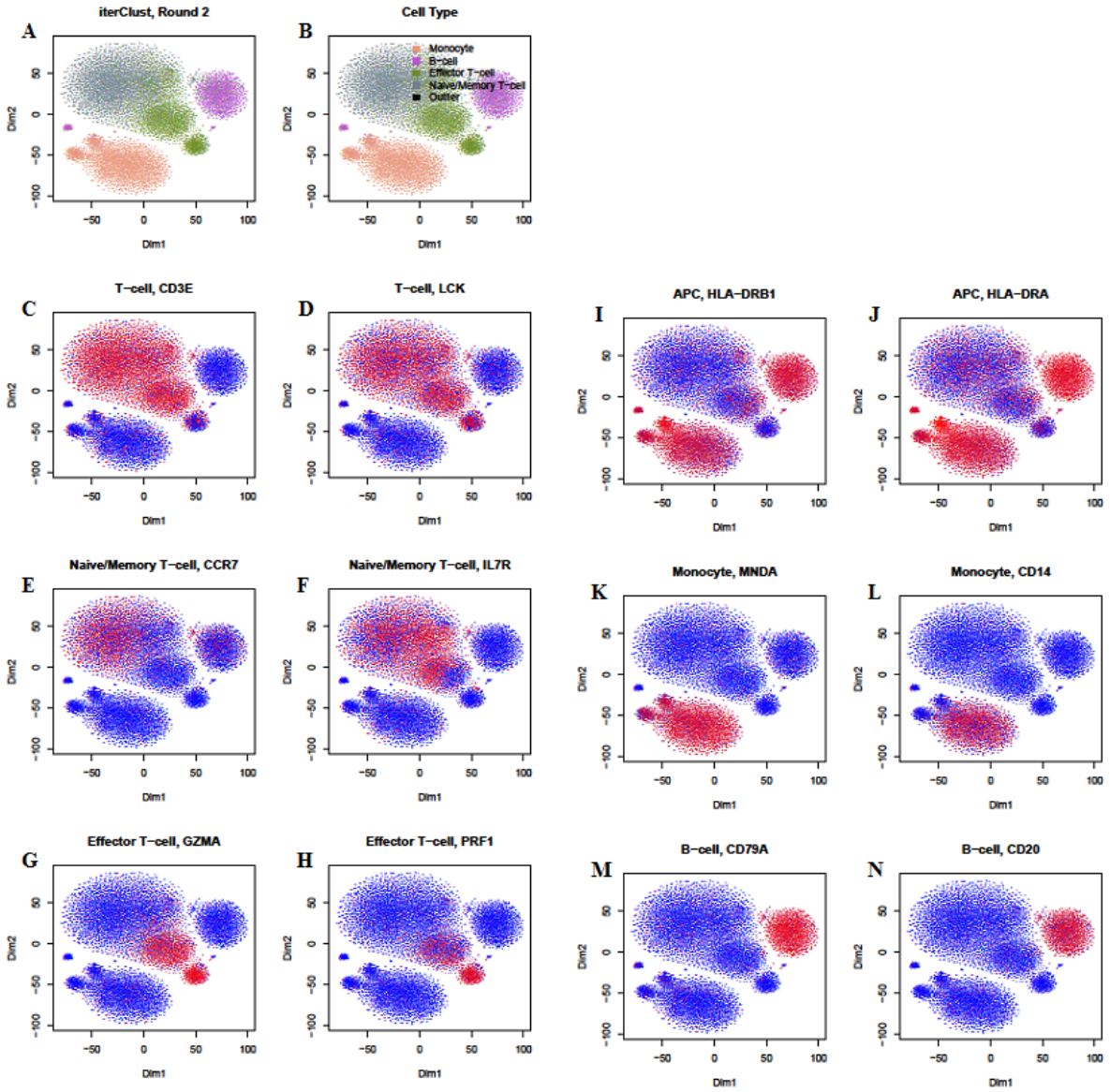

Figure 4.3 PBMC dataset lineage marker annotations. Cells were projected onto 2D-space as in Figure 4.2. Cluster IDs given by iterClust iteration 2 and annotated cell types were shown in A and B, correspondingly. (C, D, I, J) Markers for T-cells and APCs, which were separated in 1st iteration. (E-H) Markers for naïve/memory T-cells and effector T-cells, which were separated in 2nd iteration among the T-cell cluster. (K-N) Markers for monocytes and B-cells, which were separated in 2nd iteration among the APC cluster.

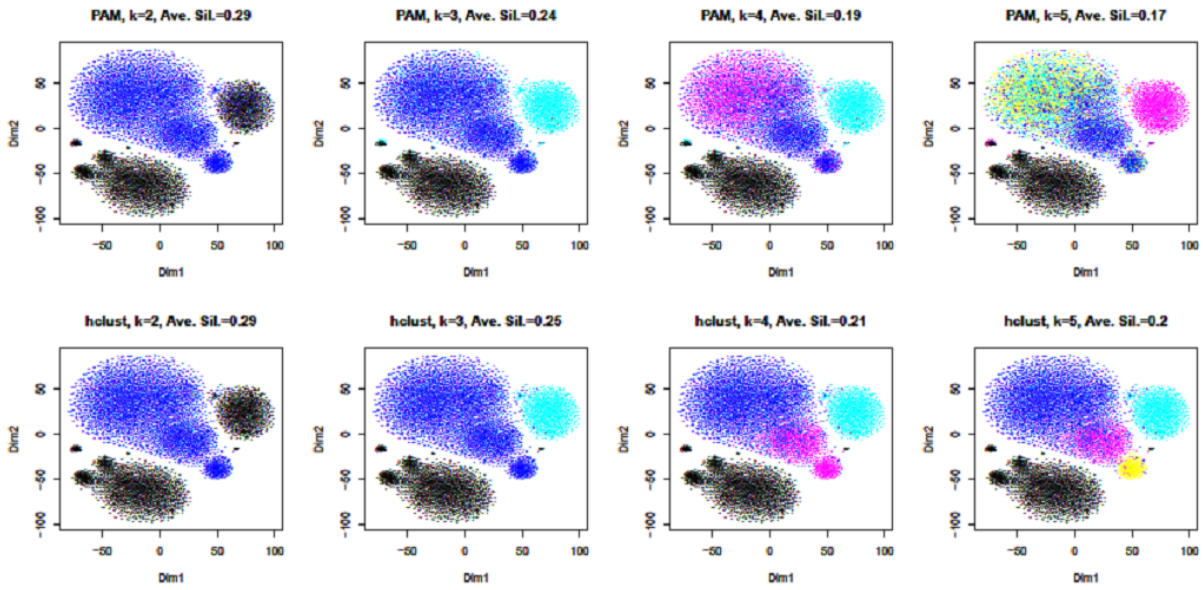

Figure 4.4 Single pass analysis on the PBMC dataset. Cells were projected onto 2D-space as in Figure 4.2. PAM (Partition Around Medoids) clustering, as well as hierarchical clustering with predefined 2-5 clusters were performed as single pass clustering methods. The optimal clustering scheme was determined by maximum average silhouette score. In both cases k=2 was considered as optimal clustering scheme, where finer grain sub-division of T-cells and APCs were not detected.

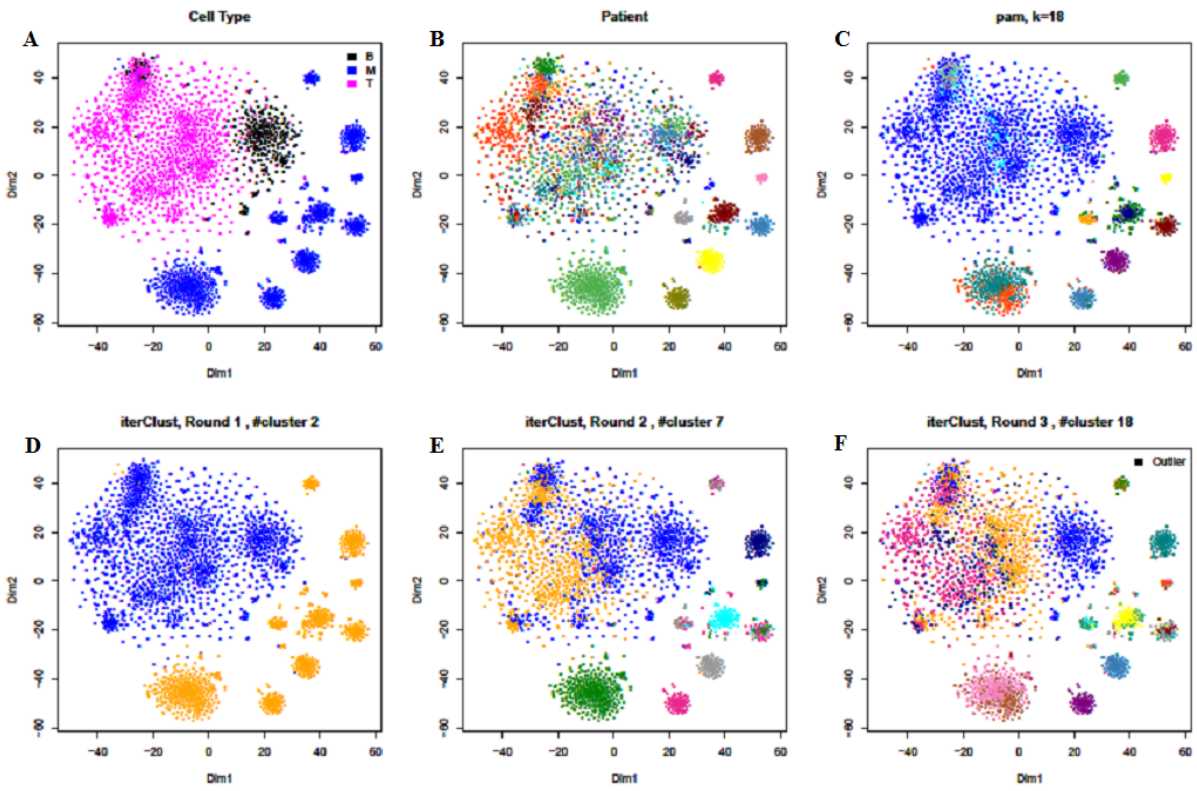

Figure 4.5 Tirosh human cutaneous melanoma (SKCM) dataset. This dataset contains single SKCM tumor cells, as well as infiltrating lymphocytes dissociated from different primary patient samples. For illustration purpose, the data was projected on 2D-space with t-SNE plots [Maaten & Hinton, 2008]. The states of tumor cells are reported to be highly patient-specific while infiltrating lymphocytes are less patient-specific but heterogeneous [Tirosh, 2016]. Dataset was visualized as in main figure. Patient ID and annotated cell types were shown in A and B, correspondingly, B: B-cell, M: SKCM cell, T: T-cell. iterClust searched 3 iterations (D-F), the 1st iteration separated immune cells from tumor cells, followed by reveling patient-specificity among tumor cells, and the 3rd iteration dissected immune cells, particularly separated B-cells from T-cells. As a comparison, result of single pass PAM clustering (used as coreClust() in iterClust) with pre-defined 18 clusters (same with iterClust round 3) was shown in C, where B and T cells cannot be separated.

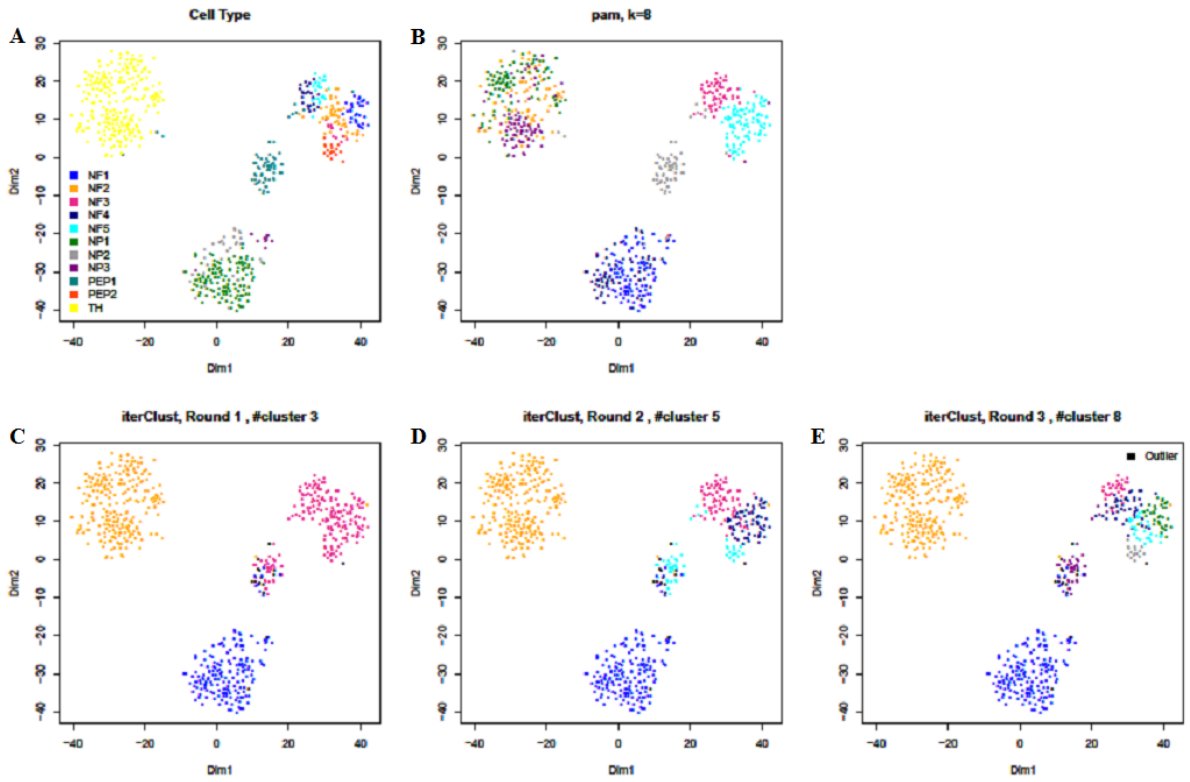

Figure 4.6 Usoskin mouse sensory neuron dataset [Usoskin, 2015]. For illustration purpose, the data was projected on 2D-space with t-SNE plots [Maaten & Hinton, 2008]. Annotated cell types were shown in A. iterClust searched 3 iterations (C-E), the 1st iteration separated TH, NP and NF-PEP cell populations, followed by revealing subpopulations among NF and PEP cells in 2nd and 3rd iterations. As a comparison, result of single pass PAM clustering (used in coreClust() in iterClust) with pre-defined 8 clusters (same with iterClust round 3) was shown in B, where NF heterogeneity were not fully appreciated, while falsely separated TH and NP cells.

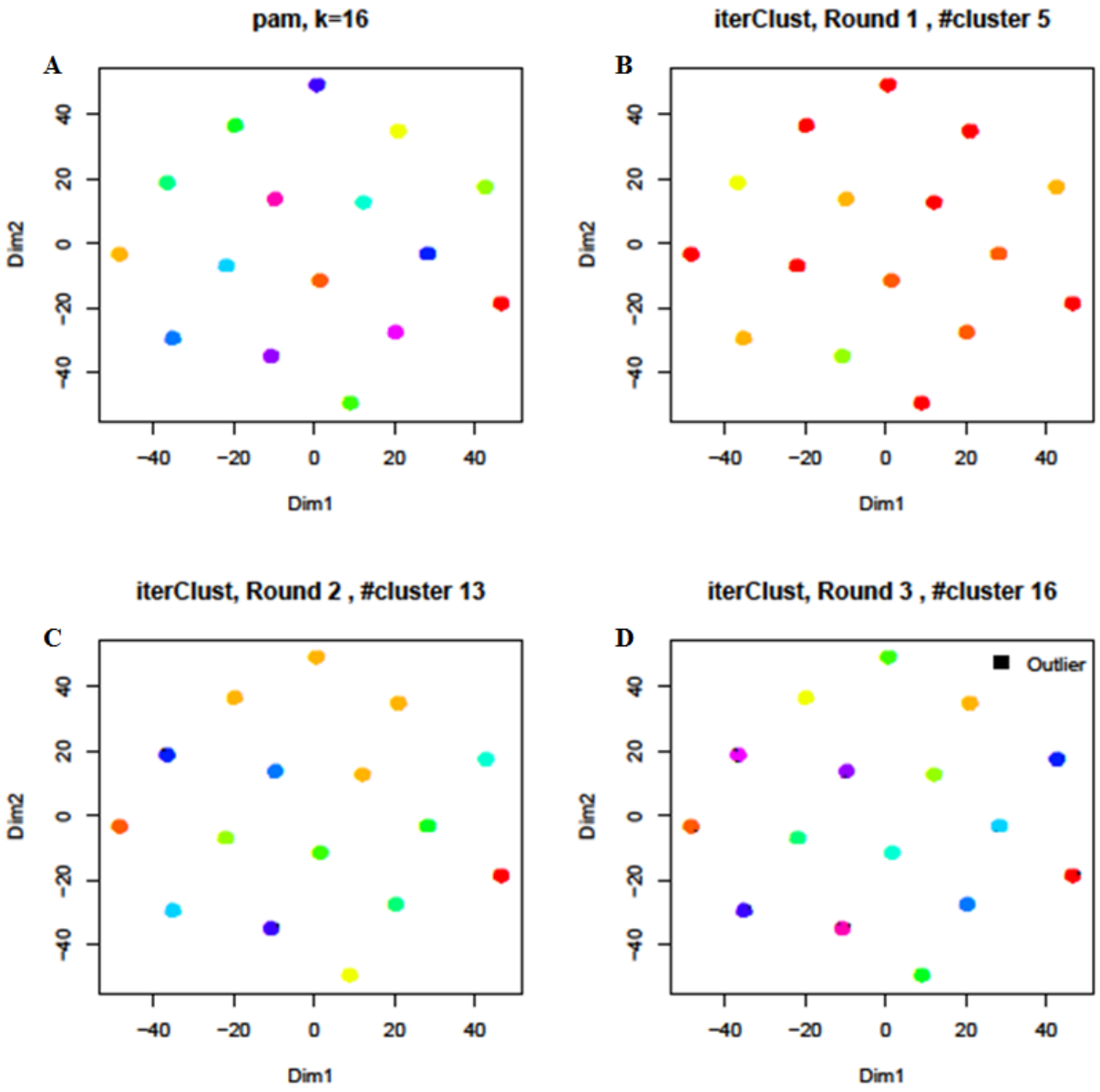

Figure 4.7 Dim1024 dataset [Fränti, 2006]. The dataset contains N=1024 observations, D=1024 features and k=16 Gaussian clusters, which represents general clustering problem. For illustration purpose, the data was projected on 2D-space with t-SNE plots [Maaten & Hinton, 2008]. iterClust searched 3 iterations (B-D), and all 16 clusters were correctly discovered at 3rd iteration. As a comparison, result of single pass PAM clustering (used in coreClust() in iterClust) with pre-defined 16 clusters was shown in A.

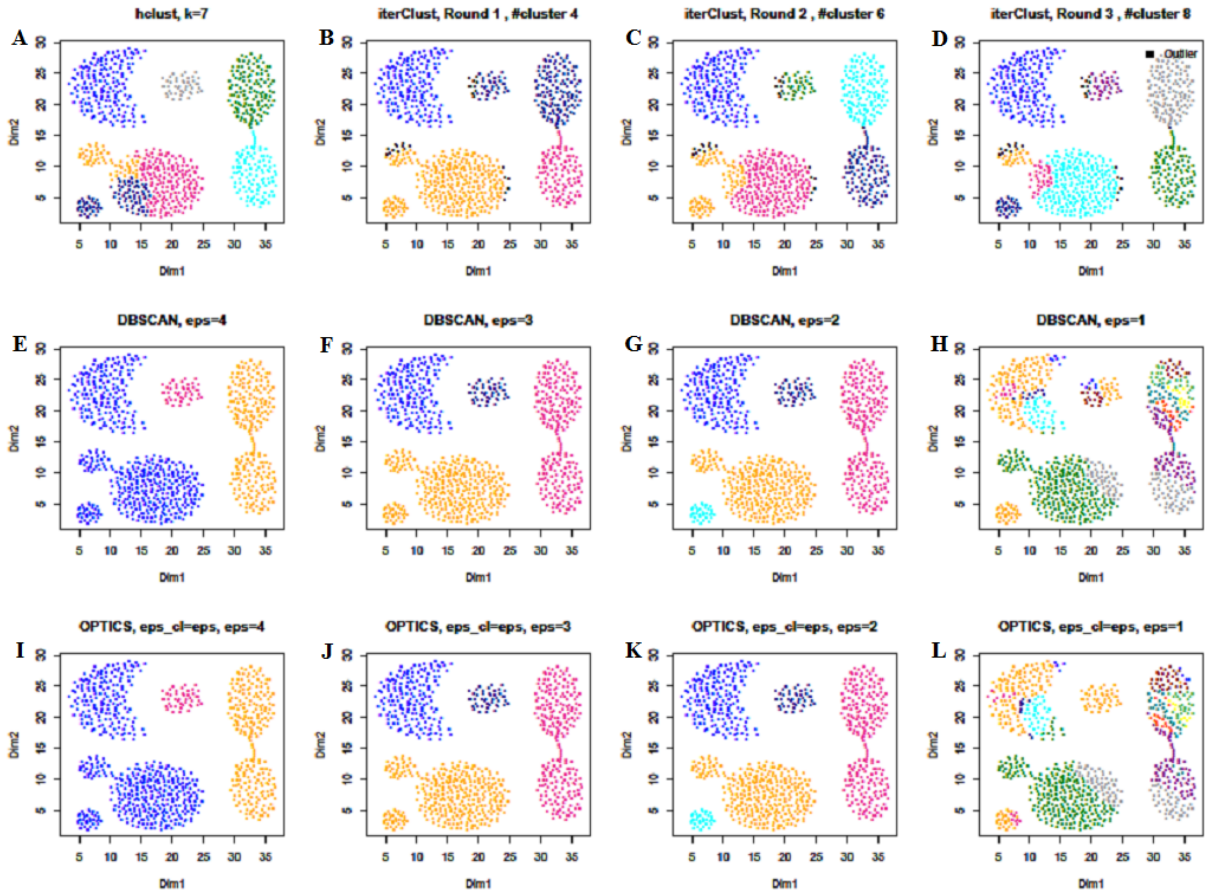

Figure 4.8 Aggregation dataset [Gionis, 2007]. The dataset contains N=788 observations, D=2 features and k=7 clusters, which represents low-dimensional clustering problem. Density-based clustering algorithms DBSCAN [Ester, 1996] and successors OPTICS, which takes into account local point density [Ankerst, 1999] were designed for low-dimensional clustering problem. Here, I compared iterClust (B-D), with single pass hierarchical clustering (used in coreClust() in iterClust, A), DBSCAN (E-H) and OPTICS (I-L).

Part V

Inter- and Intra-Tumor Molecular Subclasses of Human Glioblastoma

5.1 Introduction

Glioblastoma multiforme (GBM) is the most common malignant brain tumor in adults with unsatisfactory prognosis [Ohgaki & Kleihues, 2005]. Prognosis of GBM, as previously shown, is associated with tumor molecular features [Ceccarelli, 2016; Freije, 2004; Liang, 2005; Murat, 2008; Nutt, 2003; Phillips, 2006; Verhaak, 2010; Wang, 2017]. Based on these molecular features, GBM has been clustered in distinct subtypes based on two major competing schemes, as described in [Phillips, 2006] and [Ceccarelli, 2016; Verhaak, 2010] with the most recently correction [Wang, 2017]. I propose that the observed inter-tumor heterogeneity, as defined by these classification schemas, is the direct result of intra-tumor heterogeneity at the single cell level [Burrell, 2013]. In addition, all previously proposed classification analysis were done within specific subsets of whole transcriptome, therefore may not be effectively recapitulated by truly unsupervised clustering methods. Even more troubling, while all classification algorithms are probabilistic in nature, the specific probabilities representing the likelihood that individual tumors may belong to one of the classes, e.g. Wang Proneural (wPN), Mesenchymal (wMES) or Classical (wCL), are not generally disclosed. Indeed, as shown by my analysis, a large number of tumor samples or single cells have probabilities that are almost identical for two and even all of the three classes (Figure 5.1 and 5.2). As a result, their assignment to a specific class is at best misleading.

