UNIVERSITY OF CALIFORNIA
SANTA CRUZ

DEEP LEARNING GENE REGULATORY NETWORK DYNAMICS FROM TRANSCRIPTOMES

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

DOCTOR OF PHILOSOPHY

in

BIOINFORMATICS AND BIOMOLECULAR ENGINEERING

by

Nathan Maulding

June 2024

The Dissertation of Nathan Maulding
is approved:

Professor Benedict Paten, Chair

Professor Vanessa Jonsson

Professor Josh Stuart

Marc Hafner, PhD

Dean Peter F. Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

Abstract

Deep Learning Gene Regulatory Network Dynamics from Transcriptomes

by

Nathan Maulding

Regulation of gene expression is critical for all life processes. RNAseq allows for the investigation of the expression state and the inference of gene regulatory networks (GRNs). The GRN reveals mechanisms of disease and development both at an organismal and cellular level. Constructing GRNs from expression data is done by detecting statistical relationships over a series of samples. However, these relationships are noisy and do not take into account many important variables. To explore one such limitation, I implemented dual RNAseq to investigate the regulatory mechanisms that occur between both host and virus transcriptomes in SARS-CoV-2 infected samples (Chapter 3). To address the temporal aspect of GRNs, I developed a statistical framework for detecting relationships between transcription factors and their targets from inferred trajectory progressions (”pseudotime”) in single-cells (Chapter 4). Finally, I developed an interpretable, time-aware deep learning transformer for modeling disease progression and demonstrate its ability to identify gene programs that promote breast cancer resistance post-treatment (Chapter 5).

Chapter 1

Acknowledgements

I have many people to thank for supporting me throughout my education. First and foremost I want to express gratitude to my family. My parents David and Diane, my sister Hope, and brother Noah have all been instrumental in my life and have been constants that I can rely on through my highs and lows. I am also grateful to have a great set of friends to inspire me both in my education and in pursuing adventure in life.

Josh Stuart, Lucas Seninge, Bianca Yue, and Chris Wong and their freely offered insights have often shown me the way to move beyond the current limitations I am facing to make a more impactful work. I also want to thank Marc Hafner for being an excellent mentor during my internship with Genentech. Finally, I want to express gratitude to my committee members, Benedict Paten, Vanessa Jonsson, Josh Stuart, and Marc Hafner, for taking an interest in my work and development.

To everyone I have leaned on throughout the years, thank you!

Chapter 2

Introduction

Eukaryotic life follows a common structure with DNA, RNA, and proteins taking the central role in the cell. Proteins, the units that define and coordinate a cell's functional capacity, are made as a result of translated gene expression. Expression of genes is a transient process that is influenced by the state of the cellular environment. Generegulation occurs throughout all biologic processes and is responsible for producing the unique cellular states and types. The past decade has seen a revolution in molecular biology made possible by developments in transcriptomic profiling through RNA sequencing [36]. The initial RNAseq technology, known as bulk RNAseq, which assessed cells in aggregate, have expanded research beyond anecdotal gene-level inquiries into systematic genome-wide views of cellular phenomena. [36]. RNAseq allows for the investigation of the expression state and the inference of gene regulatory networks (GRNs). The GRN reveals mechanisms of disease and development both at an organismal and cellular level. However, there are limitations to the current ability of transcriptomic methods to uncover the totality of the processes influencing the GRN of any given sample.The newer single-cell RNAseq (scRNAseq) approach, which detects transcriptome activity within individual cells, offers a higher resolution perspective on dynamic biological processes that would otherwise be obscured using bulk sequencing. Inferring GRNs using scRNAseq information shows much promise, but computational methods are still in development and represent an active area of research.

There can be multiple species within a single RNAseq that contribute to the GRN. This occurs commonly in studies of the gut microbiome, where there is an abundance of bacteria with independent transcriptomes that interact with the host [117]. Additionally, host-pathogen studies are also highly reliant on understanding how the two transcriptomes interact and regulate

each other at a genetic level to produce a GRN [117]. This has become an increasingly apparent need as many articles have been published on understanding how the SARS-CoV-2 virus causes COVID-19 by initiating a genetic response within the host [16, 185]. These studies highlight the conditional differences between healthy and disease states, but do not pursue how the expression patterns of the human and SARS-CoV-2 transcriptomes associate together to form a GRN in a disease state.

Multi-transcriptome quantification has been successfully implemented through dual RNAseq methods [178, 177], but tools have not been developed to integrate this information to infer GRNs. I applied dual RNAseq analysis to both cell lines and patient samples and used multiple correlative methods including average linkage dendritic clustering, Pearson correlation networks, and Pagerank network importance. I derive a consensus network implicated by these multiple views that sheds new light on the roles of human genes and pathways in SARS-CoV-2 infection (Chapter 3).

Chronology is also important for understanding cellular processes because GRNs are temporally ordered and controlled by regulators, such as transcription factors (TFs). TFs play a central role in governing the transitions between different cellular states. These transitions are driven by the dynamic regulation of gene expression orchestrated by TFs. The emergence of ”cell trajectory” inference methods has enabled the identification of transitions between different cell states [141, 133]. By analyzing cells in the context of their ”pseudotime”, the temporal regulation of genes and pathways along a transition “branch” from one cell state or type to another can be assessed. Datasets, such as [128], have been subjected to trajectory analysis. These pseudotime inferences revealed how the temporal expression patterns along the trajec-

tory lineage lead to myeloid progenitor commitment. Tools have been developed to detect and report differentially expressed genes along the trajectory branch. However, because this is an emerging field, tools that analyze gene expression patterns in trajectories are scarce and should be expanded.

I introduce a novel method for implicating TFs to cell trajectories called DREAMIT – 'Dynamic Regulation of Expression Across Modules in Inferred Trajectories' (Chapter 4). DREAMIT aims to analyze dynamic regulatory patterns along trajectory branches, implicating transcription factors (TFs) involved in cell state transitions within scRNAseq datasets. DREAMIT performs spline smoothing on the raw data followed by calculating metrics of association between the TF and its targets. To assess DREAMIT's performance, I used the TF-Marker database, which allowed us to determine its effectiveness in identifying TFs known to play essential roles in specific tissues, previously established as high-confidence markers for those tissues [187].

In chapter 5, I leverage TraCe-Seq technology [28] to track the molecular states and fitness of clones over time. To uncover differences associated with the selection process of particular clones post-treatment, I map their pre-treatment or intermediate transcriptional states to help us understand the gene programs that allow for that clone to be resistant. Methodologies to interpret such datasets are still in their infancy. For example, clustering approaches can fail to distinguish resistant from sensitive cell states as recent clonal barcoding systems have shown that a few genes, or even a single gene, can determine a cell's fate [138]. However, incorporating clonal fitness information through a tunable parameter has been shown to benefit the identification of features associated with cell outcomes [138].