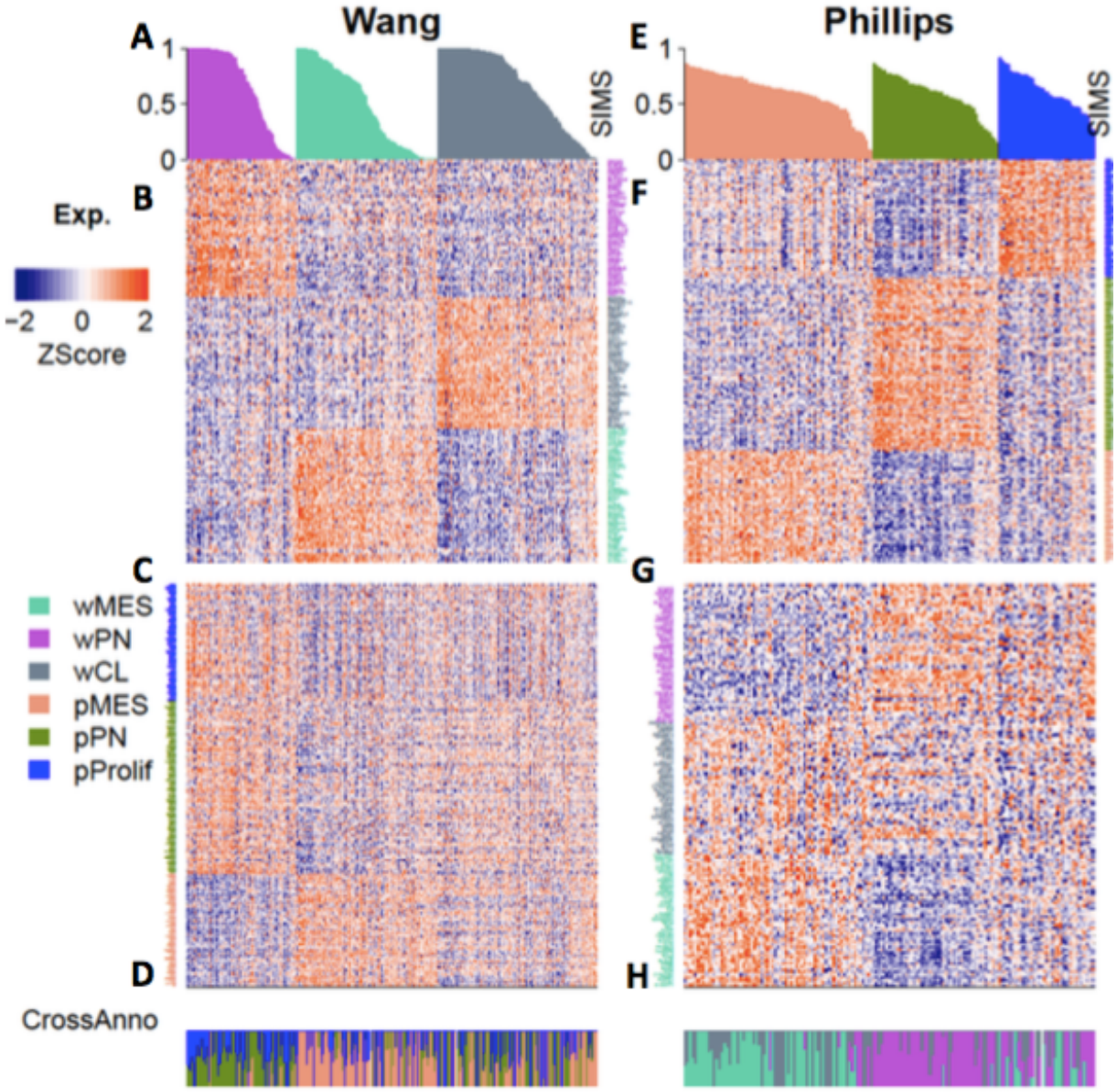

Figure 5.1 Cross-annotation between Wang and Phillips classifiers. For both Wang (A) and Phillips (A) datasets, samples were ordered according to classification confident scores (SIMS), which were determined in the original studies. The normalized expression of subtype-defining genes in their corresponding datasets (B, F) and their counterpart (C, G), was shown in the heatmaps. The cross-annotation panels showed the likelihood of being each subtype, which was measured by the relative proportion of GSEA (Gene Set Enrichment Analysis, see Methods) scores (D, H).

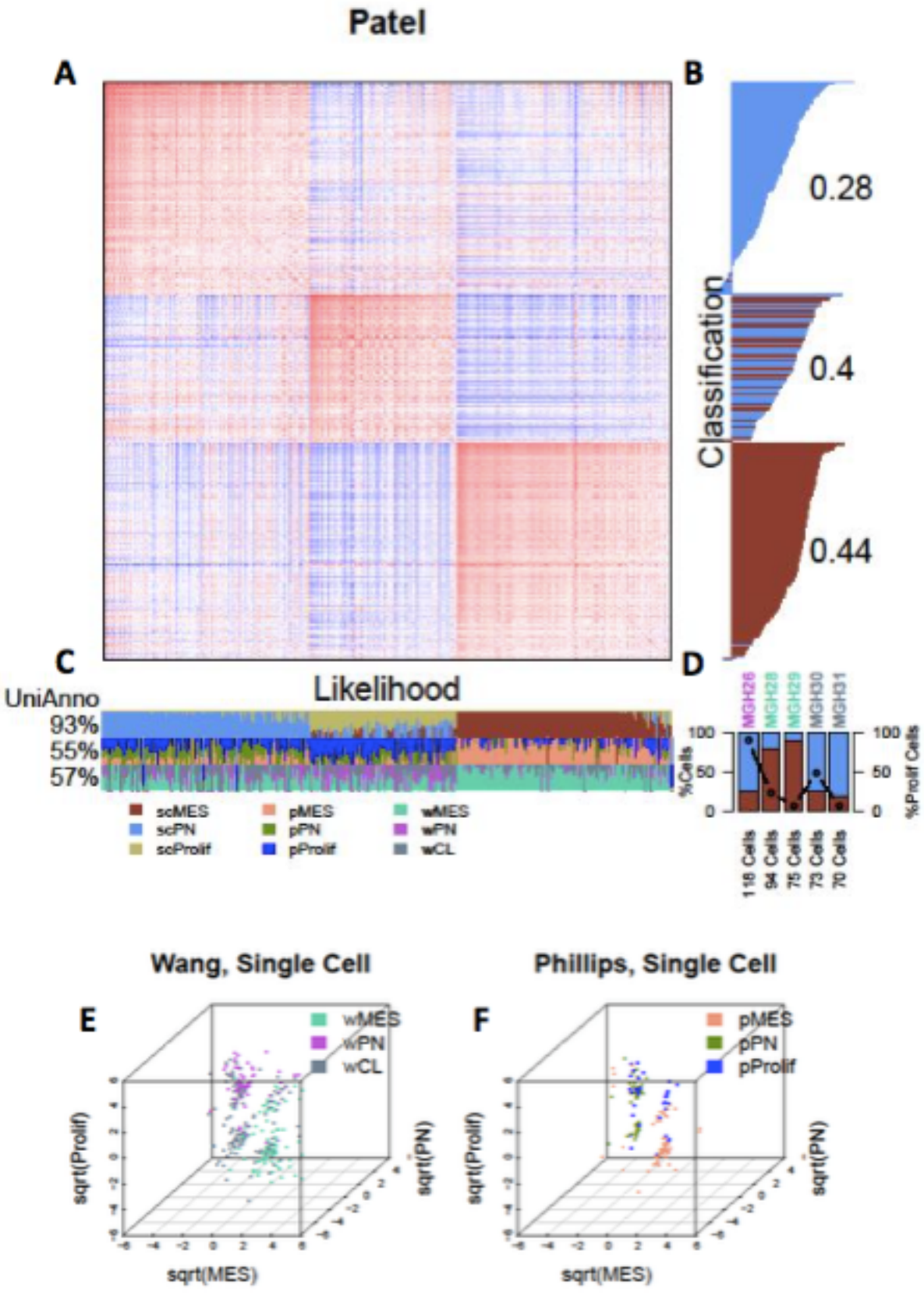
Patel
A
B
0.28
0.4
0.44
Classification
C
Likelihood
UniAnno
93%
55%
57%
scMES
scPN
scProlif
pMES
pPN
pProlif
wMES
wPN
wCL
D
MGH26
MGH28
MGH29
MGH30
MGH31
%Cells
%Prolif Cells
116 Cells
94 Cells
75 Cells
73 Cells
70 Cells
Wang, Single Cell
Phillips, Single Cell
E
F
wMES
wPN
wCL
pMES
pPN
pProlif
sqrt(Prolif)
sqrt(PN)
sqrt(MES)

Figure 5.2 Classification analysis on Patel single cell dataset. Pairwise similarity between single cells at protein activity level [Alvarez, 2016] was shown in A. Each single cell was annotated with Phillips, Wang as well as single cell subtype-specific marker sets separately (see Methods). The likelihood of being each subtype, which was measured by the relative proportion of GSEA scores, is shown in the C. Percentage of cells that were uniquely annotated was shown by UniAnno, which is considered as an evaluation of robustness of GSEA sets. The subtype with the largest relative proportion of single cell scores was considered as the single cell subtype of each cell, which was presented in B. The samples were clustered using PAM method with k = 3 as optimal scheme determined by silhouette scores (Table 5.1), which was also presented in B. Proportion of MES and PN single cells, as well as proportion of Proliferative cells in each primary patients was shown in D. Wang classification of five bulk primary patients samples, which was determined in the original study was color-coded. Annotation of Wang and Phillips datasets using derived single cell subtype-specific marker sets was shown in E and F respectively.

Table 5.1 Summary of different clustering schemes for Patel single cell dataset.

		#Cells	Mean Silhouette			#Cells	Mean Silhouette
k=2	Cluster1	236	0.32	k=3	Cluster1	158	0.28
	Cluster2	194	0.28		Cluster2	110	0.40
					Cluster3	162	0.44
	Total	430	0.30		Total	430	0.37
k=4	Cluster1	157	0.23	k=5	Cluster1	108	0.21
	Cluster2	94	0.39		Cluster2	64	0.21
	Cluster3	59	0.05		Cluster3	93	0.34
	Cluster4	120	0.32		Cluster4	57	0.03
					Cluster5	108	0.29
	Total	430	0.27		Total	430	0.23

To address some of these issues, recent work has attempted to directly tackle the issue of intra-tumor GBM heterogeneity by single cell RNA sequencing (scRNA-Seq) profiling and classification [Darmanis, 2017; Patel, 2014; Wang, 2017]. However, these approaches suffer from several limitations. First, single cells are generally profiled at low depth, making classification efforts less effective due to significant technical noise [Kolodziejczyk, 2015]. Second, most of these efforts have attempted to classify single cells according to previously

reported bulk profile schemas, rather than learning the classification from scratch. Finally, single cells taken from a human context or following passaging as neurosphere are either mixed with non-tumor subpopulations, whose identity may be different to assess, or are more likely to recapitulate specific subtypes.

To address the first two issues, I used a fully unsupervised classification methodology, based on the metaVIPER algorithm [Alvarez, 2016; Ding, 2018a], which does not require any a priori gene-set selection. Specifically, metaVIPER computes the activity of ~2,000 transcription factors (TFs) based on the expression of their transcriptional targets, as identified by the algorithm for the Accurate Reconstruction of Cellular Networks (ARACNe) algorithm. Several studies [Alvarez, 2016; Aubry, 2015; Bisikirska, 2016; Brichta, 2015; Della Gatta, 2012; Ikiz, 2015; Kushwaha, 2015; Lefebvre, 2010; Repunte-Canonigo, 2015; Rodriguez-Barrueco, 2015; Talos, 2017], including in GBM [Carro, 2010; Chen, 2014; Chudnowski, 2014; Ding, 2018a] have shown that protein activity predicted by metaVIPER or its precursor MRA (Master Regulator Analysis) from transcriptomic profiles are more robust descriptors of biological states. Indeed, fully unsupervised, metaVIPER-based single cell analysis recapitulates the TFs that were originally reported and experimentally validated as master regulators of the MES-PN transition at the single cell level [Carro, 2010]. For the third issue, I harvested tissue from PDX models established at the Mayo Clinic, by orthotopic transplantation of florescent protein labeled GBM patient tissue. I specifically selected models that had been originally classified (before transplantation) as representative of all subtypes described in [Verhaak, 2010; Wang, 2017]. This approach effectively removes contamination by microenvironment related cells, which are rapidly replaced by murine cells. Based on these analyses and on de novo classification of single cells, independent of previously reported classification schemas, I report three key findings. First,

GBM cells implement a bi-stable state characterized either by a MES or PN signature with a more or less proliferative (Prolif) phenotype, including when considering cells from samples previously classified as wCL. Second, the new analysis provides virtually unequivocal classification (>90% of single cells), compared to previous schemas (Wang and Phillips) that only unequivocally classify ~55% of individual cells. Second, previously reported bulk-tissue classification schemas can be fully recapitulated by a mixture of MES and PN cells with different degrees of proliferative potential. Finally, I report that single cell state evolves during passaging in PDX models resulting in increasingly lower single-cell heterogeneity fidelity at higher passages. Taken together, the first two findings provide a potential rationale for the current classification discrepancies, as well as a more definitive classification schema. Indeed, these findings suggest that virtually every GBM comprises a PN and a MES niche, with varying degrees of proliferative potential, thus requiring the use of subtype-targeted combination therapy supplementing the current standard of care (radiation + temozolemide), which is mostly targeting cells with the higher proliferative potential. The third finding suggests that extreme caution should be used when using PDX models of GBM, as their ability to recapitulate the original state of the patient tumor is likely affected at later passages.

5.2 Methods

5.2.1 Cell line generation

GFP-expressing derivatives of patient derived xenografts were established by transducing short-term explant cultures with lentivirus as previously described [Gupta, 2014; Gupta, 2016]. Transduction efficiency was evaluated by examining the extent of GFP positivity by fluorescent microscopy before cell expansion in an athymic nude mouse.

5.2.2 Intracranial implantation

GFP-expressing cell lines were harvested from flank tumors of athymic nude mouse and placed on Matrigel (BD Biosciences, Billerica, MA) coated plates. Once cells were adherent, GFP positivity was confirmed by fluorescent microscopy and cells were prepped for intracranial injection. Cells were stereotactically injected into the right hemisphere of each athymic nude mouse. Animals were monitored daily and sacrificed by cervical dislocation at the onset of neurological decline.

5.2.3 Tumor dissociation

Following cervical dislocation, brains were removed and tumors carved using goggles equipped with GFP visualization (BLS Limited, Budapest, Hungary). Tumors were placed into Eppendorf tubes containing DMEM media (Corning, 10-013-CV) and immediately processed for dissociation. Tumor specimens were rinsed with PBS and trypsinized (TrypLE, ThermoFisher, Waltham, MA). The tissue was then minced into small fragments using dissection scissors and placed in a 37 degree waterbath, vortexing periodically. The solution was spun at 1200 rpm for 3 minutes, trypsin was removed, and the tissue pellet was resuspended in DMEM. The specimens were aspirated into 1 mL syringes and passed through 21G and 23G needles (Cardinal Health, Dublin, OH) until no clumps were visible. This solution was then filtered using 100 and 40 μm filters (ThermoFisher, Waltham, MA). Viability and single cell state of isolated cells was conducted by trypan blue exclusion and light microscropy. Finally, cells were spun down, washed twice in PBS, and immediately submitted to Mayo Clinic's Core for Single Cell sorting.

5.2.4 Single cell RNA sequencing

The cells were counted and measured for size and viability using the Vi-Cell XR Cell Viability Analyzer (Beckman-Coulter, Brea, CA). A C1 Single-Cell Array Integrated Fluidics Circuit (IFC) for mRNA-Seq (cell size 5-10 uM, Fluidigm product number 100-5759; cell size 10-17 uM, Fluidigm part number 100-5760) was primed in the C1 Single-Cell Auto Prep System (Fluidigm product number 100-7 000). While the IFC was being primed, the lysis, reverse transcription and PCR reagents were thawed and the respective chemistries were mixed in a clean room DNA-free hood. After priming, the IFC was taken to a cell culture hood and the cells were pipetted into the IFC. The IFC was placed back into the C1 System to load and separate the cells. Once the cells were sorted into up to 96 separate chambers, the IFC was removed from the C1 System and imaged on a microscope. Cell number and viability were noted on a log sheet. The lysis, reverse transcription and pre-amplification chemistries were pipetted into the IFC in the cell culture hood. The IFC was loaded into the C1 a final time to run the mRNA-Seq script overnight.

The following morning, up to 96 individual cDNA samples were harvested from the IFC. All samples were quality control tested and quantified on a 96-capillary array Fragment Analyzer (Advanced Analytical, Ankeny, IA). Smear analyses were run using the PROSize® 2.0 software (Advanced Analytical). Only samples that passed the smear analysis thresholds were selected for library construction. Each of these samples was diluted to between 200-250 pg/ul. The Illumina Nextera XT DNA Library Preparation Kit (Illumina product number FC-131-1096) was used to create individually indexed cDNA libraries for sequencing.

5.2.5 scRNA-Seq gene expression analysis

After demultiplexing, the resulting raw reads were aligned to hg19 reference index by Bowtie2-2.2.6 [Langmead & Salzberg, 2012]. Aligned reads were sorted and indexed by samtools-1.2 [Li,

2009]. Counts matrix was measured with R package GenomicFeatures-1.24.5 and GenomicAlignments-1.8.4 [Lawrence, 2013] and TxDb.Hsapiens.UCSC.hg19.knownGene-3.2.2 [Carlson & Maintainer, 2015] from Bioconductor.

5.2.6 Microarray gene expression analysis

Informative probe clusters were assembled with the cleaner algorithm [Alvarez, 2009] and the expression data was summarized and normalized with the MAS5 algorithm as implemented in the R package affy-1.54.0 from Bioconductor [Gautier, 2004]. Differences in sample distributions were removed with the robust spline normalization procedure implemented in the R package lumi-2.28.0 from Bioconductor [Du, 2007; Du, 2008; Lin, 2008; Du, 2010].

5.2.7 Regulatory networks and protein activity profiles

All 5 brain tumor regulatory networks were reverse engineered by ARACNe [Basso, 2005; Lachmann, 2016] with 100 bootstrap iterations using 1,813 transcription factors (genes annotated in Gene Ontology molecular function database. as GO:0003700, 'transcription factor activity', or as GO:0003677, 'DNA binding', and GO:0030528, 'transcription regulator activity', or as GO:00034677 and GO: 0045449, 'regulation of transcription'). Parameters were set to 0 DPI (Data Processing Inequality) tolerance and MI (Mutual Information) p-value threshold of 10^{-8}. Protein activity profiles were inferred from expression matrix by integrating the 5 regulatory networks with metaVIPER [Alvarez, 2016; Ding, 2018].

5.2.8 Gene set enrichment analysis (GSEA)

Phillips and Wang subtype-specific highly expressed gene sets were downloaded from the original studies [Phillips, 2006; Wang, 2017]. Single cell subtype-specific activated protein sets

were built with metaVIPER-inferred protein activity profiles from dataset GSE57872 [Patel, 2014]. Only characteristic single cells (silhouette score > 0.5) were included. 10e-3 was used as p-value cutoff.

In dimension #1 (MES-PN):

MES protein set = t-test(MES, PN, alternative = "greater")

PN protein set = t-test(PN, MES, alternative = "greater")

In dimension #2 (Prolif):

Prolif protein set = t-test(Prolif, non-Prolif, alternative = "greater")

With the above-mentioned signatures, GSEA was performed with a costumed R function.

5.2.9 Data and software availability

The expression data reported in this study has been deposited in GEO under the accession number GSE107978.

5.3 Results

5.3.1 Cross-annotation between Phillips and Wang classifiers clarifies GBM inter-tumor heterogeneity

To compare the Phillips [Phillips, 2006] and Wang [Ceccarelli, 2016; Verhaak, 2010; Wang, 2017] classification schemes, in Figure 5.1, I used the published Phillips classifier to annotate Wang dataset and vice versa (see Methods). I first ordered the samples according to their classification confident scores, which were determined in the original studies (Figure 5.1A and

E). Then I showed the normalized expression of subtype-specific markers (Figure 5.1B and F), recapitulating results presented in the original studies. The expression of subtype-specific markers in cross-annotation analyses (Figure 5.1C and G) showed 1) Wang Mesenchymal (wMES) overlaps with Phillips Mesenchymal (pMES); 2) Wang Proneural (wPN) samples express both Phillips Proneural (pPN) and Proliferative (pProlif) markers, while Wang Classical (wCL) samples tend to only have higher expression on pPN markers; 3) pPN samples express both wPN and wCL markers; 4) pProlif samples express all three Wang marker sets. The cross-annotation results were further summarized in Figure 4.1D and H, where the likelihood of being each subtype was presented. Taken together, cross-annotation results indicated MES (including wMES and pMES) and PN (including wPN, wCL and pPN) are real GBM inter-tumor molecular subclasses, while Prolif potential is an independent dimension besides MES-PN heterogeneity. Specifically, high Prolif potential with PN feature constitutes wPN subtype, while low Prolif potential with PN feature constitutes wCL and pPN subtypes.

5.3.2 Classification analysis with single cells reveals GBM intra-tumor heterogeneity

Previously studies in Dr. Califano lab have shown that transcriptional regulators activity profiles inferred from gene expression are more robust descriptors of biological states, especially for single cells, whose transcriptomic profiles are usually confounded by high level of technical and biological noises [Alvarez, 2016; Ding, 2018]. Therefore, to better study the molecular subclasses of GBM tumor cells, I converted published GBM single cell expression data obtained from primary tumor classified as all three Wang subtype (Figure 5.2D) [Patel, 2014; Wang, 2017] to protein activity profiles (see Methods) using metaVIPER integration of 5 available GBM ARACNe networks (Table 5.2). PAM (Partitioning Around Medoids) clustering was performed, and silhouette score suggested k=3 to be the optimal scheme (Table 4.1). To further study the

characters of each single cell subclasses, I used published Phillips and Wang classifiers to annotate each cluster (Figure 5.2A and C). Result showed two clusters were annotated as MES-like (annotated as wMES or pMES) and PN-like (annotated as wPN/wCL or pPN), while the third cluster was composed of cells showed high Prolif potential, suggesting regardless of inter-tumor subtypes, GBM is essentially composed of MES-like and PN-like tumor cells (I named them MES and PN single cells) with varying Prolif potential. I confirmed the above conclusions by analyzing an independent GBM scRNA-Seq dataset (Darmanis, 2017; see Figure 5.3A). Based on this, I built subtype-specific activated protein sets from single cells (see Methods) and previously validated MES-associated master regulators were recapitulated in MES-specific activated regulator sets, including *STAT3, C/EBPB, FOSL2, bHLH-B2* and *RUNX1* [Carro, 2010].

Table 5.2 Summary of used brain tumor regulatory networks.

Expression Source	# Samples	# Regulators	# Targets	# Interactions
TCGA RNA-Seq (Verhaak)	166	1811	18354	259025
REMBRANDT	804	1265	11909	329695
Phillips	176	835	8263	252082
TCGA affymetrix	202	938	9162	272520
TCGA agilent	202	1675	16115	748366

To further associate GBM intra- and inter-tumor subclasses, I first annotated Wang and Phillips samples with classifier derived from above single cell analysis (Figure 4.2E and F, Figure 5.4). Consistent with cross-annotating analysis of Wang and Phillips datasets in previous session, MES (wMES and pMES) and wPN/wCL/pPN tumors are essentially MES and PN subtypes, while wPN/wCL tumors are PN with high/low Proliferative potential, respectively. I then, on the other hand, decomposed bulk primary tumors by showing the proportion of MES and PN cells, as well as proportion of Proliferative cells described on independent dimension besides MES-PN

heterogeneity, in each patient (Figure 5.2D). Results suggested that wMES patients (MGH28 and 29) are dominated (>80%) by MES cells, while wPN and wCL patients (MGH26, 30 and 31) are dominated (>70%) by PN cells. Results also suggested wPN patient (MGH26) has higher proportion of Prolif cells (~90%) compared to other patients (on average ~30%). Taken together, these results suggested inter-tumor MES and PN subtypes are caused by dominating proportion of MES and PN tumor cells, respectively. Also, consistent with cross-annotating analysis of Wang and Phillips datasets in previous section, usually MES tumors are less Proliferative, while wPN/wCL tumors are caused by high/low proportion of Proliferative PN tumor cells, respectively.

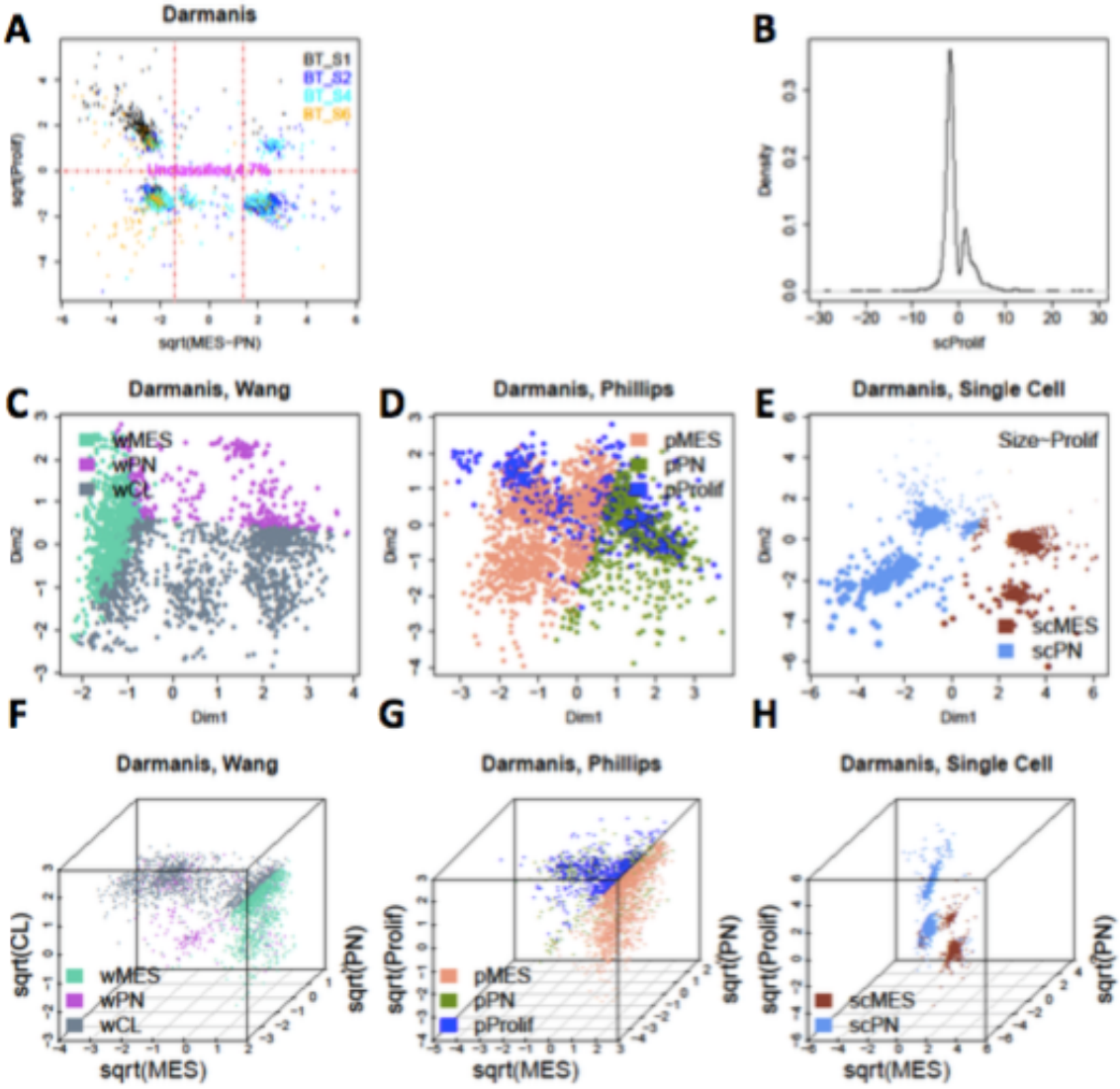

Figure 5.3 Classification analysis on Darmanis single cell dataset. A) Single cells were presented on a scatterplot showing heterogeneity and proliferating. Specifically, the distribution Prolif scores was showed in B. (C-E) Single cells were projected on 2-D space according to their GSEA scores towards Phillips subtypes (C), Wang subtypes (D), and single cell features (E) using MDS respectively (F-H, corresponding original 3-D space before projection), and color-coded according to subtypes determining by the corresponding classifiers. Prolif potential of cells was represented by the size of dots (D).

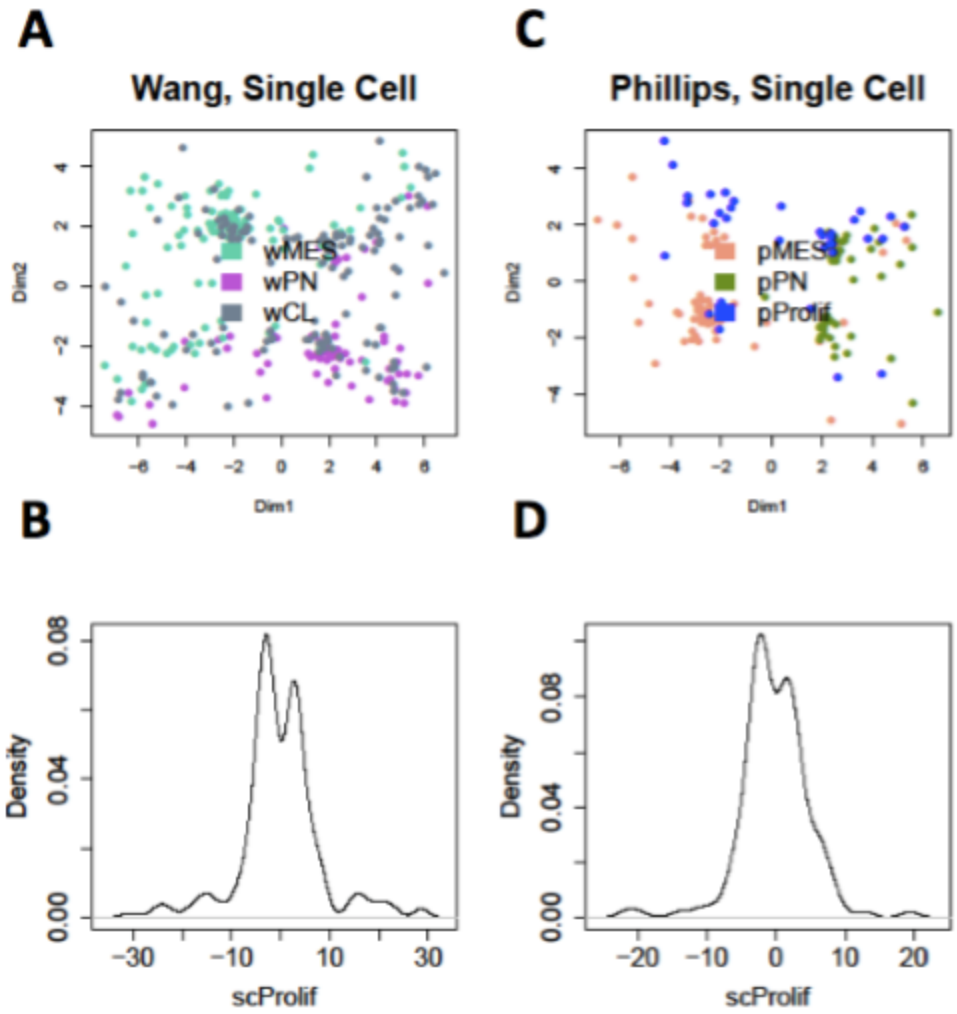

Figure 5.4 Classification analysis on bulk datasets. Wang and Phillips dataset was visualized using 2-D MDS-projection. Specifically, the distribution of Prolif scores was showed in B and D, respectively.

5.3.3 Evaluation of Phillips, Wang and single cell classifiers

To evaluate the performance of Phillips, Wang and single cell classifiers, I measured their capability in giving mutual-exclusively annotation towards single cells in [Patel, 2014] dataset. Specifically, on scatterplots I show the GSEA scores towards Wang (Figure 5.5A-C) and Phillips subtypes (Figure 5.5D-F), in a pairwise manner. Both Wang and Phillips classifiers gave "double

positive" cells, e.g. wMES-wPN cells in 1st quadrant of Figure 5.5A, and pMES-pPN cells in 1st quadrant of Figure 5.5B. One the other hand, with my single cell classifier, ~95% single cells either fall into 2nd or 4th quadrant of Figure 4.5G, suggesting they were mutual-exclusively annotated as PN or MES, which is consistent with my previous conclusion that MES and PN are real subtypes of GBM tumor cells. Since I also concluded Prolif potential as independent dimension besides MES-PN heterogeneity in describing GBM tumor cells, I visualized the two dimensions in Figure 5.5H, giving a comprehensive description of the tumor cell states.

I also projected single cells on 2-D MDS (Multi-Dimensional Scaling) space according to their GSEA scores towards Wang, Phillips, and my single cell classifiers, respectively (Figure 5.6B, D and F. Positions of cells in original 3-D GSEA space before MDS-projection were shown in A, C and E). Ideally, a good classifier should give clear separation of single cells on 2-D space, agreeing with their corresponding subtypes determined by this specific classifier, e.g. the single cell classifier, where the 1st dimension denoted MES-PN heterogeneity and 2nd dimension presented Prolif potential on 2-D MDS plot (Figure 5.6F). In contrast, annotation given by both Phillips and Wang classifiers didn't agree well with 2-D cell projection, e.g. in Wang classifier, annotated wMES cells subdivided into two groups, among which one group co-segregated with annotated wPN cells and the other group co-segregated with annotated wCL cells (Figure 5.6B); in Phillips classifier, for the two major cell clusters, one contained cells annotated as all Phillips subtypes, and the other one was composed of pPN and pProlif cells (Figure 5.6D). Taking together, results suggested single cell classifier gave clear classification thereby outperformed the other two. I further confirmed the above conclusions by analyzing an independent GBM scRNA-Seq dataset (see Figure 5.3B-H) [Darmanis, 2017].

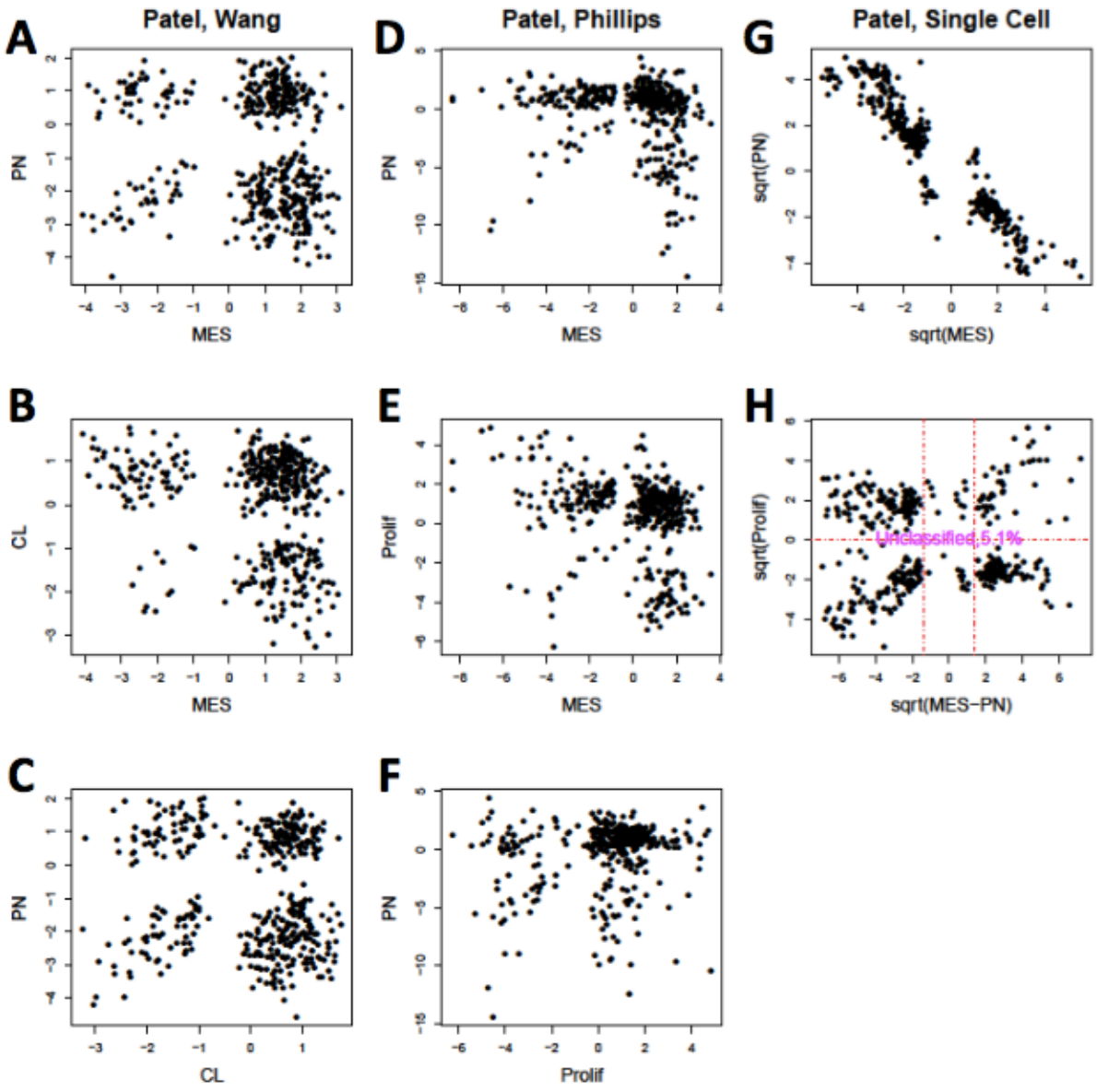

Figure 5.5 Evaluation of Phillips, Wang and single cell classifiers. GSEA scores towards Wang subtypes (A-C) and Phillips subtypes (D-F) were shown in a pairwise manner on scatterplots. For single cell classifier, GSEA scores towards PN and MES subtypes were shown in G, and projections of single cells onto Prolif potential dimension (GSEA score of Prolif) and MES-PN heterogeneity dimension (GSEA score towards MES minus that of PN) were shown in H. If the absolute projection on heterogeneity dimension was smaller than 2 (within the two vertical lines, corresponds to p-value < 0.05), then the cell was considered as "unclassified". Horizontal line denoted Prolif-score equals to 0. Cells reside above which were considered to have proliferating potential, and vice versa.

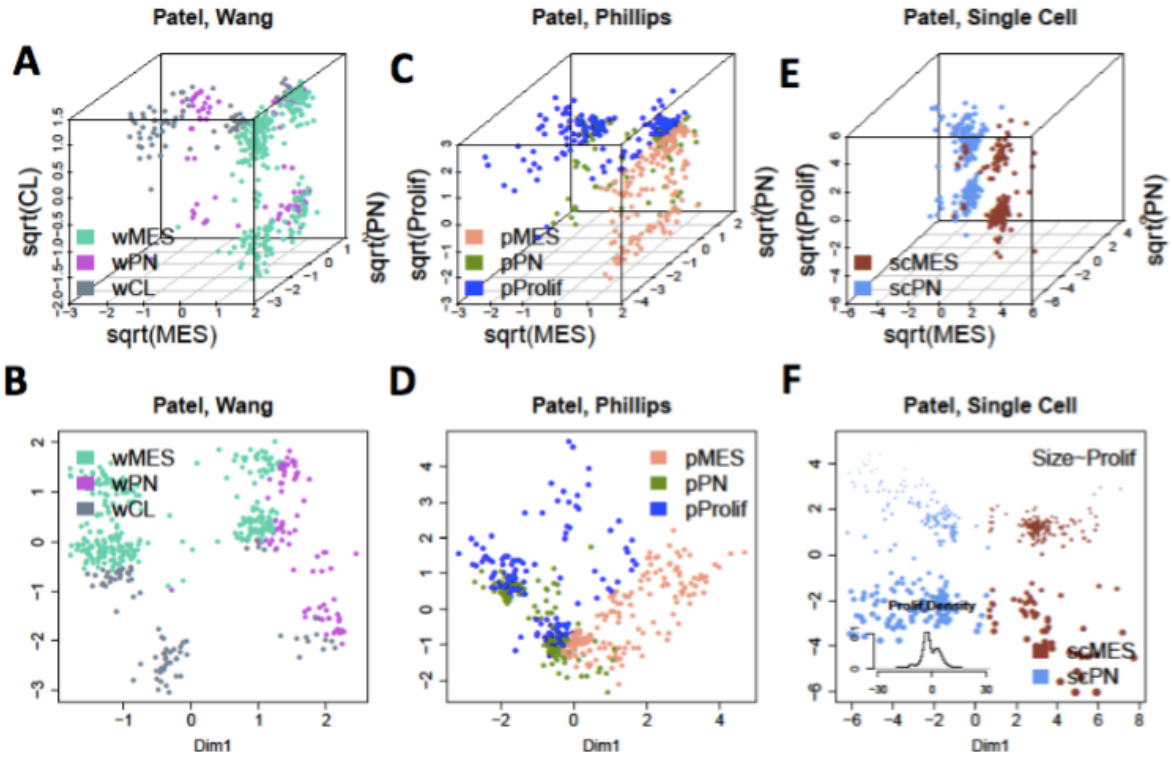

Figure 5.6 Patel dataset visualized in original 3-D GSEA space and 2-D MDS- space.

5.3.4 "Passage-effect" when passaging GBM in mouse primary xenograft models

GBM mouse patient derived xenograft (PDX) models, which are established by engrafting patient tumor specimens into the flank of nude mice and subjected to subsequent serial passaging, are important for advancement of basic and translational biology [Carlson, 2011]. Therefore it's critical to examine 1) whether PDX models preserve GBM heterogeneity and 2) whether serial passaging deviates tumor features. To explore both points, Dr. Sarkaria lab generated scRNA-Seq profiles from mouse PDX models (Mayo dataset, Table 5.3, Figure 5.7A and B). I annotated these single cells with Phillips, Wang and previously built single cell classifier (Figure 5.8). Followed the same analysis done in Figure 5.6, I showed only single cell classifier gave clear separation of the profiled cells (MES and PN), confirming its superior performance against

Phillips and Wang classifiers, as well as GBM is essentially composed of MES and PN cells, further suggesting mouse PDX models preserve both MES and PN tumor single cells. However, as passaged in mouse PDX, I observed decreasing MES proportion (Figure 5.7C) and increasing Prolif potential, compared to primary cells (Figure 5.5H and Supplemental Figure 5.3A where ~60% profiled cells were annotated as non-Prolif cells), suggesting deviation in GBM molecular feature when undergoing serial passaging.

Table 5.3 Summary of processed GBM samples provided by Mayo Clinic.

Sample ID	Source	Data Type	Verhaak Subtype	Passage	#Cells
G12	PDX	scRNA-Seq	NL	>P10	87
G22	PDX	scRNA-Seq	NL	P6	52
G38	PDX	scRNA-Seq	CL	P1	56
G38	PDX	scRNA-Seq	CL	P6	31
G43	PDX	scRNA-Seq	MES	P15	43
G84	PDX	scRNA-Seq	NL	P3	64
G85	PDX	scRNA-Seq	PN	P6	71

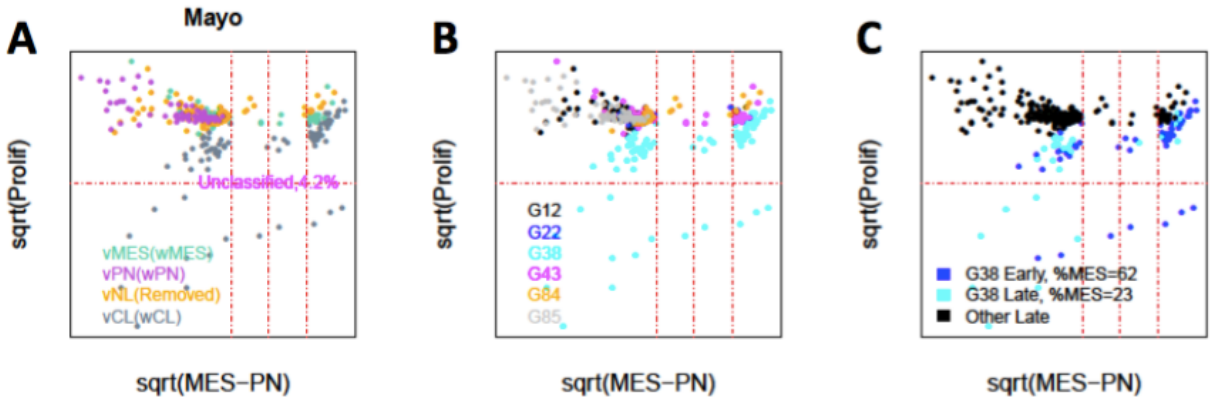

Figure 5.7 Annotating single cells obtained from mouse PDX models. (A) Profiled single cells (Table 4.3) were presented on scatterplots. Verhaak and Wang classification of primary patients where the mouse PDX models were derived (A), IDs (B) and passage status (C) were also shown. Specifically, in C, I highlighted the early (P1) and late (P6) passages for Classical mouse PDX model G38, and the corresponding proportion of MES cells.