I developed a semi-supervised approach with partial least squares regression (PLSR) to identify pre-treatment markers of innate resistance in TraCe-seq datasets. This approach incorporates both the transcriptional similarities of cells together with a phenotypic response to reveal signatures that might otherwise be hidden from traditional methods. Next, to identify gene programs associated with adaptive resistance, I introduce TraCSED (Transformer-based modeling of Clonal Selection and Expression Dynamics). TraCSED learns a dynamical process of clonal selection using a clone trajectory's inferred pseudotime from TraCe-Seq single-cell RNA-seq data as the time series input for model fitting. TraCSED uncovers interpretable gene programs that are directly related to the selection process and estimates time periods at which these may be critical for resistance.

Chapter 3

Chapter I: Multi-transcriptomic profiling and codifying SARS-CoV-2 characteristics

3.1 Background

The following dual RNAseq analysis of human and SARS-CoV-2 genes was published in *Scientific Reports* by Maulding et. al. [104]. At the time of this writing, the SARS-CoV-2 pandemic, which is caused by the coronavirus disease 2019 (COVID-19), has a global mortality rate that is still unknown [45, 113]. Its first appearance was reported in December 2019 and has since spread to 213 countries and territories and has caused over 100 million confirmed cases [45, 123, 83](worldometers.info). The SARS-CoV-2 infection has been reported to cause a variety of symptoms including fever, cough, fatigue, shortness of breath, and abnormalities in the chest as determined by CT [74, 33, 58]. Severe cases manifest with acute respiratory distress syndrome and lung injury, leading to morbidity due to damage to the alveolar lumen leading to inflammation and pneumonia [183, 189].

A variety of model systems and tissue types have been used to study the transcriptional response to SARS-CoV-2 infection in an effort to better understand the molecular basis of COVID-19 [16, 185]. These studies have revealed a cytokine, chemokine, and immune response to SARS-CoV-2 infection. This gene signature has been useful in understanding the biology of the COVID-19 disease. While differential expression of gene families has been thoroughly investigated in SARS-CoV-2 infection, there has been no attempt to date to correlate SARS-CoV-2 genes with the human host gene expression program.

We developed an unbiased multi-transcriptome read alignment pipeline to investigate the transcriptomes of human and virus together in the same sample. Previous studies analyzed only the human transcriptome, which provided important information about what genes and

pathways are differentially expressed between infected compared to uninfected samples. However, as illustrated by countless gene expression analyses from microarrays to RNA-seq, a complementary systems-level view can be achieved by looking at the co-expression of genes that may pick up on more subtle, but still significant associations missed by differential expression. Our pipeline leverages dual-RNAseq to quantify transcripts from both the host and pathogen together, which has shown promise in other systems [177, 178]. Dual-RNAseq originally required additional library enrichment to detect rare classes of transcripts. However, modern sequencing now yields a high enough read depth to provide accurate quantification of the entire host and pathogen transcriptomes without the need for additional library enrichment steps.

Thus, the main aim of the work presented here is to investigate the utility of the dual-RNAseq to study samples infected with the SARS-CoV-2 virus that makes it possible to correlate human genes with specific viral genes. We applied the analysis to both cell lines and patient samples and used multiple correlative methods including average linkage dendritic clustering, Pearson correlation networks, and Pagerank network importance. We derive a consensus network implicated by these multiple views that sheds new light on the roles of human genes and pathways in SARS-CoV-2 infection.

3.2 Method

3.2.1 Dual RNA-seq to simultaneously map virus-host transcriptomes

In order to quantify the host and pathogen transcriptomes, we implemented a dual RNA-seq [177, 178] analysis pipeline (dRAP). dRAP takes a series of reference FASTA files

and their corresponding GTFs and concatenates them into a single FASTA and GTF, which is subsequently used to create a mapping index. The human reference FASTA and ENSEMBL GTF for hg38 and the SARS-CoV-2 reference FASTA (`NC_045512v2`) and REFSeq GTF were downloaded from the UCSC Genome Browser [77]. RNAseq reads were trimmed using Trimmomatic [18] to filter out low quality and adapter sequences. STAR was then used with the parameters runMode='genomeGenerate', sjdbOverhang=100, and genomeSAindexNbases=6, to create a merged hg38/`NC_045512v2` index [41, 78]. Following the creation of the merged index, each sample was then mapped to the index using STAR to get transcription counts with the parameters outSAMtype, twopassMode, outFilterMultimapNmax, and quantMode set as 'BAM SortedByCoordinate', 'Basic', 1, and 'GeneCounts', respectively. This generated a 'ReadsPerGene' for each sample, which was then used as input for DESeq2 count normalization and differential gene expression (DGE) determination [93]. Gene expression levels, as read counts, were estimated and filtered by Cook's distance and nominal p values were corrected for False Discovery Rates (FDR) and a significance threshold was set at FDR < 0.05.

3.2.2 Differential expression (DE) analysis

To test whether dRAP is capable of detecting and quantifying host and pathogen transcripts, we used two previously published datasets on patients and cell lines infected with SARS-CoV-2 [16, 185]. NHBE and A549 cell lines (Wild-type n=3 and SARS-Cov-2 infected n=3 each) and postmortem Lung Biopsies (Healthy n=2 and SARS-CoV-2 infected n=2) were processed from the Tenoever data. Healthy (n=3) and SARS-CoV-2 infected (n=3) PBMC's and healthy (n=3) and SARS-CoV-2 infected (n=4) BALF patient samples were processed from the

Chen data. DESeq2 DGE of SARS-CoV-2 transcripts was determined for each tissue. Kendall rank correlations were calculated for the padj values of SARS-CoV-2 transcripts between each tissue type. Genes that did not qualify for statistics based on the DESeq2 cook's distance criteria were included as padj of 1, i.e. as the last rank in the set.

3.2.3 Comparing human differential expression between samples and cell lines

Because of the differences in SARS-CoV-2 transcripts detected between tissue and sample types, we wanted to determine if these differences also were reflected by the host transcriptional response. A one-to-one comparison of the log2 Fold-Change for each tissue against each cell line was plotted to determine common differential gene expression patterns. Each panel displays the log2 Fold-Change of one group versus the log2 Fold-Change of another by plotting genes that exceed significance and fold change thresholds in each group. Each plotted gene is sized according to its significance (-log10 adjusted p value) for the x-axis group and colored according to its significance for the y-axis group. Genes that were up-regulated and met threshold criteria for both groups were recorded as up-regulated matches. Figures display genes that met the threshold criteria of $|\text{Fold-Change}| > 1.5$ and $\text{padj} < 0.05$. A Chi-square test was then used to determine the likelihood of the distribution of genes with matched and mismatched expression directionality between the two groups. These DGE were then used as input for the gProfiler:GOSt functional profiling [137]. Common genes and ontologies in SARS-CoV-2 infected groups were recorded for use in downstream analysis.

3.2.4 Co-expression (CE) analysis

Because the differences in SARS-CoV-2 expression appeared to result in notable changes in host transcriptome response, we hypothesized that gene co-expression with SARS-CoV-2 may illuminate mechanisms of action and therapeutic targets. To this end, we utilized three different strategies as part of a Co-Expression (CE) analysis. These included analyses of genes demonstrating high correlation with SARS-CoV-2 (CE-Net), genes clustered with SARS-CoV-2, and genes with high weighted-PageRank influence (CE-PageRank).