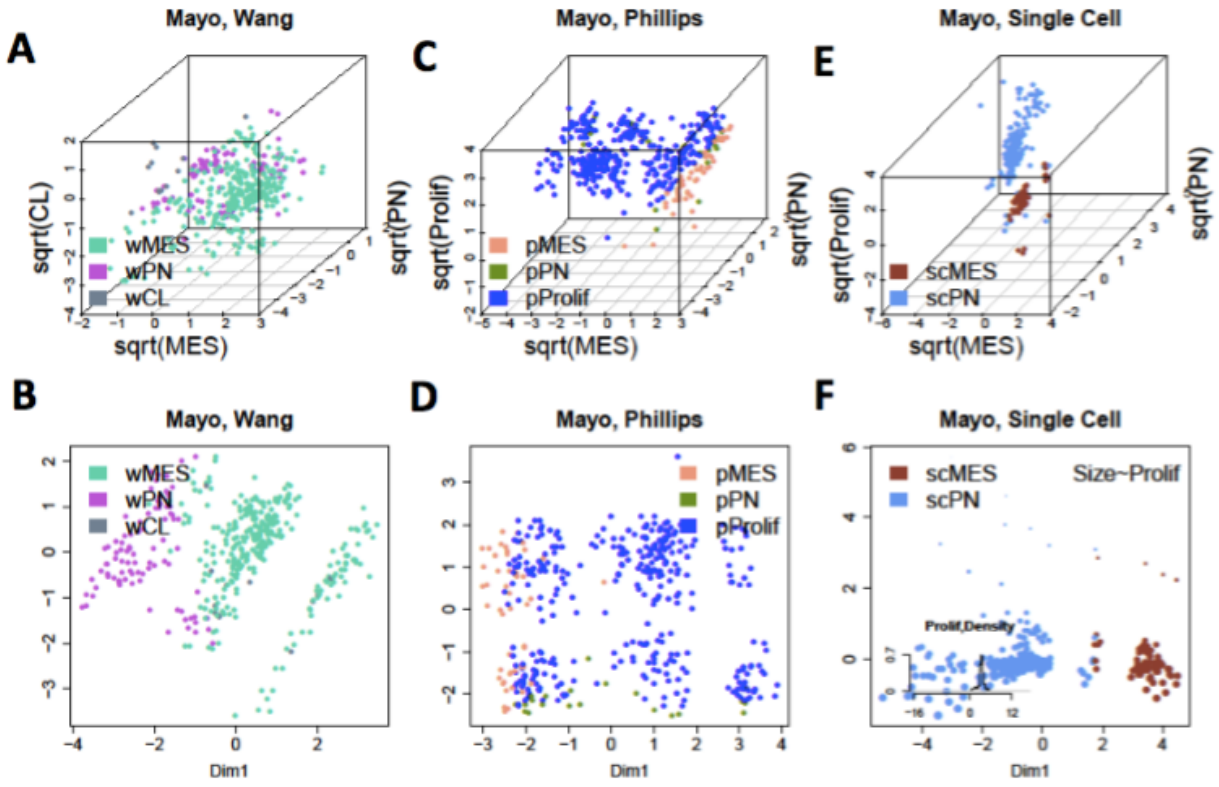

Figure 5.8 Mayo dataset visualized in original 3-D GSEA space and 2-D MDS- space.

5.4 Discussion

Taken together, the study suggested that GBM is composed of MES and PN tumor cells with different Prolif potentials, these cells mix at varying ratio and cause different inter-tumor features (Figure 4.9A). Considering previous study in Califano lab `suggested MES and PN are interchangeable by perturbing five master regulators [Carro, 2010], therefore therapeutic strategies should target MES and PN tumor cells at the same time as well as considering the Prolif potential, and these findings would be great value for developing novel therapeutic strategies. I also showed “passage-effect” (Figure 4.9B) that molecular features of GBM drifts as

mouse PDX passaged, suggesting mouse PDX model should be used with cautious and early passage is preferred.

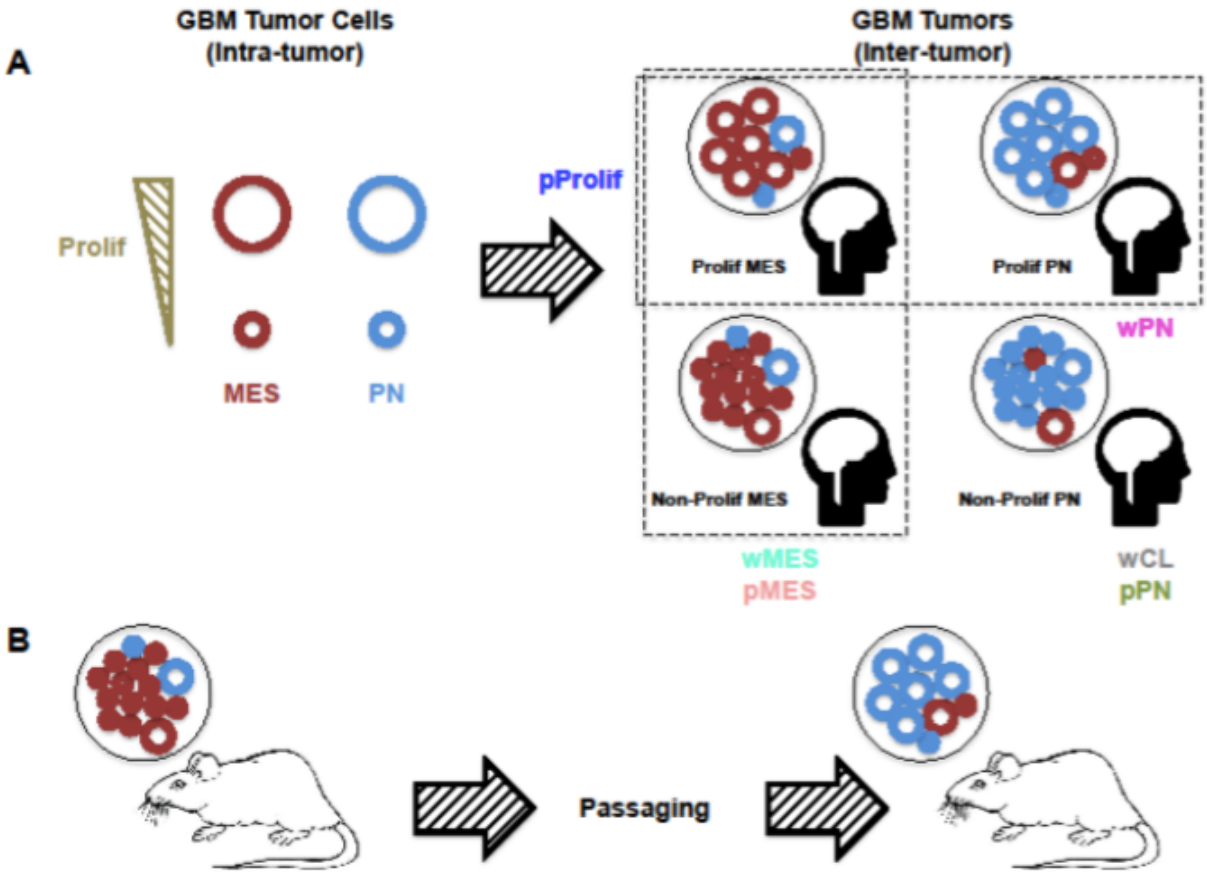

Figure 5.9 Graphic summary. (A) Inter-/intra-tumor heterogeneity: GBM is composed of MES and PN tumor cells. Besides heterogeneity, tumors cells also have various Prolif potential, which is considered a second dimension of tumor features. These single cells mix at various ratios thereby causes GBM inter-tumor heterogeneity. Specifically, tumors dominated by MES cells are considered as wMES or pMES subtypes, tumors dominated by cells with high Prolif feature are considered as pProlif subtypes, tumors dominated by Prolif PN cells are considered as wPN subtypes, and tumors dominated by non-Prolif PN cells are considered as wCL or pPN subtypes. (B) "Passage-effect": PN cell proportion and Prolif potential increases as mouse GBM PDX model is passaged.

Part VI

Elucidating Regulators of β-cell De-Differentiation in Type-2 Diabetes

6.1 Introduction

Type 2 Diabetes (T2D) is associated with defective β-cell insulin secretion and subsequent reductions of β-cell mass. Unveiling the mechanisms that control β-cell dysfunction is of fundamental importance towards understanding the pathophysiology of human T2D [Prentki & Nolan, 2006]. Conventionally, apoptosis has been thought of as a major pathway to loss of β-cell mass in T2D [Butler, 2003]. Recent studies using rodent models have shown that dedifferentiation of mature β-cells into endocrine progenitor-like cells can also play a role in this process [Talchai, 2012]. However, it is unclear whether similar de-differentiation process can cause β-cell failure in human T2D. To solve this problem, I combined single cell RNA sequencing (scRNA-Seq) technique [Kolodziejczyk, 2015] with master regulator analysis [Califano & Alvarez, 2017] to have higher resolution characterization of β-cells under de-differentiation state, as well as discovering de-differentiation related master regulators. Dr. Son also performed functional validation of the discovered mater regulators by testing whether they can reprogram health β-cells to de-differentiated cells, or even α-cells.

6.2 Methods

6.2.1 Sample collection and islet dissociation

In total Dr. Son collected islet samples from 11 donors, see Table 6.1 for detail. Human primary Islets were dissociated into single cells by adding 0.025% trypsin solution.

Table 6.1 Sample information

Donor ID	Sex	Age	BMI	HbA1C
Normal20151026	M	25	26.0	NA
Normal20160224	F	51	20.1	NA
Normal20160310	F	53	34.0	NA
Normal20160427	F	36	23.5	5.00%
Patient20160121	M	48	43.7	6,60%
Patient20160202	M	51	24.6	6.90%
Patient20160303	F	42	27.6	6.70%
Patient20160314	M	59	32.5	6.60%
Patient20160707	M	58	39.3	8.90%
Patient20160712	M	59	31.6	9.60%
Patient20160714	F	62	30.4	6.60%

6.2.2 Enrichment of β-cells

The freshly dispersed islet cells were submitted to auto fluorescence-activated cell sorting using an influx cell sorter (BD Biosciences), illuminating the cells at 488 nm, so that the emission at 510-550 nm could be taken as a parameter for their flavin adenine dinucleotide (FAD) content. Selection of appropriate gates allowed the simultaneous isolation of single beta and single non-beta islet cells since β-cells displayed higher FAD fluorescence than single non-β cell.

6.2.3 Over-Seq experiment

Over-Seq was designed for gain-of-function studies, testing whether putative de-differentiation master regulators can reprogram health β-cells to de-differentiated cells, or even α-cells. In contrast to Cas9-mediated knockout, Cas9-mediated gene activation in the endogenous locus often results in variable and non-robust outcomes in different systems, possibly due to difficulty of accessing compacted chromatin of transcriptionally inactive genes (heterochromatin) [Gilbert, 2014]. Therefore Dr. Son generated tools to express bicistronic candidate genes and BFP reporter

followed by a unique 18-nt barcode. Similar to Perturb-Seq [Dixit, 2016], by introducing a mixture of viruses with individual barcodes for each cDNA in individual cells, I can determine the effect of overexpression on β-cell fate conversion. To facilitate the cloning of these libraries, Dr. Son used pENTR gateway system (Thermo Fisher) that enables systematic and simple insertion of candidate cDNA into adenovirus expression plasmids.

6.2.4 Single cell RNA sequencing

Primary single cells were processed following standard C1 Single-Cell Auto Prep System protocol, using medium (10-17 uM) 800-capture-sites C1 Single-Cell Array Integrated Fluidics Circuit (IFC). Over-Seq perturbed single cells were processed following standard 10x Genomics Chromium protocol.

6.2.5 scRNA-Seq gene expression analysis

For primary single cells profiled by C1 system, after demultiplexing, the resulting raw reads were aligned to hg19 reference index by Bowtie2-2.2.6 [Langmead & Salzberg, 2012]. Aligned reads were sorted and indexed by samtools-1.2 [Li, 2009]. Counts matrix was measured with R package GenomicFeatures-1.24.5 and GenomicAlignments-1.8.4 [Lawrence, 2013] and TxDb.Hsapiens.UCSC.hg19.knownGene-3.2.2 [Carlson & Maintainer, 2015] from Bioconductor. For Over-Seq perturbed single cells profiled 10x Genomics Chromium, gene expression profiles were quantified using standard 10x Genomics Cell Ranger pipeline.

6.2.6 Over-Seq barcode quantification

10x Genomics Cell Ranger pipeline allows multi-mappers, therefore not able to give precise quantification of Over-Seq barcodes, due to identical regions among the barcode sequences. To

solve this problem I used customized analysis pipeline to quantify Over-Seq barcodes. Since Chromium follows similar barcoding and library construction strategy as scPLATE-Seq (the only differences are length of cell barcode, 16nt for Chromium and 8 for scPLATE-Seq, and length of UMI, 10nt for Chromium and 8nt for scPLATE-Seq), therefore I used scPLATE-Seq analysis pipeline for demultiplexing and UMI quantification with correct cell barcode and UMI length parameters. After demultiplexing, reads were aligned and mapped to UCSC hg38 genome using the STAR 2.6.1a [Dobin, 2013] with option --quantMode TranscriptomeSAM so that output alignments are translated into transcript coordinates for downstream analysis, --clip5pNbases 10 so that the UMI sequences are not mapped, and --outFilterMultimapNmax 1 so that only reads that uniquely mapped are considered. All other options for STAR are set as default. Then the mapped transcripts per gene were quantified with hamming distance threshold 1. Note that the above procedures are only for quantifying Over-Seq barcodes. Gene expression of the Over-Seq perturbed single cells were quantified by standard 10x Genomics Cell Ranger pipeline.

6.2.7 Regulatory networks and transcriptional regulator activity inference

The 11 (4 healthy and 7 diabetic) donor-specific regulatory networks were reverse engineered by ARACNe [Basso, 2005]. ARACNe was run with 100 bootstrap iterations using 1,813 transcription factors (genes annotated in Gene Ontology molecular function database, as GO:0003700, 'transcription factor activity', or as GO:0003677, 'DNA binding', and GO:0030528, 'transcription regulator activity', or as GO:00034677 and GO: 0045449, 'regulation of transcription'), 969 transcriptional cofactors (a manually curated list, not overlapping with the transcription factor list, built upon genes annotated as GO:0003712, 'transcription cofactor activity', or GO:0030528 or GO:0045449) and 3,370 signaling pathway related genes (annotated

in GO Biological Process database as GO:0007165 'signal transduction' and in GO cellular component database as GO:0005622, 'intracellular', or GO:0005886, 'plasma membrane'). Parameters were set to 0 DPI (Data Processing Inequality) tolerance and MI (Mutual Information) p-value (using MI computed by permuting the original dataset as null model) threshold of 10^{-8}.

The transcriptional regulator activity was inferred from gene expression signature by integrating the 11 donor-specific regulatory networks using the metaVIPER algorithm [Ding, 2018]. Specifically, for primary single cells, gene expression signature was generated by internal normalization within all profiled cells so that for each gene, the relative position of a specific single cell compared to all profiled cells was represented: for raw expression matrix described in 6.2.6, first do column-wise rank transformation, followed by extracting row-wise median and mad values, then use (rank-median)/mad as final signature. For Over-Seq perturbed single cells, gene expression signature was generated by normalizing to negative controls: for raw expression matrix described in 6.2.6, first do column-wise rank transformation within the control cells, followed by extracting row-wise median and mad values, then do column-wise rank transformation within the perturbed cells, and use (rank-median)/mad as final signature.

6.2.8 Data and software availability

The expression data of primary single cells reported in this study has been deposited in GEO under the accession number GSE98887. The expression data, as well as Over-Seq barcoding information of perturbed single cells reported in this study is available upon request to the authors.

6.3 Results

6.3.1 Identification of pancreatic islet cells under different biological states

Single pancreatic islet cells from primary donors were harvested (see Methods), and the corresponding expression and metaVIPER-inferred transcriptional regulator activity profiles were quantified (see Methods), based on which cells were projected onto on 2D-space with t-SNE [Maaten & Hinton, 2008]. A shown in Figure 6.1, at expression level, I see high-level of donor-specificity which is no longer seen using metaVIPER-inferred transcriptional regulator activity, allowing us to have fine-tuned annotation of cells.

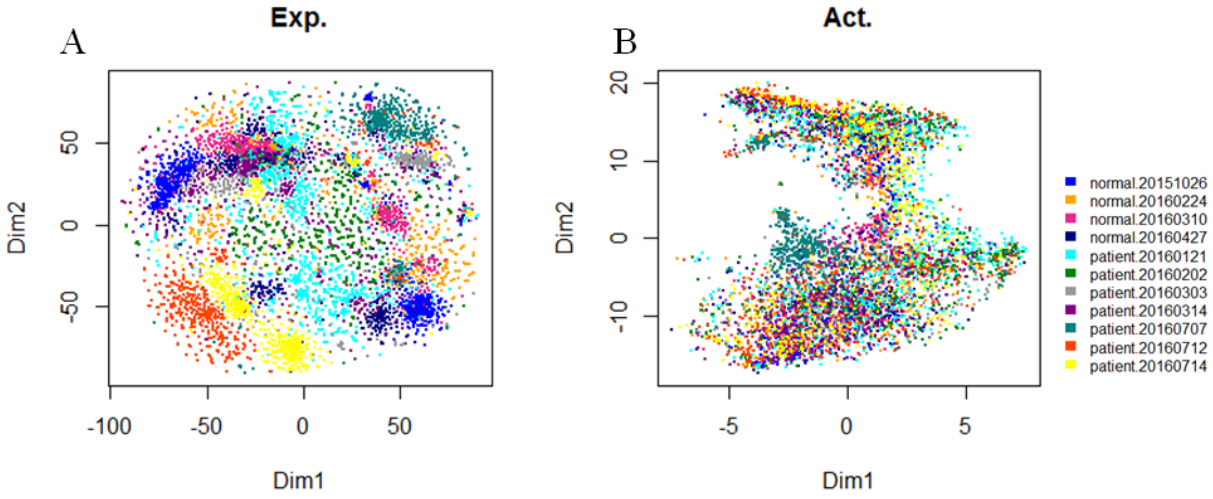

Figure 6.1 metaVIPER analysis reduces donor-specificity. Profiled single cells were projected onto on 2D-space with t-SNE [Maaten & Hinton, 2008] based on (A) expression quantified using log2(tpm+1) and (B) metaVIPER-inferred activity. Compared to expression, donor-specificity was no longer seen by using metaVIPER analysis.

It has been shown that hyperglycaemia induces metabolic stress, including oxidative, endoplasmic reticulum (ER) and hypoxic stress, leading to the dedifferentiation of β-cells, causing type-II diabetes [Kitamura, 2013]. Therefore I first checked the behavior of metabolic stress-related transcriptional regulators, including hypoxia stress-related *HIF1A* [Semenza, 1998], oxidative stress-related *NFKB1*, *HSF1*, *TP53*, *PI3K*, *AKT* and *JNK* [Finkel & Holbrook, 2000],

ER stress-related and further autophagy/apoptosis causing *ATF4*, *ATF6*, *XBP1* and *JNK* [Xu, 2005; Maiuri, 2007]. Results showed a clear cell population under metabolic stress state, marked by the activation of above mentioned transcriptional regulators (Figure 6.2).

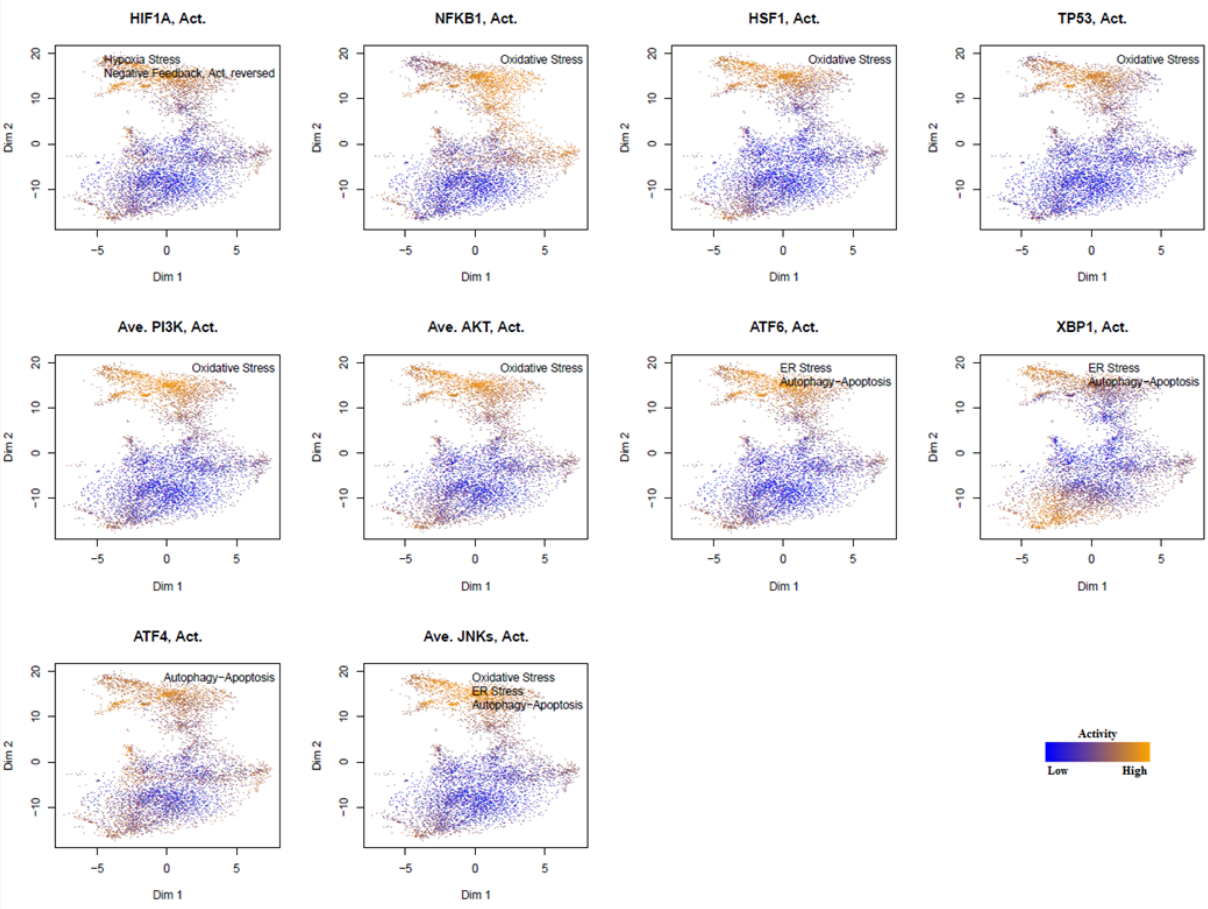

Figure 6.2 metaVIPER-inferred activity of metabolic stress-related transcriptional regulators. These regulators include hypoxia stress-related *HIF1A* [Semenza, 1998], oxidative stress-related *NFKB1*, *HSF1*, *TP53*, *PI3K*, *AKT* and *JNK* [Finkel & Holbrook, 2000], ER stress-related and further autophagy/apoptosis causing *ATF4*, *ATF6*, *XBP1* and *JNK* [Xu, 2005; Maiuri, 2007]. Specifically, due to the negative feedback regulation of *HIF1A* [Ke, 2006], the corresponding activity is reversed.

Besides metabolic status, I also checked the behavior of cell identity-related transcriptional regulators and marker genes, including activity α/β-cell-specific transcriptional regulator *IRX2* and *MAFA*, as well as expression of α/β-cell-specific markers insulin (*INS*) and glucagon (*GCG*) [Dorrell, 2011]. Results showed cell identity (majorly α/β-cell lineage) as independent dimension

besides cell metabolic status. I also discovered a small population of pancreatic exocrine cells, marked by elastase gene *CELA2A* [Kawashima, 1987] (Figure 6.3).

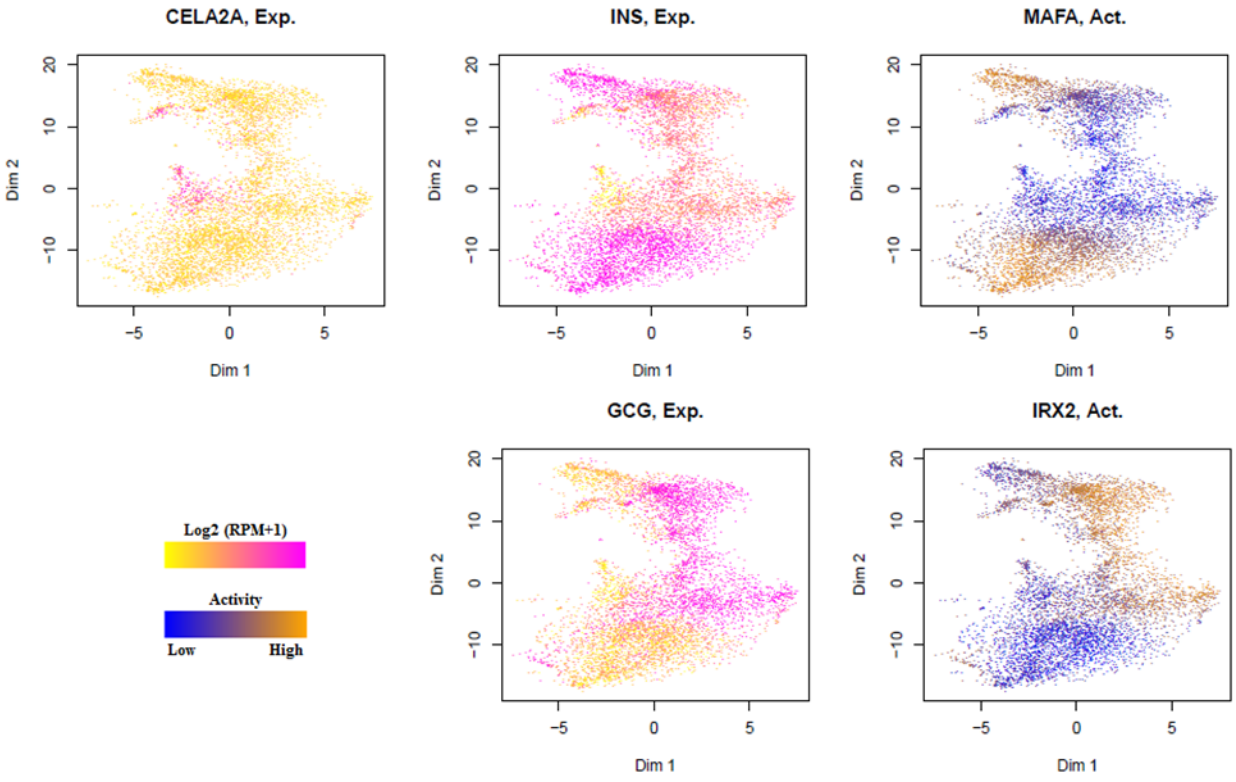

Figure 6.3 metaVIPER-inferred activity and expression of pancreatic cell lineage markers. These markers include α/β-cell-specific transcriptional regulator *IRX2* and MAFA, expression markers insulin (*INS*) and glucagon (*GCG*) [Dorrell, 2011], as well as pancreatic exocrine cell-specific expression markers *CELA2A* [Kawashima, 1987]

I then checked the behavior of de-differentiation-related transcriptional regulators, including positive regulator *FOXO1* and negative regulator *HES1* (which represses expression of de-differentiation marker *NGN3*), as well as stemness-related transcriptional regulators *NANOG*, *MYCL* and *POU5F1* [Talchai, 2012; Kitamura, 2013]. Results showed de-differentiation-related transcriptional regulators function among cells under metabolic stress. Specifically, stemness-

related transcriptional regulators were selectively activated in between strong α/β-cell states (Figure 6.4). Taking together, these results suggested metabolic stress triggers the function of de-differentiation-related transcriptional regulators, pushing cells to a more stem-like (de-differentiated) state, which has no strong α/β-cell identity.

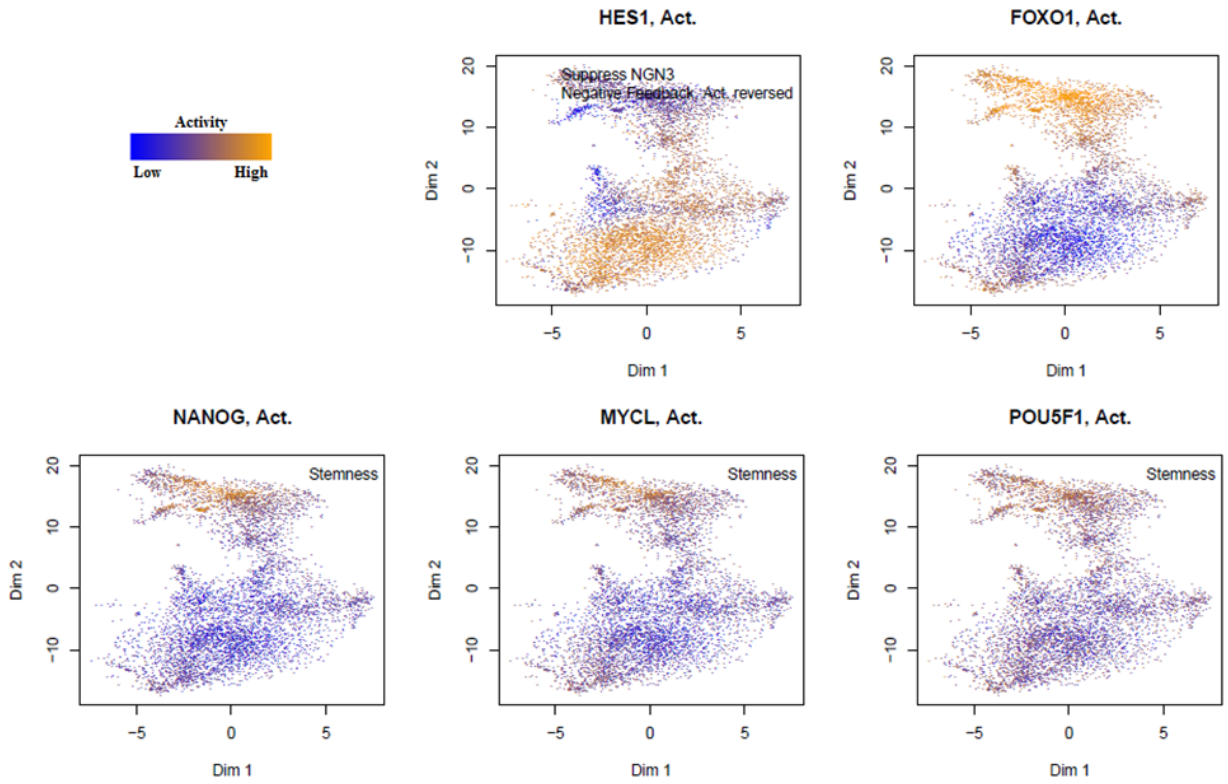

Figure 6.3 metaVIPER-inferred activity of de-differentiation/stemness-related transcriptional regulators. These regulators include positive/negative de-differentiation regulator *FOXO1*/*HES1*, and stemness-related transcriptional regulators *NANOG*, *MYCL* and *POU5F1* [Talchai, 2012; Kitamura, 2013]. Specifically, due to the negative feedback regulation of *HES1* [Hirata, 2002], the corresponding activity is reversed.

6.3.2 Construction of α/β-cell-transition pseudo-lineage, and identification of putative de-differentiation master regulators

I further used a principle curve [Hastie & Stuetzle, 1989] to describe the α/β-cell-transition pseudo-lineage among the metabolically stressed cells (Figure 6.4A). Alongside the lineage, cells

gradually losing α/β-cell identify, as well as gaining stemness feature (Figure 6.4B, shaded area). Noticeably, the shaded stage was majorly composed of cells from diabetic donors, indicating it to be the de-differentiated (stem-like) stage. I further identified putative de-differentiation master regulators that were activated in this stage (Figure 6.5).

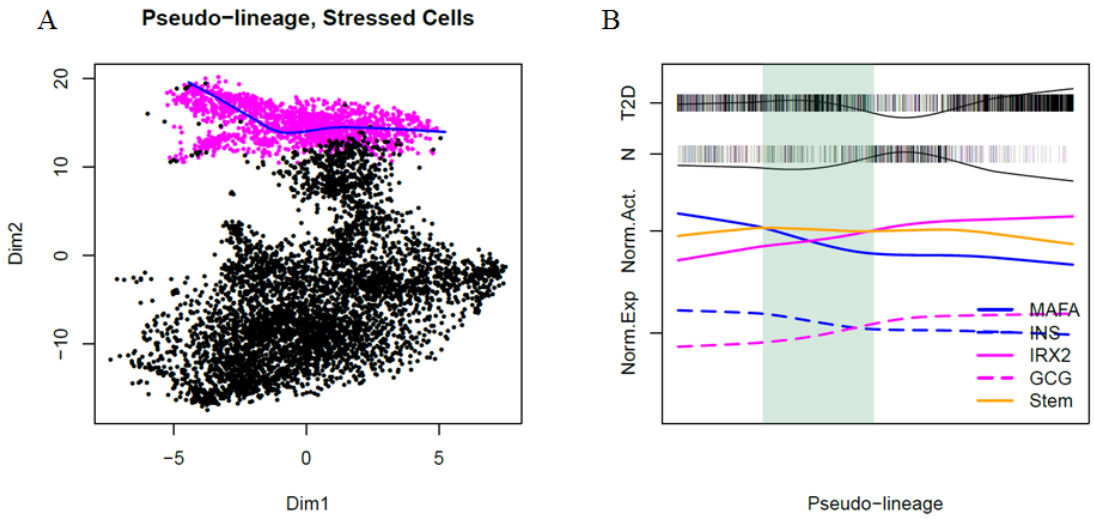

Figure 6.4 α/β-cell-transition pseudo-lineage among metabolically stressed cells. (A) Principle curve [Hastie & Stuetzle, 1989] representation of the pseudo-lineage. (B) Activity/expression of lineage markers, stemness-score, as well as normal/diabetic cell distribution alongside the pseudo-lineage.

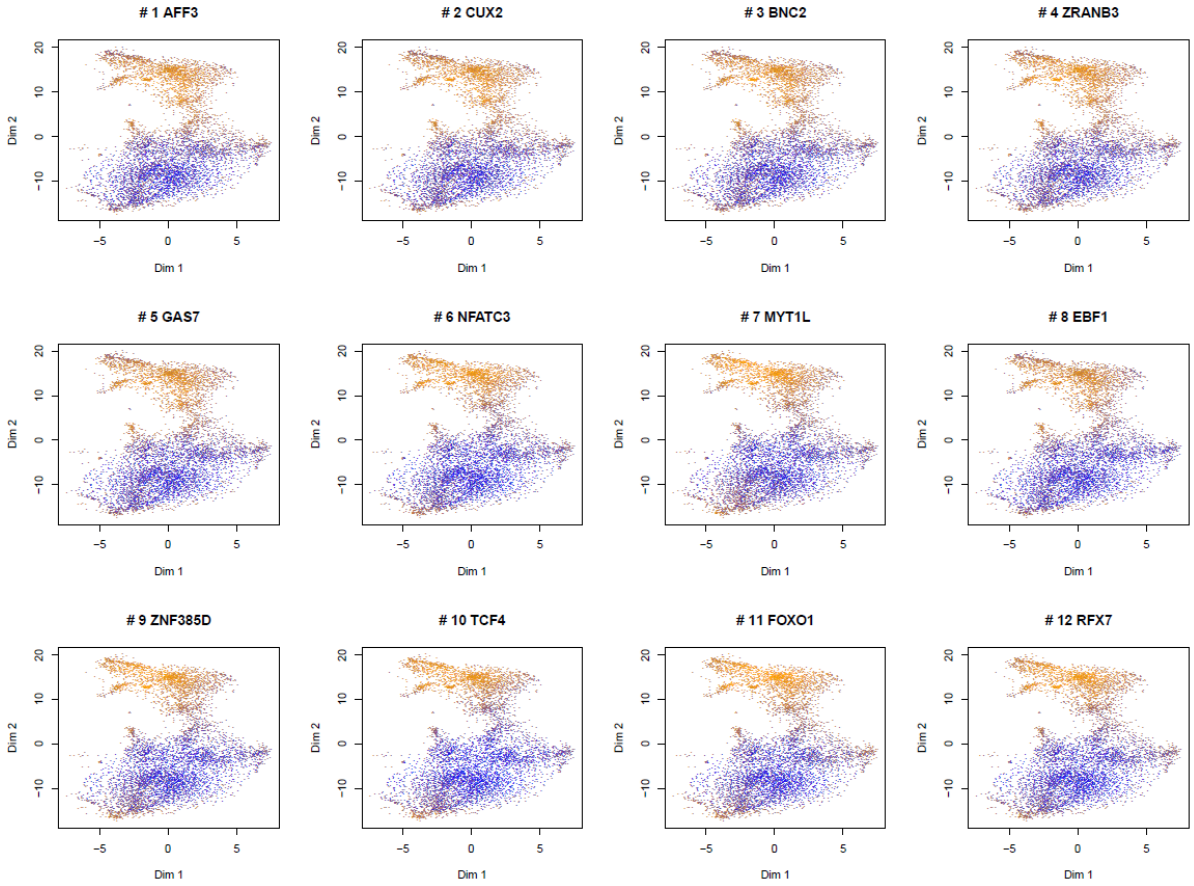

Figure 6.5 Putative de-differentiation master regulators.

6.3.3 Functional validation of putative de-differentiation master regulators

Dr. Son and I used Over-Seq to test whether putative de-differentiation master regulators can reprogram health β-cells (see Methods for details of β-cell enrichment experiment) to de-differentiated cells, or even α-cells (see Methods). I first quantified Over-Seq barcode, to assign perturbation identities of individual cells (see Methods, Figure 6.6). I then quantified gene expression profiles of the perturbed cells (specifically, cells with only 1 detected barcode), and further performed transcriptional regulator activity inference (see Methods).

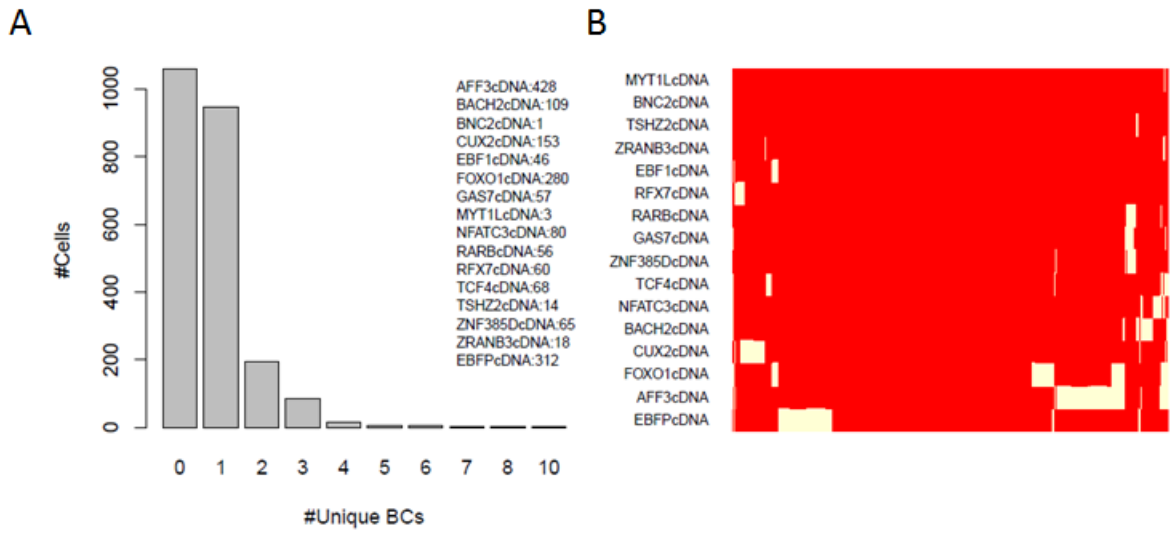

Figure 6.6 Summary of Over-Seq experiment. (A) Distribution of number of barcodes detected per cell. (B) Over-Seq identify of individual cells.

I projected the perturbed cells onto 2D-space with t-SNE [Maaten & Hinton, 2008] according to their expression profiles, and then visualized Over-Seq identity, as well as expression and metaVIPER-inferred activity of β/α-cell lineage markers. Specifically I showed health β-cells with *AFF3* overexpression (Figure 6.7). Results showed cells with *AFF3* overexpression formed a tight cluster, with losing β-cell identity (decreased *INS* expression and *MAFA* activity) and gaining α-cell identity (increased *GCG* expression and *IRX2* activity), indicating successful α-cell to β-cell reprogramming by overexpress *AFF3*.

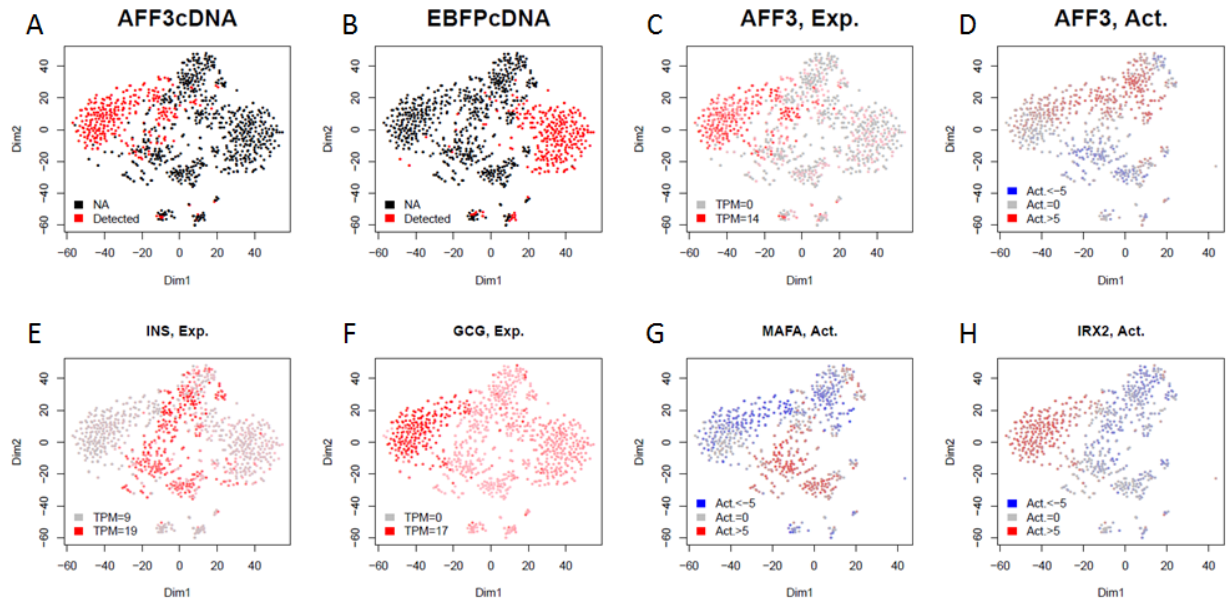

Figure 6.7 AFF3-induced α-cell identity among health β-cells. Perturbed cells were projected onto 2D-space with t-SNE [Maaten & Hinton, 2008] according to their expression profiles. (A) Cells with AFF3 Over-Seq barcode. (B) Negative control cells. (C, D) Expression and metaVIPER-inferred activity of AFF3. (E-H) Expression and metaVIPER-inferred activity of β/α-cell lineage markers.

Part VII

Developing Novel Therapeutics Targeting Cell-State Regulators of Breast Cancer Stem Cells

7.1 Introduction

Breast cancer stem cells (BCSCs) are originally considered as tumorgenic breast cancer cells: as few as 100 BCSCs are able to form tumors in mice, whereas tens of thousands of non-BCSCs failed to form tumors. Also, BCSCs could be serially passaged, by generating new tumors containing both BCSCs and non-BCSCs [Al-Hajj, 2003]. Later studies associated BCSCs with drug resistance, tumor relapse [Dean, 2005], as well as tumor metastasis [Aktas, 2009]. Therefore elucidating the molecular regulatory mechinary of BCSCs, and further developing corresponding novel targeted therapeutics would greatly help the treatment of breast cancer.

Previous studies [Alvarez, 2016; Aubry, 2015; Bisikirska, 2016; Brichta, 2015; Carro, 2010; Chen, 2014; Chudnowski, 2014; Ding, 2018a; Della Gatta, 2012; Ikiz, 2015; Kushwaha, 2015; Lefebvre, 2010; Repunte-Canonigo, 2015; Rodriguez-Barrueco, 2015; Talos, 2017] have shown that molecular regulatory mechinary of biological states could be represented by activity of transcriptional regulators, which could be predicted by metaVIPER or its precursor MRA (Master Regulator Analysis) from transcriptomic profiles. Further, abbarant biological states, e.g. tumor-related states, could be treated using drugs identified by OncoTreat analysis [Mitrofanova, 2015], which identifies drugs that can reverse abbarant transcriptional regulator activity signature.

However, BCSCs only constitute less then 1% of total cells among breast tumors, which masks them really challenging to study using conventional bulk tumor level methodologies. Therefore, by combining recently developed single cell RNA-Seq (scRNA-Seq) technology [Kolodziejczyk, 2015] with the above-mentioned metaVIPER algorithm, I identified putative BCSC population, as well as corresponding master regulators. I further applied OncoTreat analysis to each of the identified putative BCSCs, and identified potential targeted therapeutics.

7.2 Methods

7.2.1 Sample collection

In total Dr. Worley collected 5 primary samples from Columbia University Medical Center, see Table 7.1 for detail. Dr. Worley also profiled 1 mouse PDX model from Memorial Sloan Kettering Cancer Center, following IACUC guidelines.

Table 7.1 Primary sample information

ID	Diagnosis	Size, Grade, Stage	ER	PR	HER2	Ki67
FF	Carcinoma, Invasive, Ductal – NOS Poorly Differentiated (Grade 3), Total score 9 (Tubule Score 3, Nuclear Grade Score 3, Mitotic Score 3)	pT1c: Tumor more than 10 mm but not more than 20 mm in greatest dimension. pN0	0	15% weak	IHC 0, SISH negative	70%
HD	Carcinoma, Invasive with mixed ductal and lobular features Moderately Differentiated (Grade 2), Total score 7 (Tubule Score 3, Nuclear Grade Score 2, Mitotic Score 2) Ductal carcinoma in situ, nuclear grade 2, widespread 30-40 % Intraductal cribriform subtype Necrosis is present within the intraductal carcinoma Lobular neoplasia is not present	pT2(m-5) pN2a	Pos. 95% Strong	Pos. 40% intermediate	Neg. IHC 0	20%
IF	Invasive ductal carcinoma, well-differentiated with focal tubular features, modified Scarff- Bloom-Richardson grade 1 (2 + 2 + 1),	pT1c: Tumor more than 10 mm but not more than 20 mm in greatest dimension. pNX	Pos.90% Strog	Pos.5% weak	Neg. IHC 1+, SISH negative	10%
IS	microinvasive carcinoma.	pT1mi pN0	Pos. 100% Strong	Pos. 100% strong	Neg.IHC 0	2%
JC	Invasive ductal carcinoma, poorly differentiated, modified Scarff-Bloom-Richardson score 8-9 (3, 2-3, 3).	pT2: Tumor more than 20 mm but not more than 50 mm in greatest dimension. Two foci (M=2) of invasion are present, the larger is pT2. pN0(sn)	Neg. 0	Neg.0	Neg.IHC0	35%

7.2.2 Tumor dissociation

After being picked up from pathology, tumor tissue was manually separated from necrotic tissue. Next, the samples were dissociated using a Miltenyi gentleMACS Octo dissociator, according to

the manufacturer's instructions. To maintain transcriptome integrity, the samples were processed quickly and kept on ice as much as possible. Generally, the time from retrieval to lysis was approximately 4 hours, with less than 90 minutes above 0°C.

7.2.3 Enrichment of putative BCSCs

After dissociation, the single-cell suspension was filtered, treated with ACK buffer to lyse red blood cells, and then stained and sorted for CD49f(high), EpCAM(positive), DAPI(negative), and lineage(negative) using a BD Influx flow sorter. This FACS protocol removed approximately 95-99% of cells, with the remainder being enriched for BCSCs.

7.2.4 Single cell RNA sequencing

Single cells obtained from primary patients were processed following standard scPLATE-Seq protocol (see Part I for details). Single cells obtained from mouse PDXs were processed following standard 10x Genomics Chromium protocol.

7.2.5 scRNA-Seq gene expression analysis

Single cell profiled using scPLATE-Seq were analyzed following same pipeline described in section 1.2.2. Single cells profiled using 10x Genomics Chromium were analyzed using standard 10x Genomics Cell Ranger pipeline.