For groups infected with SARS-CoV-2, genes passing Cook's distance filtering with DESeq2 were clustered using average linkage with the R package 'hclust' [93]. The resulting dendritic tree was then cut into 200 clades of co-expression. Clades containing SARS-CoV-2 transcripts were then further subdivided into 5 clades. After this final subdivision, co-regulated genes participating in clades containing SARS-CoV-2 transcripts were used as input for gProfiler:GOSt functional profiling [137]. The input genes and resulting pathways were separated and used in downstream analysis.

DE genes with Benjamini–Hochberg corrected p values less than 0.05 were subsetted for PageRank analysis. We used the topological overlap matrix (TOM) 43 generated from a Pearson correlation matrix with a soft thresholding parameter of 30 to create a weighted gene network using pairwise complete observations and then ran weighted PageRanks with a damping parameter of 0.9 on each of the 5 sample groups [190]. Genes with PageRanks in the top 80th percentile of NHBE, A549, and BALF sample groups were each used as input for gProfiler:GOSt functional profiling [137]. The input genes and resulting pathways were

separated and used in downstream comparisons.

To explore co-regulated and influential components in the infection, genes with significant DE were correlated with SARS-CoV-2 transcripts for each group. Genes with an R^2 relationship > 0.95 with a SARS-CoV-2 gene were used as input for gProfiler: GOSt functional profiling. The input genes and resulting pathways were separated and used in downstream analysis. To visualize the highest correlated elements, the 'network' package in R [20] was used to plot an edge between a SARS-CoV-2 transcript and a gene if $R^2 > 0.98$. The size of the node in the network illustrates the degree of connection each node has relative to the other nodes in the network. It should be noted that, by virtue of the methodology, SARS-CoV-2 transcripts will have a higher degree of connectivity.

3.2.5 Sankey Diagrams

To visualize how the four views affirm or disagree with one another, we created Sankey diagrams. For Fig. 4B, the four views were connected to genes found in at least two of the four views. Genes were colored based on the views that they were found in: white, light blue, and dark blue indicate that DE and 1, 2, or 3 other views, respectively, and red for a gene that is not found by DE, but was by a CE view. These genes are then connected to themes, which consists of a list of gProfiler pathways in which the genes were found to be enriched (Suppl. Table 1). For Suppl. Figure 10, the four methods are connected to genes found by any of the four views and these genes.

3.2.6 Network Construction

Finally, to visualize the gene and pathway agreement between the four views, the recorded genes and gProfiler pathways from SARS-CoV-2 infected groups were used to create gene to pathway networks for each view and for results that appear through multiple views. To be in these networks, an inclusion criteria was enforced where each gene and pathway displayed is found in at least two of the three SARS-CoV-2 infected sample types, namely BALF patient samples, NHBE infected cells, and A549 infected cells. An edge was created between genes and ontologies if they were implicated together by the gProfiler:GOSt result.

3.2.7 Multi-Transcriptomic Profiling

This analysis contributed the recent publication by Maulding et. al. [104]. To test the dual RNA-seq approach for investigating SARS-CoV-2 infection, we built a dual RNA-seq analysis pipeline (dRAP) to map all transcripts of infected cells to either the host or viral genomes in an unbiased manner (Fig. 1A; see "Methods" section). We hypothesize quantifying both host and viral transcripts might enable a more sensitive and specific association of host pathways responsive to SARS-COV-2 (SARS-CoV-2) infection. To our knowledge, this represents the first attempt to estimate both host and virus transcription simultaneously in the same sample for SARS-CoV-2.

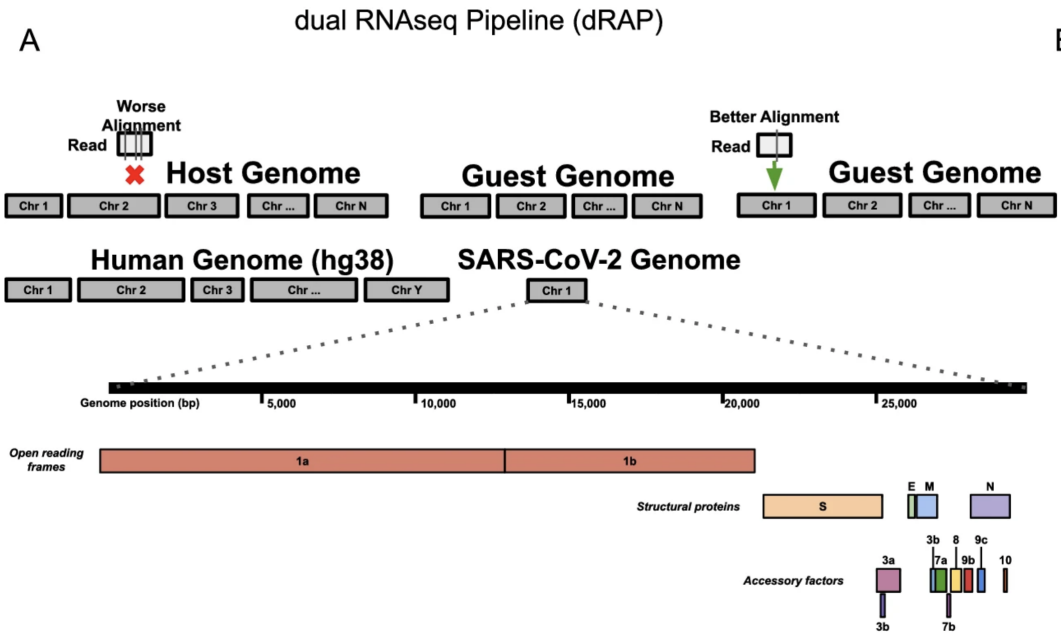


Figure 3.1: (A) Illustration of the dual RNAseq alignment pipeline (dRAP) using both host and guest genomes. Dual RNAseq creates a single reference index for RNAseq read alignment with STAR by appending "guest" genomes to the host genome allowing for simultaneous alignment of reads with multiple transcriptomes where the best overall alignment is selected. An example is shown where dRAP can resolve that a single read has a better match to the guest genome with only 1 mismatch (green arrow) compared to the host genome with 3 mismatches (red vertical lines). The SARS-CoV-2 guest genome (`NC_045512v2`) is depicted with its annotated set of genes designated as open reading frames, structural proteins, or accessory factors.

3.3 Results

3.3.1 Robust viral transcript detection in human cell lines

As a preliminary test, we first applied dRAP to previously published data on the analysis of human cell lines for which the amount of infected virus was experimentally controlled (Table 1; Fig. 2A). In the study that produced the original data, Blanco-Melo et al. [16] exposed a variety of human respiratory cell lines with SARS-CoV-2 virus. In that study, the authors found a robust host transcriptional response to infection for an ACE2 receptor-enhanced alveolar basal epithelial cell line (A549) as well as a bronchial epithelial cell line (NHBE). As the A549 and NHBE cell lines show robust host response and have established viral levels, we reanalyzed the data with dRAP to jointly analyze both the host and viral transcriptomes.