7.2.6 Regulatory networks and protein activity profiles

Protein and microRNA activity profiles were inferred from expression matrix by integrating the core TCGA regulatory networks [Giorgi, 2017] and CUPID microRNA regulatory networks [Sumazin , 2011] with metaVIPER [Alvarez, 2016; Ding, 2018], respectively.

7.2.7 OncoTreat analysis

As described in [Mitrofanova, 2015], OncoTreat analysis identifies drugs that can reverse the tumor-specific transcriptional regulator activity signature. So that, in this specific study, I either kill putative BCSCs or push them to more differentiated thereby treatable states. To archive this, I first identified the best matching cell line for each profiled single cell using gene set enrichment analysis (GSEA) [Subramanian, 2005]. I used each single cell's top 25 most activated/suppressed transcriptional regulators as gene set, and the transcriptional regulator activity profile for each cell line (see section 7.2.8, available as part of OncoTreat pipeline) as GSEA signature. Cell line that gives highest normalized enrichment score (nES) will be considered as best-matching one for each profiled single cells. I further identified best matching cell line for putative BCSCs, by checking the distribution of cell lines' nESs alongside the constructed BCSC-differentiated breast cancer cell pseudo-lineage. I then identified drugs that can reverse the transcriptional regulator activity signature of the BCSCs using GSEA. I used each single cell's top 25 most activated/suppressed transcriptional regulators as gene set, and the transcriptional regulator activity profile for each drug perturbation of the best-matching cell line (see section 7.2.8, available as part of OncoTreat pipeline) as GSEA signature. Drug perturbation that gives lowest normalized nES will be considered as best candidate for treating putative BCSCs (see Figure 7.1 for schematic workflow of OncoTreat).

7.2.8 Data and software availability

The expression data of primary single cells reported in this study has been deposited in GEO under the accession number GSE108540. The expression data of mouse PDXs single cells reported in this study is available upon request to the authors. The transcriptional regulator

activity profiles for native cell lines, and cell lines with drug perturbations used in OncoTreat analysis in this study is available upon request to the authors.

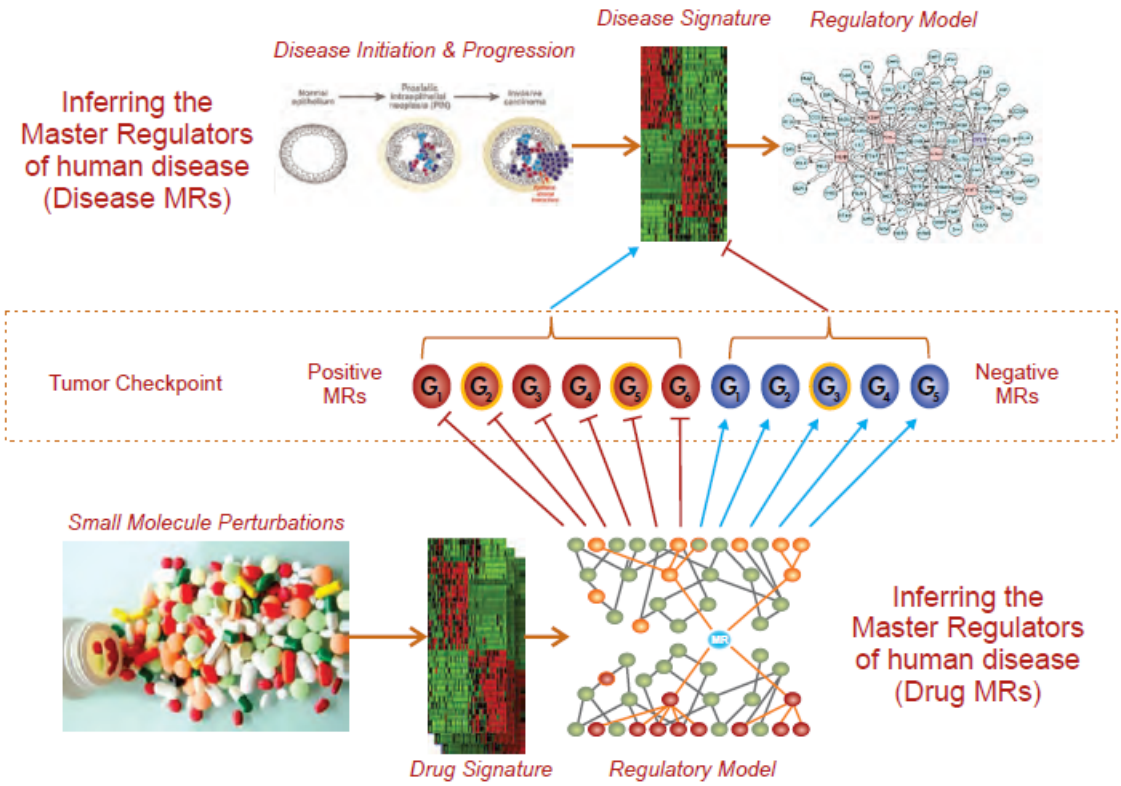

Figure 7.1 Schematic workflow of OncoTreat.

7.3 Results

7.3.1 Construction of BCSC-differentiated breast cancer cell pseudo-lineage

Single breast cancer stem cells from primary patients were harvested (see Methods), and the corresponding expression and metaVIPER-inferred transcriptional regulator activity profiles were quantified (see Methods), based on which cells were projected onto on 2D-space with t-SNE [Maaten & Hinton, 2008]. A shown in Figure 7.2, at expression level, I see high-level of

patient-specificity (cells clustered according to their corresponding patient IDs, e.g. the upper-left cluster is composed of jc cells), as well as slight batch effect (among patient-specific cell clusters, e.g. upper-left jc cluster, cells profiled within the same scPLATE-Seq plate, e.g. jc1, tend to co-segregate). These were no longer seen using metaVIPER-inferred transcriptional regulator activity, allowing us to have fine-tuned annotation of cells. With metaVIPER-analysis, cells from different patients/batches mixed together and formed a "microphone" shape. As shown in Figure 7.3, metaVIPER-inferred activity of well-established BCSC-specific transcriptional regulators, including *WNT1*, *NITCH1*, *SHH* and *BMI1* [Al-Hajj & Clarke, 2004] indicated putative BCSCs were enriched in the head of the "microphone", while differentiated breast cancer cells were enriched in the handle of the "microphone". Specifically, I checked the activity of microRNAs 200c (miR-200c, see Methods, Figure 7.4). Since not poly-adenylated, the abundance of microRNAs is not detectable using regular scRNA-Seq. However given proper regulatory networks, the activity of microRNAs can be correctly inferred. In this case, the activity of miR-200c goes opposite with BCSC-specific transcriptional regulators *BMI1*, in consistence with the fact that miR-200c was considered as suppressor of *BMI1* [Shimono, 2009]. Taking together, metaVIPER analysis revealed a BCSC-differentiated breast cancer cell pseudo-lineage, which was further described by a principle curve [Hastie & Stuetzle, 1989] on the t-SNE plot (Figure 7.2B).

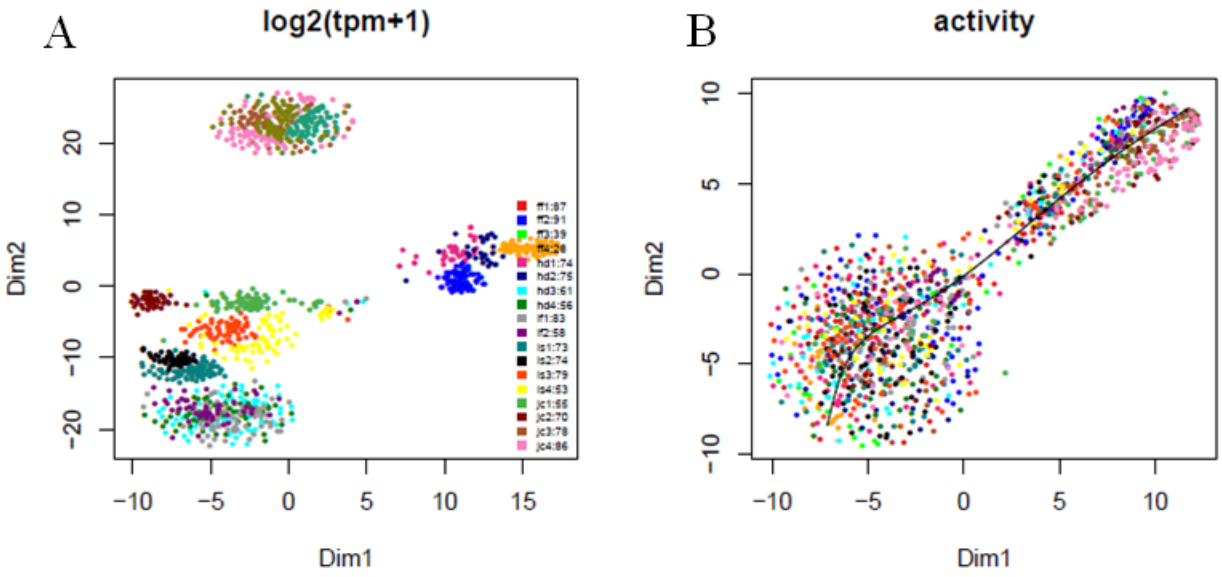

Figure 7.2 metaVIPER analysis reduces patient-specificity and batch effect, which facilitated the construction of BCSC-differentiated breast cancer cell pseudo-lineage. Profiled single cells were projected onto on 2D-space with t-SNE [Maaten & Hinton, 2008] based on (A) expression quantified using log2(tpm+1) and (B) metaVIPER-inferred activity. Compared to expression, patient-specificity and batch effect were no longer seen by using metaVIPER analysis, which further revealed a BCSC-differentiated breast cancer cell pseudo-lineage described by a principle curve [Hastie & Stuetzle, 1989].

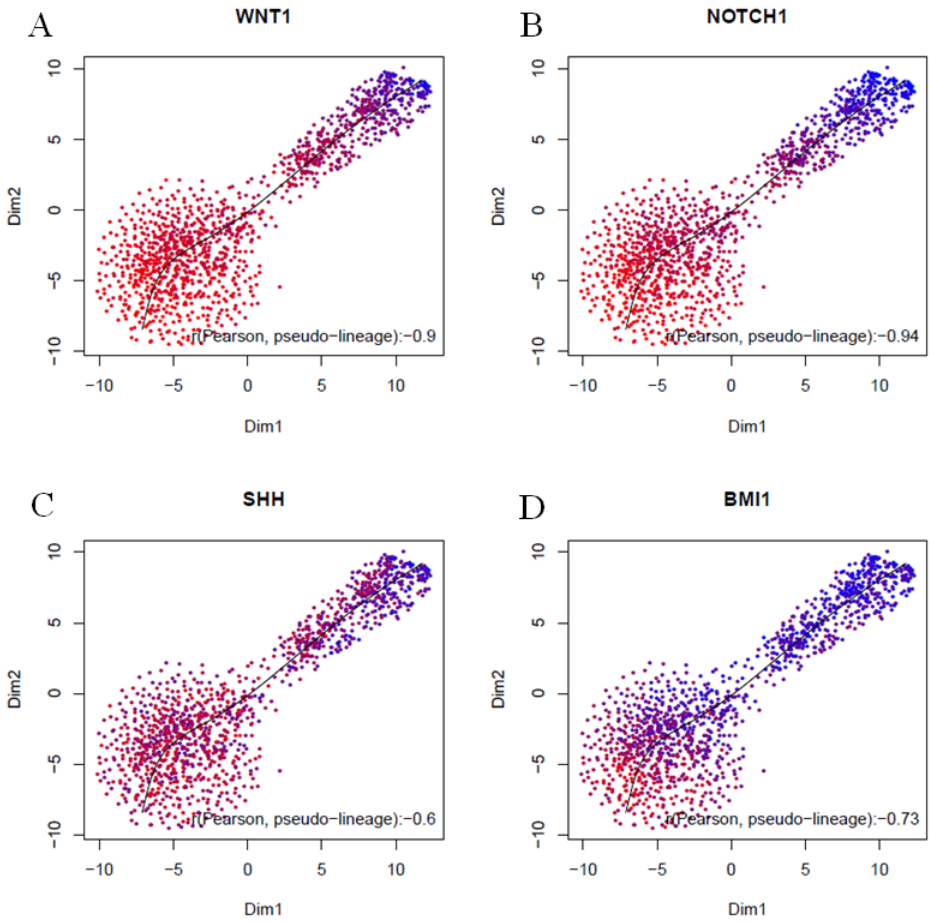

Figure 7.3 metaVIPER-inferred activity of well-established BCSC-specific transcriptional regulators (non-microRNAs) among primary cells. Blue-to-red color gradient indicated low-to-high metaVIPER-inferred activity.

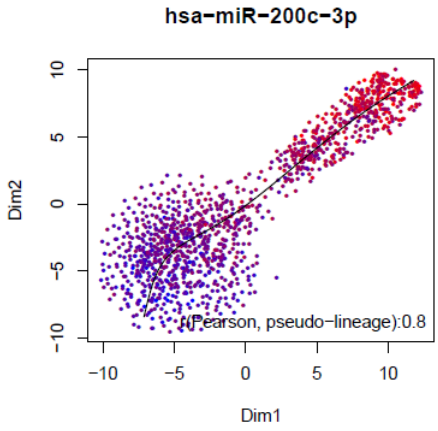

Figure 7.4 metaVIPER-inferred activity of miR-200c among primary cells. Blue-to-red color gradient indicated low-to-high metaVIPER-inferred activity.

7.3.2 Identification of novel BCSC-associated transcriptional regulators

I further identified BCSC-associated transcriptional regulators (Figure 7.5), including microRNAs (Figure 7.6), according to the enrichment of regulator-specific highly-activated cells alongside the pseudo-lineage. Specifically, I recapitulated miR-200c among the identified top BCSC-associated microRNAs.

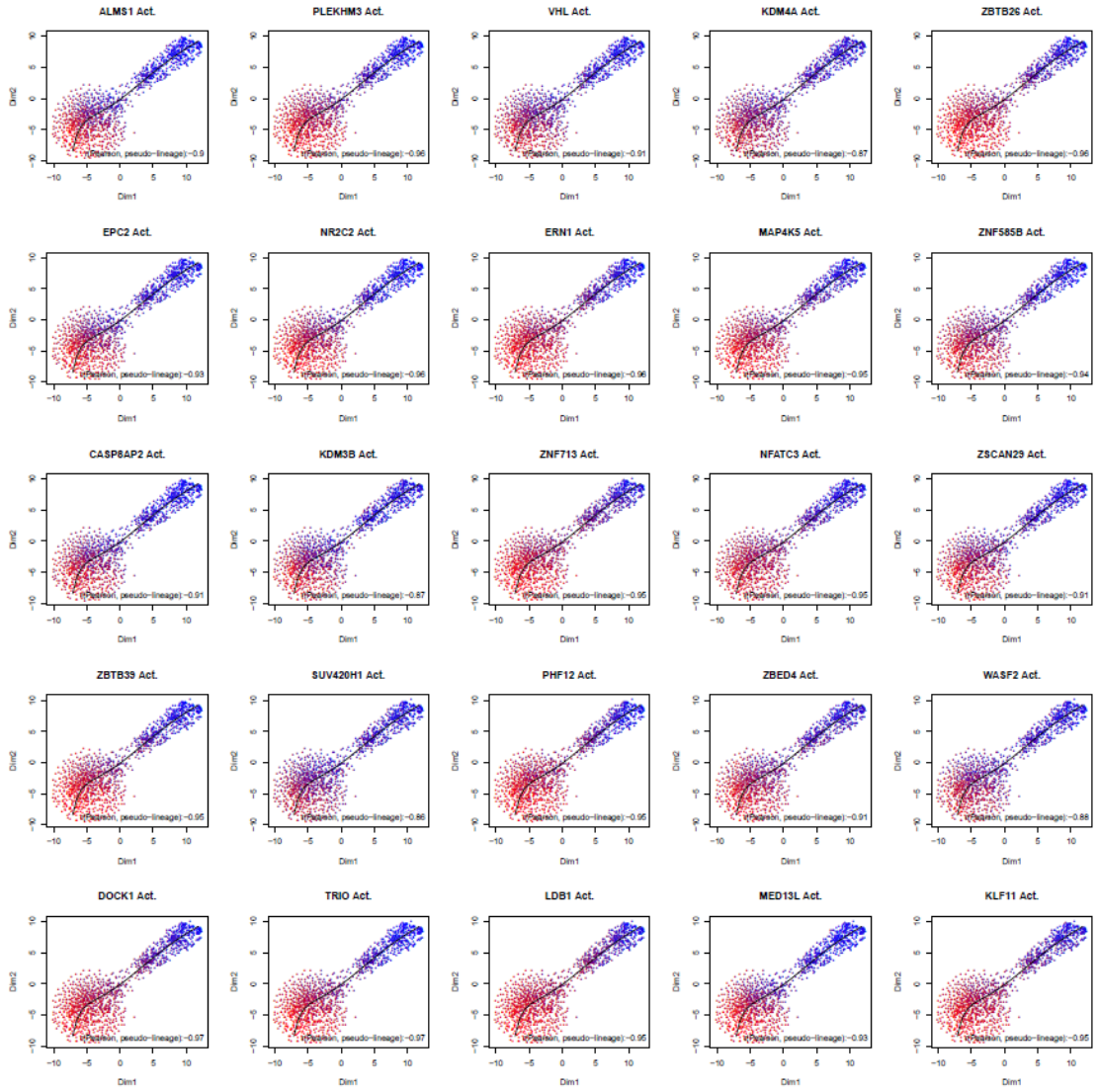

Figure 7.5 Novel BCSC-associated transcriptional regulators (non-microRNAs). Blue-to-red color gradient indicated low-to-high metaVIPER-inferred activity.

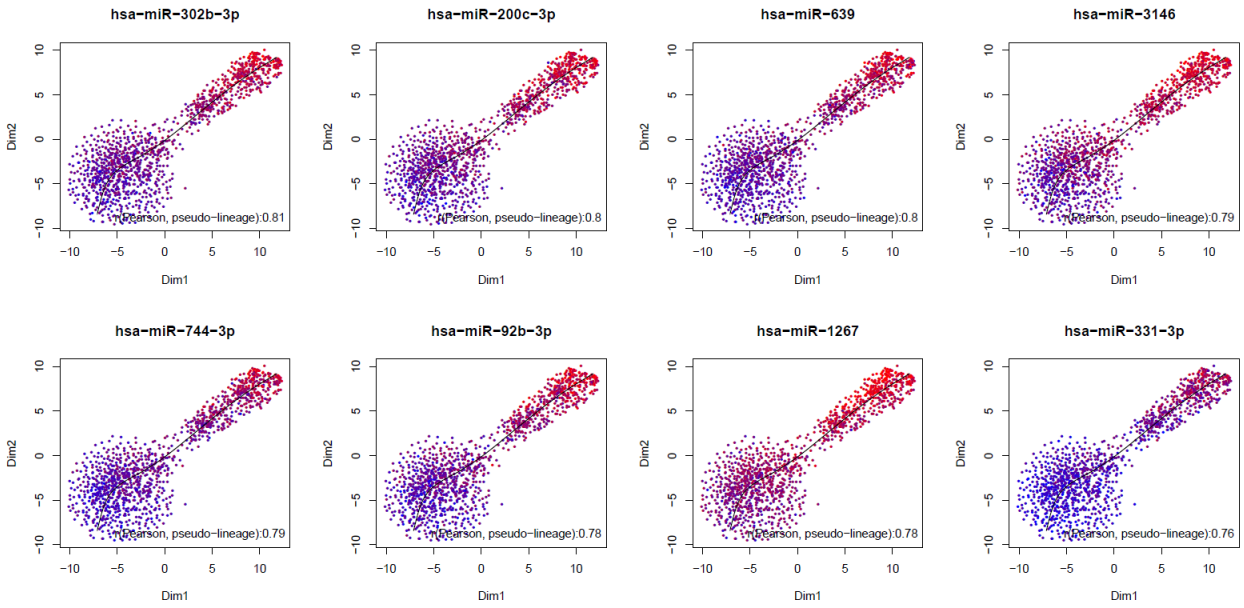

Figure 7.6 Novel BCSC-associated microRNAs. Blue-to-red color gradient indicated low-to-high metaVIPER-inferred activity.

7.3.3 BT20 as representative cell line for putative BCSCs

To identify drugs using OncoTreat that can reverse the putative BCSC-specific transcriptional regulator activity signature so that I either kill or push them to more differentiated thereby treatable states, I first need to identify cell lines that best represent putative BCSCs, by measuring the similarity between single cells and cell lines using GSEA normalized enrichment score (nES). For each profiled single cell, I used top 25 most activated/suppressed transcriptional regulators as gene set, and the transcriptional regulator activity profile for each cell line as signature for gene set enrichment analysis (GSEA) [Subramanian, 2005], see Methods for detail. Result showed breast cancer cell line BT20 [Lasfargues & Ozzello, 1958] to be the best match with previously identified putative BCSCs (Figure 7.7).

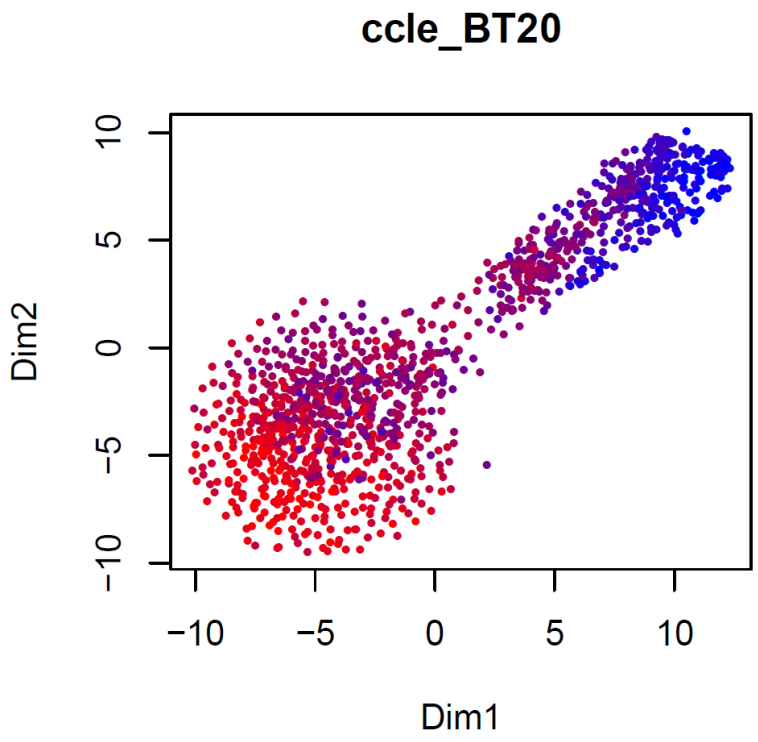

Figure 7.7 BT20 as representative cell line for putative BCSCs obtained from primary samples. Blue-to-red color gradient indicated low-to-high GSEA nES.

7.3.4 Crizotinib, Etopiside and Gemcitabine as best candidate for treating putative BCSCs

I then performed OncoTreat analysis (see Methods) on putative BCSCs against drug perturbation (different drugs, IC_{20} and IC_{50} for each drug, 6h and 24h treatment time) profiles of BT20 cell line, which was used as substitution of drug perturbation profiles of putative BCSCs, because dealing with primary cells is usually hard. In theory, if the reversion of putative BCSC-specific transcriptional regulator activity signature can be made by certain perturbation, the transcriptional regulator activity profile for that specific perturbation should anti-correlate with the transcriptional regulator activity profile of BT20 cell line. Based on this, the signature reversion was quantified by dis-similarity between profiled single cells and BT20 perturbations using GSEA nES. For each profiled single cell, I used the top 25 most activated/suppressed transcriptional regulators as gene set, and the transcriptional regulator activity profile for each

BT20 perturbation as GSEA signature. Result showed regardless of treating concentration and time, Crizotinib, Etopiside and Gemcitabine consistently reversed BCSC-specific transcriptional regulator activity signature, therefore considered as best candidate for treating putative BCSCs (Figure 7.8).

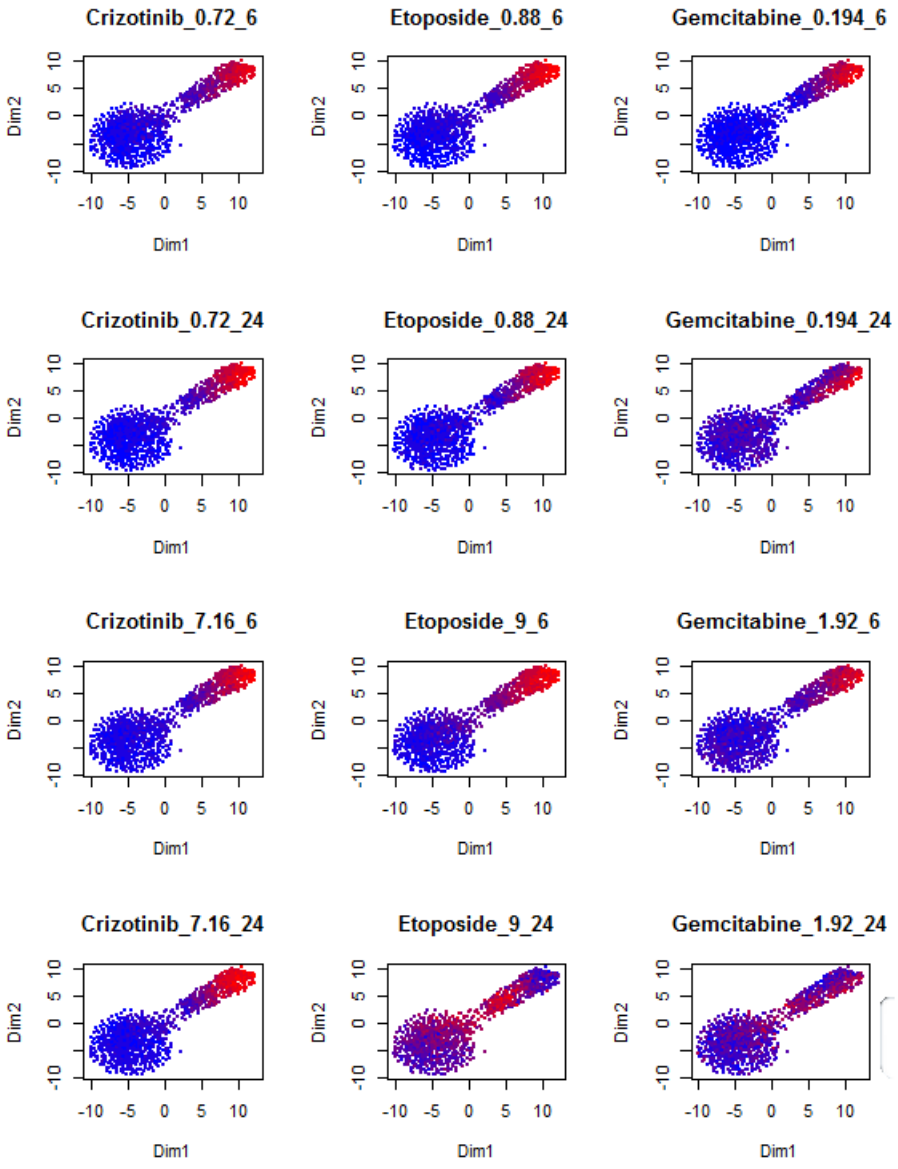

Figure 7.8 Crizotinib, Etopiside and Gemcitabine as best candidate for treating putative BCSCs obtained from primary samples. Blue-to-red color gradient indicated low-to-high GSEA nES.

7.3.5 Identification and treatment of putative BCSCs among mouse PDX models

Similar with primary patient samples, single breast cancer stem cells from mouse PDX models were harvested and analyzed. As shown in Figure 7.9, metaVIPER-inferred activity of *WNT1*, *NITCH1*, *SHH* and miR-200c indicated putative BCSCs were recapitulated, further suggested mouse PDXs as proper models for studying the regulatory mechanism and therapeutic strategies of putative BCSCs. Similarly, BT20 was shown to be representative cell line for putative BCSCs (Figure 7.10), and Crizotinib, Etopiside, Gemcitabine were shown to be best candidate for treating putative BCSCs (Figure 7.11) among mouse PDX models.

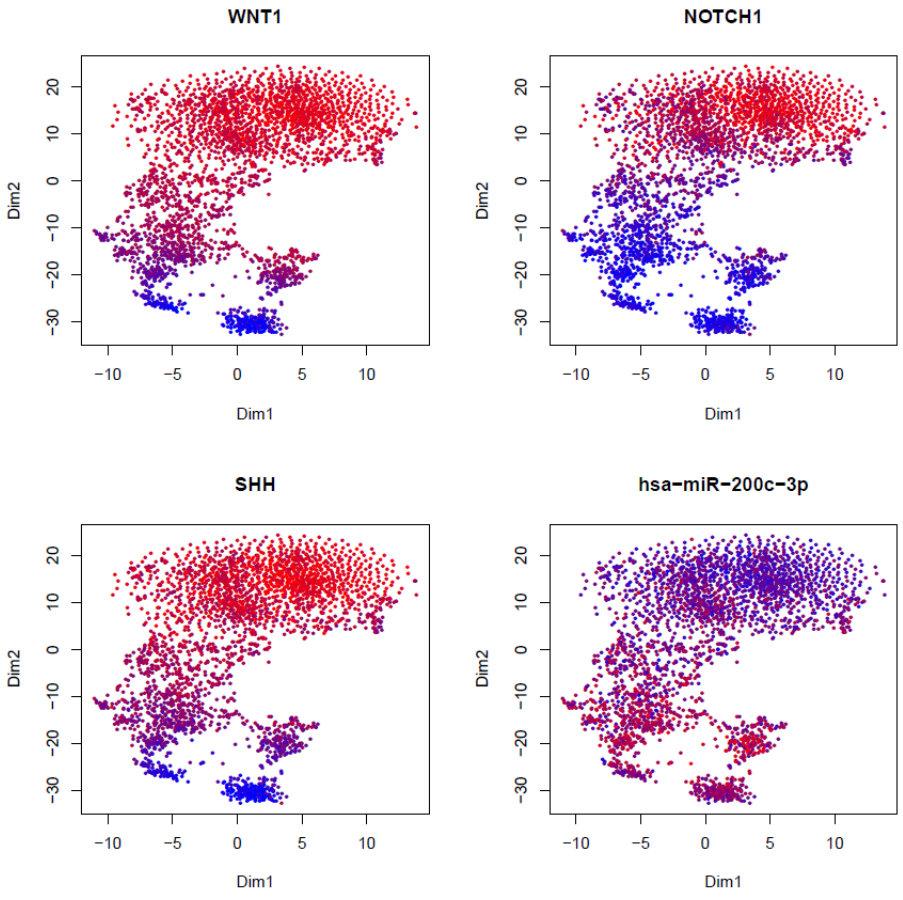

Figure 7.9 metaVIPER-inferred activity of well-established BCSC-specific transcriptional regulators among single cells obtained from mouse PDX models. Blue-to-red color gradient indicated low-to-high metaVIPER-inferred activity.

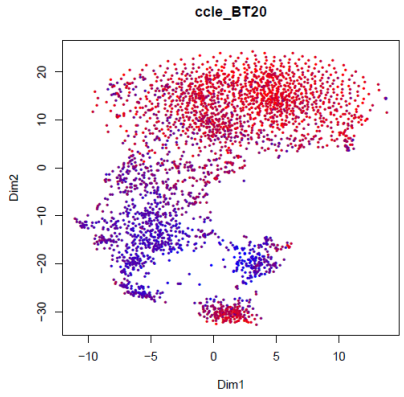

Figure 7.10 BT20 as representative cell line for putative BCSCs obtained from mouse PDX models. Blue-to-red color gradient indicated low-to-high GSEA nES.

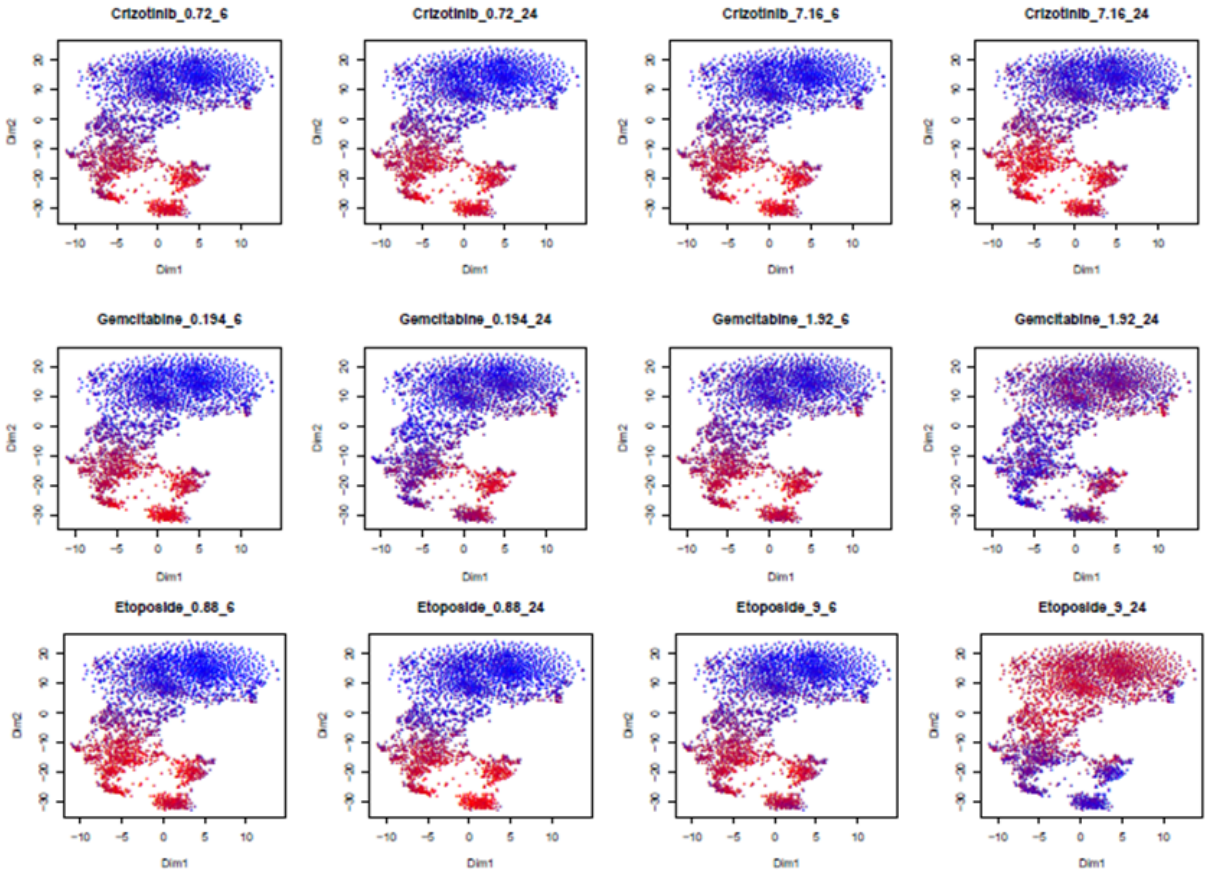

Figure 7.11 Crizotinib, Etopiside and Gemcitabine as best candidate for treating putative BCSCs obtained from mouse PDX models. Blue-to-red color gradient indicated low-to-high GSEA nES.

Part VIII

Conclusion

In this thesis I describe the development of two analytical tools for better understanding biological states of single cells, including metaVIPER and iterClust; the development of single cell RNA-Seq platform, scPLATE-Seq. Combining these, I answered three specific biological questions, including 1) understanding the inter- and intra-tumor molecular subclasses of human glioblastoma, 2) elucidating regulators of β-cell de-differentiation in type-2 diabetes and 3) developing novel therapeutics targeting cell-state regulators of breast cancer stem cells. Currently, metaVIPER and iterClust have been published. scPLATE-Seq and the glioblastoma studies are under submission. Discoveries in the type-2 diabetes and breast cancer stem cell studies are under experimental validation.

Part IX

References

[Aktas, 2009] Aktas, Bahriye, et al. "Stem cell and epithelial-mesenchymal transition markers are frequently overexpressed in circulating tumor cells of metastatic breast cancer patients." *Breast cancer research* 11.4 (2009): R46.

[Al-Hajj, 2003] Al-Hajj, Muhammad, et al. "Prospective identification of tumorigenic breast cancer cells." *PNAS* 100.7 (2003): 3983-3988.

[Al-Hajj & Clarke, 2004] Al-Hajj, Muhammad, and Michael F. Clarke. "Self-renewal and solid tumor stem cells." *Oncogene* 23.43 (2004): 7274.

[Alvarez, 2016] Alvarez, Mariano J., et al. "Functional characterization of somatic mutations in cancer using network-based inference of protein activity." *Nature genetics* 48.8 (2016): 838.

[Alvarez, 2009] Alvarez, Mariano Javier, et al. "Correlating measurements across samples improves accuracy of large-scale expression profile experiments." *Genome biology* 10.12 (2009): R143.

[Anders & Huber, 2010] Anders, Simon, and Wolfgang Huber. "Differential expression analysis for sequence count data." *Genome biology* 11.10 (2010): R106.

[Ankerst, 1999] Ankerst, Mihael, et al. "OPTICS: ordering points to identify the clustering structure." *ACM Sigmod record*. Vol. 28. No. 2. ACM, 1999.

[Aubry, 2015] Aubry, Soline, et al. "Assembly and interrogation of Alzheimer's disease genetic networks reveal novel regulators of progression." *PloS one* 10.3 (2015): e0120352.

[Aytes, 2014] Aytes, Alvaro, et al. "Cross-species regulatory network analysis identifies a synergistic interaction between FOXM1 and CENPF that drives prostate cancer malignancy." *Cancer cell* 25.5 (2014): 638-651.

[Bain, 1994] Bain, Gretchen, et al. "E2A proteins are required for proper B cell development and initiation of immunoglobulin gene rearrangements." *Cell* 79.5 (1994): 885-892.

[Basso, 2005] Basso, Katia, et al. "Reverse engineering of regulatory networks in human B cells." *Nature genetics* 37.4 (2005): 382.

[Bisikirska, 2016] Bisikirska, Brygida, et al. "Elucidation and pharmacological targeting of novel molecular drivers of follicular lymphoma progression." *Cancer research* 76.3 (2016): 664-674.

[Bosl & Motzer, 1997] Bosl, George J., and Robert J. Motzer. "Testicular germ-cell cancer." *NEJM* 337.4 (1997): 242-254.

[Bray, 2016] Bray, Nicolas L., et al. "Near-optimal probabilistic RNA-seq quantification." *Nature biotechnology* 34.5 (2016): 525.

[Brichta, 2015] Brichta, Lars, et al. "Identification of neurodegenerative factors using translatome–regulatory network analysis." *Nature neuroscience* 18.9 (2015): 1325.

[Burrell, 2013] Burrell, Rebecca A., et al. "The causes and consequences of genetic heterogeneity in cancer evolution." *Nature* 501.7467 (2013): 338.

[Bussemaker, 2001] Bussemaker, Harmen J., Hao Li, and Eric D. Siggia. "Regulatory element detection using correlation with expression." *Nature genetics* 27.2 (2001): 167.

[Butler, 2003] Butler, Alexandra E., et al. "β-cell deficit and increased β-cell apoptosis in humans with type 2 diabetes." *Diabetes* 52.1 (2003): 102-110.

[Butte, 2000] Butte, Atul J., and Isaac S. Kohane. "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements." *Biocomputing 2000*. 1999. 418-429.

[Califano & Alvarez, 2017] Califano, Andrea, and Mariano J. Alvarez. "The recurrent architecture of tumour initiation, progression and drug sensitivity." *Nature reviews cancer* 17.2 (2017): 116.

[Canoll, 1996] Canoll, Peter D., et al. "GGF/neuregulin is a neuronal signal that promotes the proliferation and survival and inhibits the differentiation of oligodendrocyte progenitors." *Neuron* 17.2 (1996): 229-243.

[Carlson, 2011] Carlson, Brett L., et al. "Establishment, maintenance, and in vitro and in vivo applications of primary human glioblastoma multiforme (GBM) xenograft models for translational biology studies and drug discovery." *Current protocols in pharmacology* (2011): 14-16.

[Carlson & Maintainer, 2015] Carlson, M. and Maintainer, B. P. TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s). R package version 3.2.2. DOI: 10.18129/B9.bioc.TxDb.Hsapiens.UCSC.hg19.knownGene (2015).

[Carro, 2010] Carro, Maria Stella, et al. "The transcriptional network for mesenchymal transformation of brain tumours." *Nature* 463.7279 (2010): 318.

[Ceccarelli, 2016] Ceccarelli, Michele, et al. "Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma." *Cell* 164.3 (2016): 550-563.

[Chen, 2014] Chen, James C., et al. "Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks." *Cell* 159.2 (2014): 402-414.

[Chudnovsky, 2014] Chudnovsky, Yakov, et al. "ZFHX4 interacts with the NuRD core member CHD4 and regulates the glioblastoma tumor-initiating cell state." *Cell reports* 6.2 (2014): 313-324.

[Clevers, 2006] Clevers, Hans. "Wnt/β-catenin signaling in development and disease." *Cell* 127.3 (2006): 469-480.

[Darmanis, 2017] Darmanis, Spyros, et al. "Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma." *Cell reports* 21.5 (2017): 1399-1410.

[Dean, 2005] Dean, Michael, Tito Fojo, and Susan Bates. "Tumour stem cells and drug resistance." *Nature reviews cancer* 5.4 (2005): 275.

[Della Gatta, 2012] Della Gatta, Giusy, et al. "Reverse engineering of TLX oncogenic transcriptional networks identifies RUNX1 as tumor suppressor in T-ALL." *Nature medicine* 18.3 (2012): 436.

[Ding, 2018a] Ding, Hongxu, et al. "Quantitative assessment of protein activity in orphan tissues and single cells using the metaVIPER algorithm." *Nature communirlsontions* 9.1 (2018): 1471.

[Ding, 2018b] Ding, Hongxu, Wanxin Wang, and Andrea Califano. "iterClust: a statistical framework for iterative clustering analysis." *Bioinformatics* 1 (2018): 2.

[Dixit, 2016] Dixit, Atray, et al. "Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens." *Cell* 167.7 (2016): 1853-1866.

[Dobin, 2013] Dobin, Alexander, et al. "STAR: ultrafast universal RNA-seq aligner." *Bioinformatics* 29.1 (2013): 15-21.

[Dorrell, 2011] Dorrell, C., et al. "Transcriptomes of the major human pancreatic cell types." *Diabetologia* 54.11 (2011): 2832.

[Du, 2007] Du, Pan, Warren A. Kibbe, and Simon M. Lin. "nuID: a universal naming scheme of oligonucleotides for illumina, affymetrix, and other microarrays." *Biology direct* 2.1 (2007): 16.

[Du, 2008] Du, Pan, Warren A. Kibbe, and Simon M. Lin. "lumi: a pipeline for processing Illumina microarray." *Bioinformatics* 24.13 (2008): 1547-1548.

[Du, 2010] Du, Pan, et al. "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis." *BMC bioinformatics* 11.1 (2010): 587.

[Dutta, 2016] Dutta, Aditya, et al. "Identification of an NKX3. 1-G9a-UTY transcriptional regulatory network that controls prostate differentiation." *Science* 352.6293 (2016): 1576-1580.

[Ester, 1996] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." *Kdd*. Vol. 96. No. 34. 1996.

[Finak, 2015] Finak, Greg, et al. "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data." *Genome biology* 16.1 (2015): 278.

[Finkel & Holbrook, 2000] Finkel, Toren, and Nikki J. Holbrook. "Oxidants, oxidative stress and the biology of ageing." *Nature* 408.6809 (2000): 239.

[Fränti, 2006] Fränti, Pasi, Olli Virmajoki, and Ville Hautamaki. "Fast agglomerative clustering using a k-nearest neighbor graph." *IEEE transactions on pattern analysis and machine intelligence* 28.11 (2006): 1875-1881.

[Freije, 2004] Freije, William A., et al. "Gene expression profiling of gliomas strongly predicts survival." *Cancer research* 64.18 (2004): 6503-6510.

[Friedman, 2004] Friedman, Nir. "Inferring cellular networks using probabilistic graphical models." *Science* 303.5659 (2004): 799-805.

[Gautier, 2004] Gautier, Laurent, et al. "affy—analysis of Affymetrix GeneChip data at the probe level." *Bioinformatics* 20.3 (2004): 307-315.

[Gensert, 2001] Gensert, JoAnn M., and James E. Goldman. "Heterogeneity of cycling glial progenitors in the adult mammalian cortex and white matter." *Developmental neurobiology* 48.2 (2001): 75-86.

[Gilbert, 2014] Gilbert, Luke A., et al. "Genome-scale CRISPR-mediated control of gene repression and activation." *Cell* 159.3 (2014): 647-661.

[Gionis, 2007] Gionis, Aristides, Heikki Mannila, and Panayiotis Tsaparas. "Clustering aggregation." *ACM transactions on knowledge discovery from data (TKDD)* 1.1 (2007): 4.

[Giorgi, 2017] Giorgi, F. M. aracne.networks: ARACNe-inferred gene networks from TCGA tumor datasets. R package version 1.4.0. DOI: 10.18129/B9.bioc.aracne.networks (2017).

[Gupta, 2014] Gupta, Shiv K., et al. "Discordant in vitro and in vivo chemopotentiating effects of the PARP inhibitor veliparib in temozolomide-sensitive versus-resistant glioblastoma multiforme xenografts." *Clinical cancer research* 20.14 (2014): 3730-3741.

[Gupta, 2016] Gupta, Shiv K., et al. "Delineation of MGMT hypermethylation as a biomarker for veliparib-mediated temozolomide-sensitizing therapy of glioblastoma." *JNCI* 108.5 (2016).

[Hanahan & Weinberg, 2011] Hanahan, Douglas, and Robert A. Weinberg. "Hallmarks of cancer: the next generation." *Cell* 144.5 (2011): 646-674.

[Hastie & Stuetzle, 1989] Hastie, Trevor, and Werner Stuetzle. "Principal curves." *Journal of the american statistical association* 84.406 (1989): 502-516.

[Hecker, 2009] Hecker, Michael, et al. "Gene regulatory network inference: data integration in dynamic models—a review." *Biosystems* 96.1 (2009): 86-103.

[Hirata, 2002] Hirata, Hiromi, et al. "Oscillatory expression of the bHLH factor Hes1 regulated by a negative feedback loop." *Science* 298.5594 (2002): 840-843.

[Hori, 2003] Hori, Shohei, Takashi Nomura, and Shimon Sakaguchi. "Control of regulatory T cell development by the transcription factor Foxp3." *Science* 299.5609 (2003): 1057-1061.

[Ikiz, 2015] Ikiz, Burcin, et al. "The regulatory machinery of neurodegeneration in in vitro models of amyotrophic lateral sclerosis." *Cell reports* 12.2 (2015): 335-345.

[Islam, 2014] Islam, Saiful, et al. "Quantitative single-cell RNA-seq with unique molecular identifiers." Nature methods 11.2 (2014): 163.

[Kaufman & Rousseeuw, 2009] Kaufman, Leonard, and Peter J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons, 2009.

[Kawashima, 1987] Kawashima, Ichiro, et al. "Characterization of pancreatic elastase II cDNAs: two elastase II mRNAs are expressed in human pancreas." *DNA* 6.2 (1987): 163-172.