Importantly, dRAP provides a quantification of the amount of viral transcript in a host cell, which could be leveraged for downstream correlation-based analyses to implicate pathways of host response (Fig. 1B). Thus, while there is high concordance of the particular transcripts detected, dRAP reports a range of fold changes, with more dramatic overexpression detected in the A549 cell line compared to NHBE, potentially due to contamination of the mock-treated NHBE cells with SARS-CoV-2. Whether the differences are due to technical artifact or biological factors, these observations support the idea that dRAP's quantifiable differences can serve as the basis for studying regulation dynamics associated with infection by jointly analyzing the viral and host transcripts together.

The relative statistical significance of genes was also found to be consistent between the cell lines (Kendall rank correlation 0.69, $p < 0.01$). For example, the genes core to the infec-

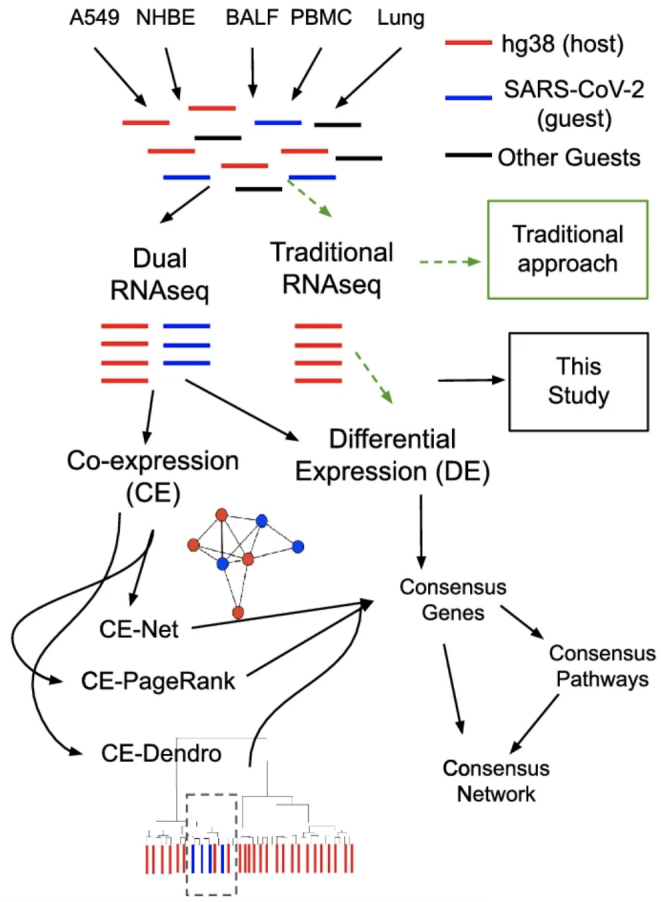


Figure 3.2: (B) Overview of dRAP application to SARS-CoV-2 analysis enabling the detection of coexpression associations between transcripts originating from human (red lines) and virus (blue lines). RNAseq reads from SARS-CoV-2 infected samples from human cell lines (A549 and NHBE) and patients (BALF, PBMC, and Lung) were collected from public datasets (Blanco-Melo et. al. 2020 and Xiong et. al. 2020 [16, 185]). Traditional RNAseq (dashed green arrows), which does not quantify both human and viral transcripts, allows for only differential expression analysis, while dRAP (black arrows) enables both downstream differential and co-expression analyses between host and virus. A549, NHBE, and BALF samples were selected for downstream analyses as they contained SARS-CoV-2 transcripts. Several co-expression analyses (CE) were performed in linear (CE-Net) and nonlinear (CE-dendro) relationships between the human and SARS-CoV-2 transcriptomes and genes of influence (CE-PageRank) in the gene regulatory network. Each CE view, along with DE, produced a set of results for A549, NHBE, and BALF groups. Genes implicated in two or more CE views were collected and used to determine enriched pathways (denoted "consensus pathways"). A "consensus network" was determined by including genes and pathways found by two or more views.

tion of the virus—ORF10, S, N, and M—were found to be the highest expressed genes in both A549 and NHBE. Out of the eleven SARS-CoV-2 transcripts, eight were detected consistently in both cell lines, suggesting viral transcripts can be determined robustly by dRAP in multiple host cell types. In addition, there is significant concordance of differentially expressed human genes between SARS-CoV-2 and RSV infections (Chi-Square 45.1, $p < 1e-10$) consistent with the report by the original authors11 (Suppl. Figure 1).

3.3.2 Robust detection of viral transcripts in human cell lines and BALF tissue

We searched for human cells that express the most consistent levels of SARS-CoV-2 virus. We expected that lung tissue would show the highest amount of viral transcription based on previous reports that the virus invades alveolar epithelial cells [67]. However, to our surprise, the most striking finding was that many SARS-CoV-2-infected patient samples from the lung or blood showed no SARS-CoV-2 expression. Specifically, PBMC samples had one patient that showed a normalized transcript count of 1.1 for ORF1ab and S, while all other samples and all SARS-CoV-2 genes had transcript counts of 0 (Table 1; Fig. 2A). No SARS-CoV-2 gene passed DESeq2 Cook's distance filtering for the PBMC samples. Similarly, lung biopsy samples showed inconsistent SARS-CoV-2 expression (Table 1; Fig. 2A). One patient showed SARS-CoV-2 expression while the other patient had none. However, even for the patient that did display SARS-CoV-2 gene expression, the expression levels were far less robust than that observed in cell lines.

In stark contrast with these other patient samples, SARS-CoV-2 patient samples from Bronchoalveolar lavage fluid (BALF) showed very robust SARS-CoV-2 expression that even

	BALF		A549		NHBE		PBMC		Lung	
	Padj	Log2(FC)	Padj	Log2(FC)	Padj	Log2(FC)	Padj	Log2(FC)	Padj	Log2(FC)
Cov2_ORF10	1.11E-32	18.3	2.11E-17	10.9	7.83E-177	4.7	DNQ	DNQ	1.28E-05	11.9
Cov2_N	2.78E-50	21.4	5.83E-26	12.8	2.11E-58	6.9	DNQ	DNQ	1.75E-05	10.1
Cov2_ORF8	7.45E-41	18.3	1.60E-11	9.7	5.93E-17	4.3	DNQ	DNQ	DNQ	9.6
Cov2_ORF7b	2.51E-04	11	DNQ	0.3	DNQ	-1.3	DNQ	DNQ	DNQ	DNQ
Cov2_ORF7a	4.86E-37	18.6	2.02E-13	9.7	8.81E-17	5.5	DNQ	DNQ	DNQ	9.3
Cov2_ORF6	2.81E-17	14.9	DNQ	5.9	DNQ	1.7	DNQ	DNQ	DNQ	DNQ
Cov2_M	1.51E-42	19.1	1.71E-16	10.6	2.52E-21	6.5	DNQ	DNQ	DNQ	9.6
Cov2_E	1.20E-24	15.2	DNQ	7.7	DNQ	3.6	DNQ	DNQ	DNQ	DNQ
Cov2_ORF3a	8.40E-40	19	8.07E-14	9.8	3.12E-13	7	DNQ	DNQ	DNQ	9
Cov2_S	1.16E-58	22.7	5.49E-18	11.1	2.24E-73	6.3	DNQ	0.4	DNQ	9.9
Cov2_ORF1ab	4.04E-69	24.7	5.01E-13	9.6	1.34E-20	4.8	DNQ	0.4	1.57E-03	10.6

Figure 3.3: SARS-CoV-2 differential gene expression for infected patient tissue and cell line samples compared with non-infected samples. Patient tissue types show dramatically different expression profiles with PBMC and Lung biopsy tissue rarely ever passing detection limits while BALF tissues show robust expression in infected patients. Cell lines also display strong SARS-CoV-2 expression although the magnitude of fold change was far less than that observed in BALF samples. “DNQ” stands for “did not qualify”, which indicates genes that did not pass Cook’s distance filtering in DESeq2 analysis.

exceeded the levels observed in infected cell lines (Table 1; Fig. 2A). The most significant SARS-CoV-2 gene was ORF1ab, followed by the S, N, and M genes. The overall profile of BALF samples had a few notable differences from that of the infected cell lines, including a much more significant overexpression of the SARS-CoV-2 ORF1ab gene and that the E, ORF6, and ORF7b genes were also significantly overexpressed in BALF samples but not in cell lines. Outside of these differences, the cell lines show similar features to the BALF SARS-CoV-2 profile at a lower expression level, including the predominant overexpression of the S, N, and M genes. Overall, the relative significance of SARS-CoV-2 genes were highly concordant between BALF and A549 (Kendall rank correlation 0.49, $p < 0.05$) and NHBE (Kendall rank correlation 0.56, $p < 0.05$) samples. In contrast, BALF and PBMC (unable to compute Kendall rank correlation because all transcripts were undetected in PBMC samples) and Lung (Kendall rank correlation 0.23, p=0.37) were shown to be discordant. Because of the similarities between BALF patient samples and cell lines and the absence of a robust SARS-CoV-2 expression profile in PBMC and lung biopsy samples, we used BALF, NHBE, and A549 samples in the following host-virus joint analysis to implicate host pathways associated with SARS-CoV-2 infection.

3.3.3 Concordant expression changes in BALF and cell lines

Given that the virus is detected in BALF and not PBMCs and lung, we reasoned that host expression changes in BALF would reflect a higher degree of direct responses to infection, compared to lung and PBMCs. To test this, we asked if the BALF DE genes were more comparable to the cell line DE genes relative to the DE genes derived from PBMCs and lung.

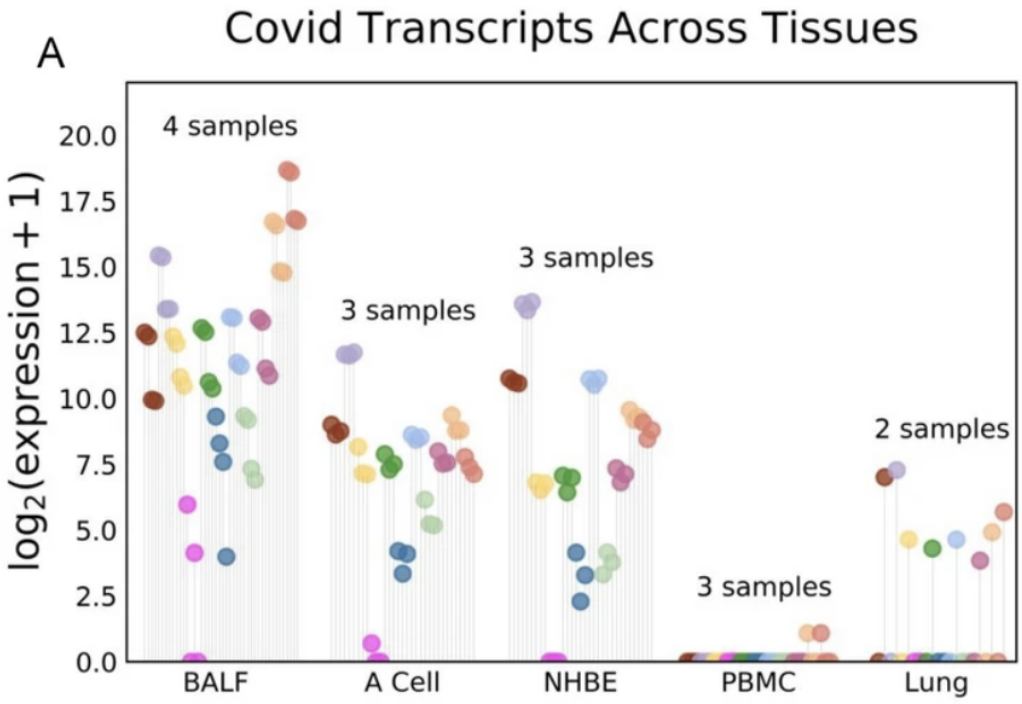


Figure 3.4: (A) dRAP is sensitive enough to detect subtle differences in SARS-CoV-2 transcripts quantities resulting in differential expression within the SARS-CoV-2 transcriptome. SARS-CoV-2 expression is also shown to be highly dependent on the system being studied. Patient BALF samples show high amounts of SARS-CoV-2, while PBMC and Lung patient samples display low or no SARS-CoV-2.