[Ke, 2006] Ke, Qingdong, and Max Costa. "Hypoxia-inducible factor-1 (HIF-1)." *Molecular pharmacology* 70.5 (2006): 1469-1480.

[Kharchenko, 2014] Kharchenko, Peter V., Lev Silberstein, and David T. Scadden. "Bayesian approach to single-cell differential expression analysis." *Nature methods* 11.7 (2014): 740.

[Kitamura, 2013] Kitamura, Tadahiro. "The role of FOXO1 in β-cell failure and type 2 diabetes mellitus." *Nature reviews endocrinology* 9.10 (2013): 615.

[Kolodziejczyk, 2015] Kolodziejczyk, Aleksandra A., et al. "The technology and biology of single-cell RNA sequencing." *Molecular cell* 58.4 (2015): 610-620.

[Kramer, 2014] Krämer, Andreas, et al. "Causal analysis approaches in ingenuity pathway analysis." *Bioinformatics* 30.4 (2013): 523-530.

[Kushwaha, 2015] Kushwaha, Ritu, et al. "Interrogation of a Context-Specific Transcription Factor Network Identifies Novel Regulators of Pluripotency." *Stem cells* 33.2 (2015): 367-377.

[Lachmann, 2010] Lachmann, Alexander, et al. "ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments." *Bioinformatics* 26.19 (2010): 2438-2444.

[Lachmann, 2016] Lachmann, Alexander, et al. "ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information." *Bioinformatics* 32.14 (2016): 2233-2235.

[Langmead & Salzberg, 2012] Langmead, Ben, and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2." *Nature methods* 9.4 (2012): 357.

[Lasfargues & Ozzello, 1958] Lasfargues, Etienne Y., and Luciano Ozzello. "Cultivation of human breast carcinomas." *Journal of the national cancer institute* 21.6 (1958): 1131-1147.

[Lawrence, 2013] Lawrence, Michael, et al. "Software for computing and annotating genomic ranges." *PLoS computational biology* 9.8 (2013): e1003118.

[Lefebvre, 2010] Lefebvre, Celine, et al. "A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers." *Molecular systems biology* 6.1 (2010): 377.

[Lei, 2011] Lei, Liang, et al. "Glioblastoma models reveal the connection between adult glial progenitors and the proneural phenotype." *PloS one* 6.5 (2011): e20041.

[Levy, 2006] Levy, Carmit, Mehdi Khaled, and David E. Fisher. "MITF: master regulator of melanocyte development and melanoma oncogene." *Trends in molecular medicine* 12.9 (2006): 406-414.

[Li, 2009] Li, Heng, et al. "The sequence alignment/map format and SAMtools." *Bioinformatics* 25.16 (2009): 2078-2079.

[Li, 2010] Li, Long, Mark Leid, and Ellen V. Rothenberg. "An early T cell lineage commitment checkpoint dependent on the transcription factor Bcl11b." *Science* 329.5987 (2010): 89-93.

[Li & Dewey, 2011] Li, Bo, and Colin N. Dewey. "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." *BMC bioinformatics* 12.1 (2011): 323.

[Liang, 2005] Liang, Yu, et al. "Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme." *PNAS* 102.16 (2005): 5814-5819.

[Lin, 2008] Lin, Simon M., et al. "Model-based variance-stabilizing transformation for Illumina microarray data." *Nucleic acids research* 36.2 (2008): e11-e11.

[Lin, 2010] Lin, Yin C., et al. "A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate." *Nature immunology* 11.7 (2010): 635.

[Liu, 2017] Liu, Zehua, et al. "Reconstructing cell cycle pseudo time-series via single-cell transcriptome data." *Nature communications* 8.1 (2017): 22.

[Lotze, 2005] Lotze, Michael T., and Kevin J. Tracey. "High-mobility group box 1 protein (HMGB1): nuclear weapon in the immune arsenal." *Nature reviews immunology* 5.4 (2005): 331.

[Maaten & Hinton, 2008] Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research* 9.Nov (2008): 2579-2605.

[Macosko, 2015] Macosko, Evan Z., et al. "Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets." *Cell* 161.5 (2015): 1202-1214.

[Maiuri, 2007] Maiuri, M. Chiara, et al. "Self-eating and self-killing: crosstalk between autophagy and apoptosis." *Nature reviews molecular cell biology* 8.9 (2007): 741.

[Margolin, 2006] Margolin, Adam A., et al. "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context." *BMC bioinformatics*. Vol. 7. No. 1. BioMed Central, 2006.

[Mitrofanova, 2015] Mitrofanova, Antonina, et al. "Predicting drug response in human prostate cancer from preclinical analysis of in vivo mouse models." *Cell reports* 12.12 (2015): 2060-2071.

[Monti, 2003] Monti, Stefano, et al. "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data." *Machine learning* 52.1-2 (2003): 91-118.

[Murat, 2008] Murat, Anastasia, et al. "Stem cell–related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma." *Journal of clinical oncology* 26.18 (2008): 3015-3024.

[Newman & Girvan, 2004] Newman, Mark EJ, and Michelle Girvan. "Finding and evaluating community structure in networks." *Physical review E* 69.2 (2004): 026113.

[Nutt, 1999] Nutt, Stephen L., et al. "Commitment to the B-lymphoid lineage depends on the transcription factor Pax5." *Nature* 401.6753 (1999): 556.

[Nutt, 2003] Nutt, Catherine L., et al. "Gene expression-based classification of malignant gliomas correlates better with survival than histological classification." *Cancer research* 63.7 (2003): 1602-1607.

[Ohgaki & Kleihues, 2005] Ohgaki, Hiroko, and Paul Kleihues. "Epidemiology and etiology of gliomas." *Acta neuropathologica* 109.1 (2005): 93-108.

[Patel, 2014] Patel, Anoop P., et al. "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma." *Science* (2014): 1254257.

[Phillips, 2006] Phillips, Heidi S., et al. "Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis." *Cancer cell* 9.3 (2006): 157-173.

[Piovan, 2013] Piovan, Erich, et al. "Direct reversal of glucocorticoid resistance by AKT inhibition in acute lymphoblastic leukemia." *Cancer cell* 24.6 (2013): 766-776.

[Prentki & Nolan, 2006] Prentki, Marc, and Christopher J. Nolan. "Islet β cell failure in type 2 diabetes." *The Journal of clinical investigation* 116.7 (2006): 1802-1812.

[Ramsköld, 2011]Ramsköld, Daniel, et al. "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells." *Nature biotechnology* 30.8 (2012): 777.

[Repunte-Canonigo, 2015] Repunte-Canonigo, Vez, et al. "Identifying candidate drivers of alcohol dependence-induced excessive drinking by assembly and interrogation of brain-specific regulatory networks." *Genome biology* 16.1 (2015): 68.

[Rodriguez-Barrueco, 2015] Rodriguez-Barrueco, Ruth, et al. "Inhibition of the autocrine IL-6–JAK2–STAT3–calprotectin axis as targeted therapy for HR−/HER2+ breast cancers." *Genes & development* 29.15 (2015): 1631-1648.

[Rousseeuw, 1987] Rousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics* 20 (1987): 53-65.

[Rubinfeld, 1997] Rubinfeld, Bonnee, et al. "Stabilization of β-catenin by genetic defects in melanoma cell lines." *Science* 275.5307 (1997): 1790-1792.

[Saeys, 2007] Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. "A review of feature selection techniques in bioinformatics." *Bioinformatics* 23.19 (2007): 2507-2517.

[Semenza , 1998] Semenza, Gregg L. "Hypoxia-inducible factor 1: master regulator of O 2 homeostasis." Current opinion in genetics & development 8.5 (1998): 588-594.

[Shimono, 2009] Shimono, Yohei, et al. "Downregulation of miRNA-200c links breast cancer stem cells with normal stem cells." *Cell* 138.3 (2009): 592-603.

[Sonabend, 2013] Sonabend, Adam M., et al. "Murine cell line model of proneural glioma for evaluation of anti-tumor therapies." *Journal of neuro-oncology* 112.3 (2013): 375-382.

[Subramanian, 2005] Subramanian, Aravind, et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *PNAS* 102.43 (2005): 15545-15550.

[Sumazin , 2011] Sumazin, Pavel, et al. "An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma." *Cell* 147.2 (2011): 370-381.

[Szabo, 2000] Szabo, Susanne J., et al. "A novel transcription factor, T-bet, directs Th1 lineage commitment." Cell 100.6 (2000): 655-669.

[Talchai, 2012] Talchai, Chutima, et al. "Pancreatic β cell dedifferentiation as a mechanism of diabetic β cell failure." *Cell* 150.6 (2012): 1223-1234.

[Talos, 2017] Talos, Flaminia, et al. "A computational systems approach identifies synergistic specification genes that facilitate lineage conversion to prostate tissue." *Nature communications* 8 (2017): 14662.

[Thiery, 2009] Thiery, Jean Paul, et al. "Epithelial-mesenchymal transitions in development and disease." *Cell* 139.5 (2009): 871-890.

[Thiery, 2002] Thiery, Jean Paul. "Epithelial–mesenchymal transitions in tumour progression." *Nature reviews cancer* 2.6 (2002): 442.

[Thorgeirsson, 2002] Thorgeirsson, Snorri S., and Joe W. Grisham. "Molecular pathogenesis of human hepatocellular carcinoma." *Nature genetics* 31.4 (2002): 339.

[Tirosh, 2016] Tirosh, Itay, et al. "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq." *Science* 352.6282 (2016): 189-196.

[Trapnell, 2014] Trapnell, Cole, et al. "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells." *Nature biotechnology* 32.4 (2014): 381.

[Usoskin, 2015] Usoskin, Dmitry, et al. "Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing." *Nature neuroscience* 18.1 (2015): 145.

[Vallejos, 2017] Vallejos, Catalina A., et al. "Normalizing single-cell RNA sequencing data: challenges and opportunities." *Nature methods* 14.6 (2017): 565.

[Verhaak, 2010] Verhaak, Roel GW, et al. "Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1." *Cancer cell* 17.1 (2010): 98-110.

[Vu, 2016] Vu, Trung Nghia, et al. "Beta-Poisson model for single-cell RNA-seq data analyses." *Bioinformatics* 32.14 (2016): 2128-2135.

[Wang, 2017] Wang, Qianghu, et al. "Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment." *Cancer cell* 32.1 (2017): 42-56.

[Weinstein, 2002] Weinstein, I. Bernard. "Addiction to oncogenes--the Achilles heal of cancer." *Science* 297.5578 (2002): 63-64.

[Wu, 2014] Wu, Angela R., et al. "Quantitative assessment of single-cell RNA-sequencing methods." *Nature methods* 11.1 (2014): 41.

[Xu, 2005] Xu, Chunyan, Beatrice Bailly-Maitre, and John C. Reed. "Endoplasmic reticulum stress: cell life and death decisions." *The Journal of clinical investigation* 115.10 (2005): 2656-2664.

Part X

Appendices

Appendix A

scPLATE-Seq Protocol

A.1 Materials and equipment:

A.1.1 Lysis Buffer

- SUPERaseIn RNase inhibitor, 20U/µl (Ambion, #AM2696)
- Triton X-100, 10% (Sigma-Aldrich, #93443)
- dNTP set, 100mM each (Thermo Scientific, #R0181)
- HyClone water (GE Healthcare, #SH30538.01)
- Oligo(dT) mRNA capture primers, AAGCAGTGGTATCAACGCAGAGTAC[8 bp barcode]$N_{(8)}T_{(29)}V$, (IDT).
- Template switch oligo (TSO), AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG (IDT)

A.1.2 RT Mix

- Betaine, ultrapure (Alpha Aesar, #J77507)
- ProtoScript II reverse transcriptase (NEB, #M0368X)
- $MgCl_2$, 2M, cell culture grade (Fisher Scientific, #BP9741-10X5)
- Dithiothreitol (DTT), ultrapure (Invitrogen, #15508013)
- HyClone water (GE Healthcare, #SH30538.01)

A.1.3 Exo I digestion Mix

- Exonuclease I, 20U/µl (Thermo Scientific, #EN0582)
- HyClone water (GE Healthcare, #SH30538.01)

A.1.4 Pooling and Concentration

- Dynabeads MyOne Silane (ThermoFisher, #37002D)
- Buffer RLT Plus (Qiagen, #1053393)
- Ethanol, 200 proof (Sigma-Aldrich, #E7023)

A.1.5 Pre-amplification PCR Mix

- KAPA HiFi HotStart ReadyMix, 2X (Kapa Biosystems, #KK2612)
- SMART PCR primer, AAGCAGTGGTATCAACGCAGAGT

A.1.6 Nextera amplification

- Nextera PCR primer,
  AATGATACGGCGACCACCGAGATCTACACGCCTGTCCGCGGAAGCAGTGGTA
  TCAACGCAGAGT*A*C

A.1.7 Equipment

- Eppendorf twin.tec PCR Plate 96, semi-skirted (Eppendorf, #951020303)
- Microseal 'F' PCR Plate Seal, foil, (Biorad, #MSF1001)
- Speedball soft rubber brayer, 3.5’’ (Speedball, #004174)

- Eppendorf DNA LoBind microcentrifuge tubes, 1.5ml and 2ml (Eppendorf, #022431021 and #022431048)
- Agilent 2100 Bioanalyzer
- Agilent high-sensitivity DNA kit (Agilent, #5067-4626)
- Qubit 2.0 fluorometer (ThermoFisher)
- Qubit dsDNA high-sensitivity kit (ThermoFisher, #Q32854)
- Magnetic separators for 1.5ml and 15ml tubes (Invitrogen, #12321D and #12301D, or equivalent)
- Vortex mixer (Fisher #02-215-414, or equivalent)
- Multi-channel Pipettes (Integra Viaflow II, #4011, or similar)
- Eppendorf cold blocks (#022510541)

A.2 Procedure

Steps 1-6: Lysis plate preparation

1. **Critical:** At all steps, be meticulous to avoid contamination. Work in a PCR hood whenever possible. Wear disposable Tyvek sleeves or coat in the hood prior to amplification. All reagents must be free of DNA, RNA, and nucleases. Single-cell library quality can be significantly affected by small changes in reagents (concentration, age, or manufacturer). Avoid multiple freeze-thaw cycles (aliquot reagents). Prior to beginning, clean all surfaces to remove nucleic acid and nuclease contamination.
2. Prepare lysis buffer mix. For each sample prepare 6µl of lysis mix: 5.33µl Hyclone H2O, 0.15µl of 10% Triton X-100, 0.38µl of SUPERaseIN, and 0.15µl of 100mM dNTP mix. Make at least 1.2X the volume required.

- In the lysis and binding steps, the concentration of Triton X-100, SUPERaseIN, and dNTP mix will be 0.2%, 1U/µl, and 2mM, respectively.

3. Load 6µl of lysis mix into each well of a 96-well plate. (This, and subsequent steps, should be carried out with multi-channel pipettes or using a liquid handling robot.)
4. Add 1.5µl of 10µM mRNA capture primers into each well. Mix thoroughly.
5. Seal plates. If plates will be stored frozen, meticulously seal using Biorad Microseal F foil or -80C safe film.
6. Centrifuge plate briefly at 4C to collect liquid. Plates can be kept at 4C for same-day use or stored at -80C for up to 1 month.

Steps 7-9: Loading samples into lysis plates

7. Centrifuge plate briefly to defrost plates and collect liquid. Put plates in Eppendorf cold blocks.
8. Load samples into plates. NOTE: If using FACS to load plates: 1) carefully align sorter using same model plates. One method of alignment is to sort a visible number of droplets onto a plate sealed with optical film. Align sorter so the droplet is dead center at all four corner wells. 2) observe stream stability. Large or irregular cells can destabilize the stream and lead to empty wells.
9. Thoroughly seal plates with -80C safe foil (or optical seal) and spin down for 2min at full speed, 4C. NOTE: plates can be stored at -80C for up to two weeks.

Steps 10-15: Reverse transcription

10. Prepare reverse transcription (RT) mix. For each sample, prepare 7.5µl of RT mix. Combine 3 µl 5X RT buffer, 3 µl of 5M betaine, 0.45µl of 200mM $MgCl_2$, 0.075 µl of 1M DTT, 0.2 µl of 200U/µl ProtoScript II reverse transcriptase, 0.2 µl of 20U/µl SUPERaseIN, 0.15µl of 100µM template switch oligo, and 0.43µl of Hyclone water. Make at least 1.2X the volume required.
    - In the final 15µl reaction, the concentrations will be: 1M betaine, 1X RT buffer, 6mM MgCl2, 5mM DTT, 40U RT enzyme, 7.5U SUPERaseIN, and 1µM template switch oligo.
    - NOTE: DTT should be fresh or from frozen aliquot that has not been repeatedly freeze-thawed. 1M DTT may precipitate in the freezer, but can be resuspended by warming for a few minutes at room temperature or 37C, and then vortexing. The template switch oligo should be aliquoted and stored at -80C.
11. Centrifuge at room temperature, full speed, for 2min.
12. Run primer anneal program on thermal cycler. 72C (3min)→4C (infinite)
13. Add RT mix to plate and mix thoroughly. As much as possible, keep mix and plate near 0C throughout.
14. Apply seal and quickly spin plate at 4C.
15. Run RT program.
    - Stage 1: 42C (90min)
    - Stage 2: 10 cycles of: 50C (2min)→42C (2min)
    - Stage 3: 75C (10min)→4C (infinite)

Steps 16-18: Exonuclease I cleavage

16. Centrifuge plate briefly at 4C.
17. Dilute Exonuclease I to 7.5U/µl in 1X Exo I buffer and add 2µl (15U) to each sample. Mix thoroughly.
18. Run Exo I program on thermal cycler. 37C (30min)→80C (15min)→4C (infinite).

Steps 19-32: Pooling and cleanup

19. Centrifuge plate briefly at 4C.
20. Pool samples in 15ml polypropylene tube.
21. Clean up pooled samples with silane beads. In 15ml tube, combine sample (e.g., 1400µl) with cleanup mix:
    - 3.5 volumes RLT Plus (e.g., 4,900µl)
    - EtOH to 33% final (e.g., 2,771µl)
    - 3µl silane beads for each 300µl solution (e.g., 91µl).
22. Vortex to mix. Incubate at room temperature for 20min on rotator.
23. Collect beads on magnetic stand. Remove most supernatant, but leave ~1ml remaining.
24. Resuspend beads in remaining 1ml supernatant and transfer to 1.5ml tube.
25. Collect beads on magnetic stand. Remove supernatant.
26. Wash beads 2X with 500µl 80% EtOH.
27. Remove remaining EtOH with p10 tip and dry pellet in hood. Do not over dry.
28. Elute in 100µl Hyclone H2O. Vortex tube for 1min, and then flick or very briefly pulse the sample in a microcentrifuge. Incubate 1min at room temperature and then collect beads on magnet and remove the supernatant into two PCR tubes (47.5µl each).

29. Add 50µl 2X Kapa HiFi Hotstart and 2.5µl 10µM SMART PCR primer into each tube. Mix and quickly spin down PCR mix.
30. Run SMART PCR program on thermal cycler. NOTE: The number of amplification cycles required can vary from 13-20, depending on cell type and state. Because over amplification can lead to both PCR bias and recombination, it is important to use the least amount of amplification that yields a sufficient library. To determine this, run the first plate for either 14 cycles (large cell lines) or 18 cycles (small primary cells); clean up, and determine concentration using Qubit DNA HS. Adjust the number of cycles for the remaining plates.
    - Stage 1: 98C (3min)
    - Stage 2: 98C (20seconds)→67C (15seconds)→72C (5min)
    - Stage 3: 72C (5min)→4C (infinite)
31. Pool the two PCR reactions into a 1.5ml tube. Measure the total volume (slightly less than 200µl).
32. Clean up using Ampure XP with 0.7:1 ratio. Elute in 22µl Hylcone water. Remove 20µl to a new tube. Take care not to transfer any beads, which can ruin the Bioanlyzer run.

Steps 32-35: Library quantification and sequencing

33. Determine library concentration using 3µl of sample for Qubit HS analysis.
    - NOTE: in some cases libraries can successfully be sequenced from low concentration samples (e.g., less than 0.07ng/µl), but such libraries may not be visible in the Bioanalyzer, and the Nextera tagmentation reaction will have to be

loaded blind. It is generally preferable to *optimize the preamplification PCR to have a Qubit concentration of 0.1-0.4ng/µl*. Do not over amplify.

- NOTE: low concentration samples can be concentrated using a Speedvac. However, do not dry the sample—reduce the volume to ~10µl.

34. Determine average library size using Bioanalyzer DNA HS. Average fragment size should be 1,600-1,900bp.
    - NOTE: When running low concentration samples, the Bioanalyzer is very sensitive to contaminants, including beads, reagents, and air bubbles. Take extra care and include a blank well.
    - NOTE: Average fragment sizes below ~1,600bp often contain degradation. If the protocol has been followed closely, it is likely that such contamination was present in the cells prior to lysis. In some cases, such as tumor samples, such degradation may be unavoidable. Libraries with average fragment sizes as low as ~1,100bp have been successfully sequenced, but data quality is affected.
35. Perform Nextera XT DNA Library Prep with the following modifications:
    - Load 0.5ng of sample into the tagmentation reaction instead of the stated 1ng.
    - Use 5µl of Index 1 (i7) adapter/primer and 5µl of 5µM Nextera PCR primer
    - Following tagmentation and amplification, cleanup with 0.6X:1 ratio Ampure XP, followed by an additional cleanup using 1:1 ratio Ampure XP. Elute in 22µl RSB.
    - Run sample on Bioanalyzer DNA HS. Average fragment size should be between 300bp and 600bp. No primer dimers should be present.

Appendix B

Data and Software Availability for scPLATE-Seq

All described expression profiles have been deposited at the Gene Expression Omnibus (GEO) under accession number GSE104493.

Demultiplexing and UMI quantification scripts for scPLATE-Seq: https://github.com/califano-lab/scPLATE-Seq

Appendix C

Data and Software Availability for metaVIPER

scRNA-Seq data for the mouse glioblastoma model described have been deposited at the Gene Expression Omnibus (GEO) under accession number GSE95157.

R-package aracne.networks is available on Bioconductor (https://bioconductor.org/packages/release/data/experiment/html/aracne.networks.html).

Tirosh human cutaneous melanoma (SKCM) dataset was downloaded from GEO with accession number GSE72056.

SKCM, B and T lymphocyte interactomes: https://figshare.com/s/565815d2e44c7e3b2b36

Brain tumor interactomes: https://figshare.com/s/97fb0a95e9fc708b8c96

TCGA expression and somatic mutation profile: http://cancergenome.nih.gov/.

REMBRANDT data set: https://gdoc.georgetown.edu/gdoc/.

COSMIC somatic mutation profile: http://cancer.sanger.ac.uk/cosmic.

Filtered PBMC scRNA-Seq expression profiles generated using 10x Genomics V2 chemistry: https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.0.1/pbmc4k

Filtered PBMC scRNA-Seq expression profiles generated using 10x Genomics V1 chemistry: https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k

metaVIPER is implemented in viper function from Bioconductor R-package VIPER: https://www.bioconductor.org/packages/release/bioc/html/viper.html.

ARACNe algorithm: http://califano.c2b2.columbia.edu/aracne.

Appendix D

Data and Software Availability for iterClust

Filtered Chromium 10X Genomics public PBMC scRNA-Seq profiles were downloaded from https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc4k and https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/pbmc8k.

Tirosh human cutaneous melanoma (SKCM) dataset was downloaded from GEO with accession number GSE72056.

Clustering analysis benchmarking datasets were downloaded from https://cs.joensuu.fi/sipu/datasets/.

iterClust R package can be found on Bioconductor (https://bioconductor.org/packages/release/bioc/html/iterClust.html), as well as GitHub (https://github.com/hd2326/iterClust).