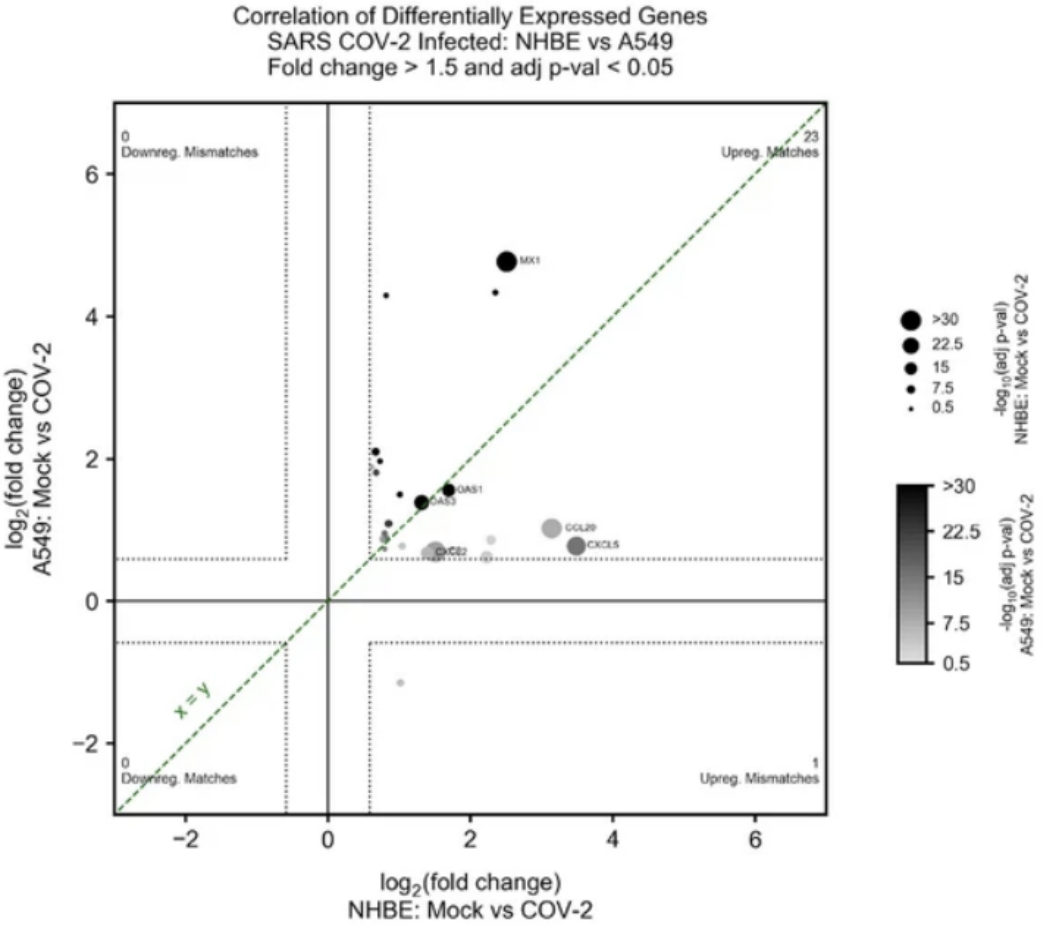


Figure 3.5: Log2 fold change comparison between differential expressions in infected samples against non-infected samples shows that the tissue specificity of SARS-CoV-2 extends to the degree of concordance observed in the human transcriptome. (B) Statistically significant concordance was observed between NHBE and A549 cell lines ($p < 1e\text{-}5$).

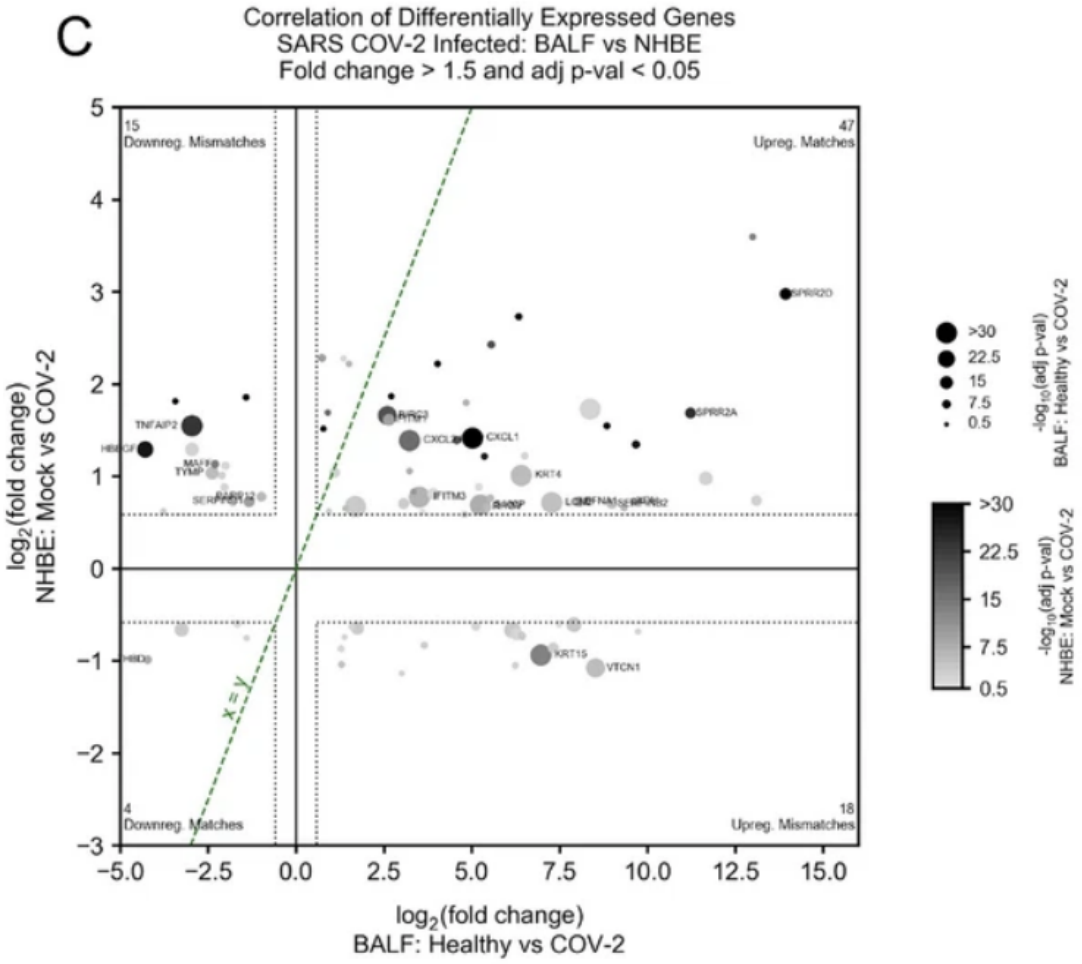


Figure 3.6: Log2 fold change comparison between differential expressions in infected samples against non-infected samples shows that the tissue specificity of SARS-CoV-2 extends to the degree of concordance observed in the human transcriptome. (C) Statistically significant concordance was observed between BALF patient samples with NHBE ($p < 0.05$).

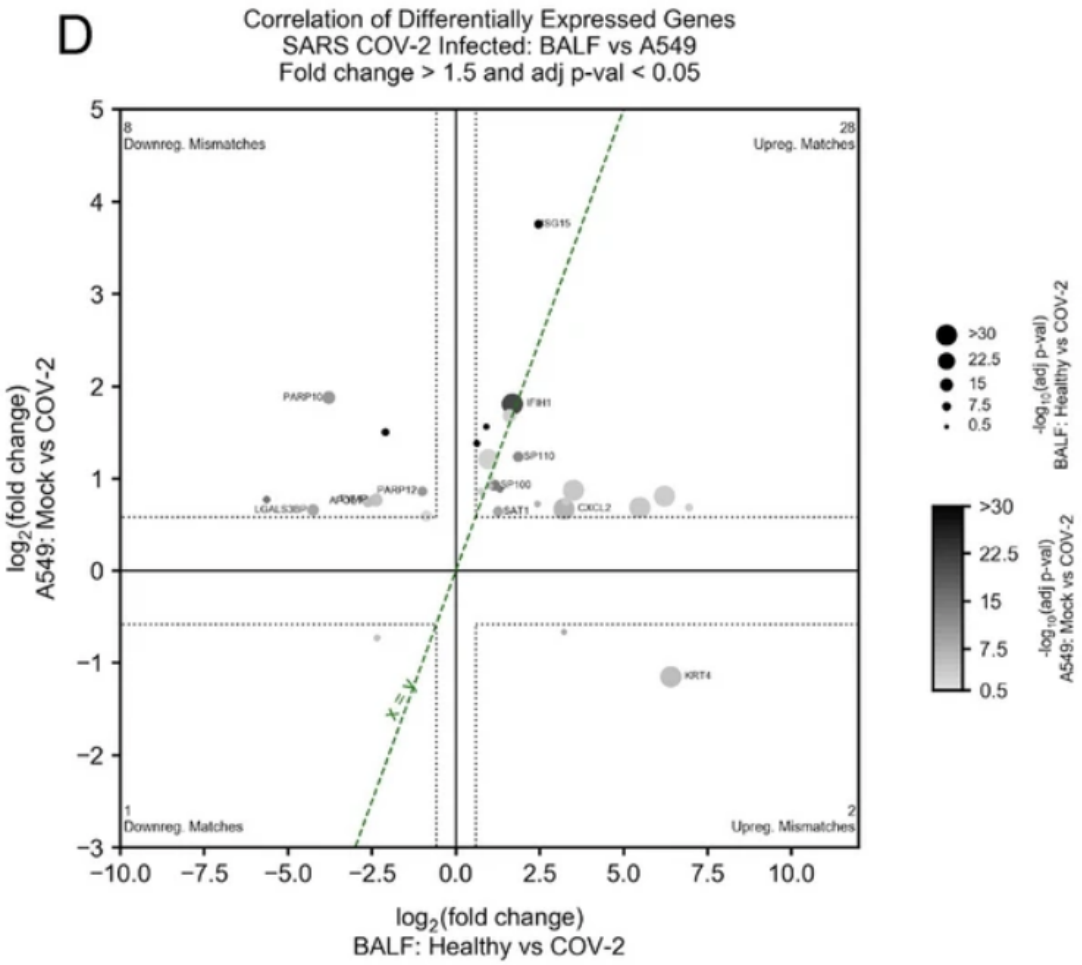


Figure 3.7: Log2 fold change comparison between differential expressions in infected samples against non-infected samples shows that the tissue specificity of SARS-CoV-2 extends to the degree of concordance observed in the human transcriptome. (D) Statistically significant concordance was observed between BALF patient samples and A549 cell lines ($p < 0.01$).

The total expression profile of these samples were then compared against each other to determine common and diverging signatures in the various patient tissue samples and cell lines. First, the profile of the NHBE and A549 cell lines were compared (Fig. 2B). Overall, there were 23 matching directional expression changes and 1 mismatched directional expression change in genes meeting Fold Change and p value thresholds for both NHBE or A549 cells, demonstrating a significant concordant relationship (Chi-Square 20.2, $p < 1e\text{-}5$). Genes showed a common up-regulation between the two cell lines, which is clearly observed in the plot through changes in size and color indicating increasing significance. Some of the most significant changes were observed in genes with roles in antiviral response (MX1, IFI27, IRF9, OAS1, OAS3), and chemokine signaling (CXCL5).

Comparing BALF patient samples with NHBE cells there were 51 matching directional expression changes and 33 mismatched directional expression changes (Fig. 2C), demonstrating significant concordance (Chi-Square 3.9, $p < 0.05$). Most genes show common up-regulation (47 genes), including SPRR2D, SPRR2A, PLAT, CXCL1, and CXCL2. The comparison of BALF with A549 cells produced a similar result (Chi-Square 9.3, $p < 0.01$) (Fig. 2D) with 29 matching directional expression changes and 10 mismatched directional expression changes.

Next, we compared BALF with lung and PBMC samples. BALF and lung samples were found to have 16 matched directional expression changes and 23 mismatched directional expression changes, demonstrating that the majority of gene expression changes are discordant, but without reaching statistical significance(Chi-Square 1.3, p=0.26) (Fig. 2E). In a more extreme manner, BALF and PBMC samples showed 275 matched directional expression changes

and 683 mismatched directional expression changes, demonstrating that the majority of gene expression changes are very significantly discordant (Chi-Square 173.7, $p < 1e\text{-}38$) with BALF samples (Fig. 2F). Therefore, for both PBMC and lung samples there was a dominant discordance in gene expression changes compared to BALF. This is consistent with our observations above in which we found robust SARS-CoV-2 expression levels in BALF compared with undetectable levels in PBMC and lung.

3.3.4 Multi-view coexpression reveals a human transcriptional network associated with SARS-CoV-2 transcripts

The estimates of viral and host RNAs for the same samples provided by dRAP enable investigating host regulatory pathways through a coexpression (CE) analysis to identify human transcripts most correlated with viral transcripts. CE could complement DE to find human genes directly associated with viral infection by detecting more subtle patterns of transcripts fluctuating concordantly with particular viral products that may suggest regulatory connections between the viral and host genes. The BALF human tissue samples and the two cell lines were used for CE as the above analyses found these samples provided robust viral RNA expression.

To prioritize candidate genes and pathways, we elected to use three different CE strategies to see if any genes and pathways were nominated by one (or more) viewpoint(s). First, we collected the top k most correlated human transcripts using Pearson correlation for each viral protein and saved the union of these human proteins. Next, we performed a hierarchical clustering analysis on the Pearson correlation matrix to determine if there existed a clade, or clades, enriched for viral proteins. Third, we performed a PageRank analysis on the Spear-

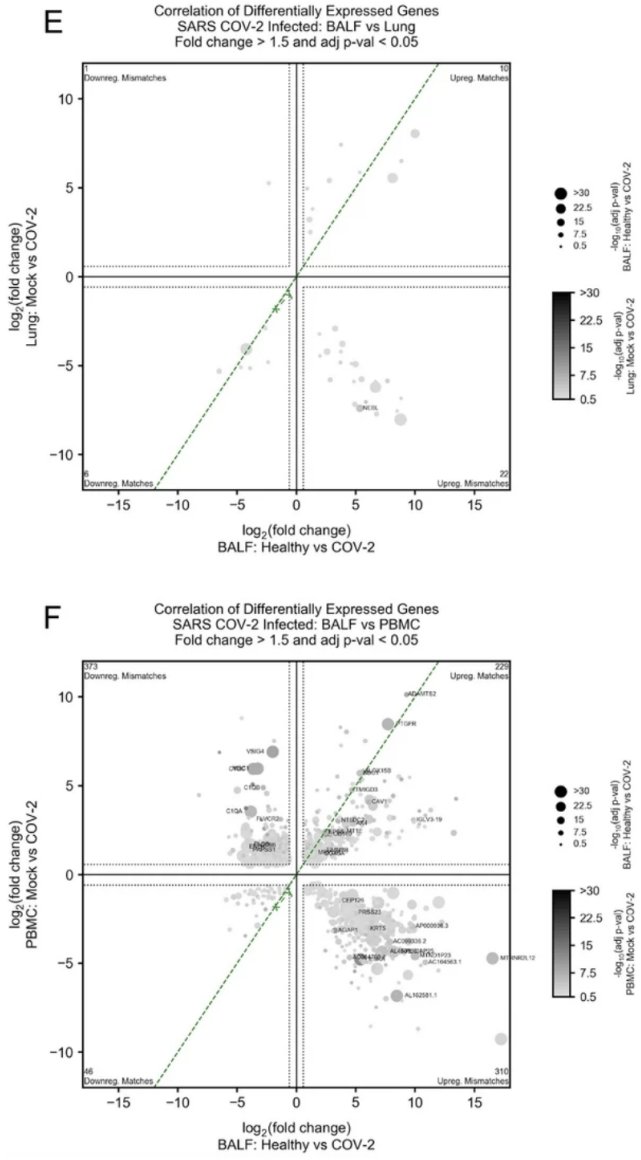


Figure 3.8: Log2 fold change comparison between differential expressions in infected samples against non-infected samples shows that the tissue specificity of SARS-CoV-2 extends to the degree of concordance observed in the human transcriptome. However, there was no concordance observed in BALF versus Lung ($p = 0.26$) (E) and a significant discordance versus PBMC ($p <$ 1e-38) (F) patient samples. This suggests that the lack of SARS-CoV-2 expression observed in Lung and PBMC samples is also associated with significantly altered human expression, making these tissue types not ideal for learning SARS-CoV-2 mechanisms.

man correlation matrix to determine which host genes are central to the host-virus correlation network. Candidate host genes were collected from each of these approaches, compared to each other as well as to the list of genes from DE analysis.

For the first CE view, we took the entirety of normalized transcript counts passing DESeq2 Cook's distance filtering and clustered their expression using an average linkage distance between pairwise Pearson correlations to create dendrograms with various clades [110]. BALF samples and NHBE and A549 cell lines infected with SARS-CoV-2 all had extremely localized co-expression of SARS-CoV-2 genes with each other. This is evidenced by the fact that in NHBE cells (Fig. 3A,B) and BALF samples (Suppl. Figure 2A) all SARS-CoV-2 genes passing Cook's distance were co-expressed in the same clade as each other. A549 cells were similar with the exception of the SARS-CoV-2 N gene, which was located in an adjacent clade (Suppl. Figure 2B). The CE-dendro analysis clustered some genes with SARS-CoV-2 genes (Suppl. Figure 3) that were not implicated by differential expression analysis as they failed to pass Benjamini–Hochberg corrected significance tests. These include MYC, NFKBIA, and DDX1 that are implicated in proto-oncogenic pathways [31], immune response in lungs [4], and host cofactor enhancement of SARS-CoV-1 replication [186], and thus of possible relevance to SARS-CoV-2 mechanisms.

The second CE view used genes with significant differential expression (padj < 0.05) to create a network of genes with an $R^2 > 0.98$ with a SARS-CoV-2 gene (see "Methods" section). The NHBE and A549 networks have a very similar architecture, whereas the BALF network displays a much denser network of expression (Suppl. Figures 4–6). All three of these networks display very similar gene signatures of Chemokines, SPRR's, S100's, viral response,

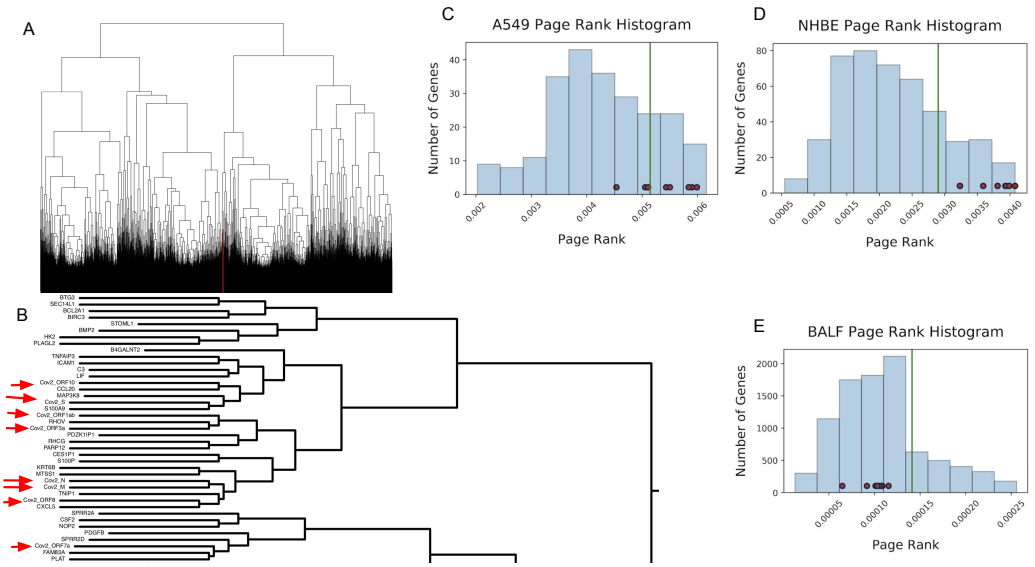


Figure 3.9: By clustering expression SARS-CoV-2 and human expression patterns concurrently we observed that SARS-CoV-2 transcripts were localized in a small clade visualized in red (A). This clade of coexpression with SARS-CoV-2 transcripts contains a set of genes associated with SARS-CoV-2 mechanisms in infection (B). The histogram distribution of PageRank values for the A549 (C) and NHBE (D) cell lines shows that the SARS-CoV-2 genes are highly influential. However, in BALF samples (E), SARS-CoV-2 genes are at the lower end of the PageRank distribution likely due to the numerous differentially expressed genes creating a much larger set than that for the cell lines. In A-C, the green line marks the 80th percentile in the distribution and the small red nodes along the distribution represent SARS-CoV-2 genes.

and interferon response genes. The CE-Net for the strong linear relationships found between SARS-CoV-2 genes and human genes displays a very modularized and sparse architecture of particular genes localized in groups of pathways (Suppl. Figure 7). This network layout enabled visually extracting gene subgroups that potentially coordinate to elicit specific responses. A few examples of important gene subgroups in Suppl. Figure 7 include the gene modules indicative of a lipopolysaccharide response and chemokine/cytokine activity (CXCL5, CXCL8, CCL20, HIF1A), cornification and epithelial cell differentiation (SPRR2A, SPRR2D, SPRR2E, PI3, KRT6B, ESF1, RHCG, MTSS1),and antiviral response (OAS1, MX1 and PARP9, DTX3L).

The final CE view used weighted PageRank to find genes with high influence among a set of significant DE genes. In A549 cells, many SARS-CoV-2 genes (S, ORF3a, ORF1ab, ORF7a, ORF10) ranked above the 80th percentile of influence among significant DE genes (Fig. 3C). Similarly, SARS-CoV-2 genes (M, S, N, ORF8, ORF10, ORF3a, ORF7a, ORF1ab) for the most part ranked above the 80th percentile of influence (Fig. 3D). However, in BALF patient samples, no SARS-CoV-2 were above the 80th percentile of influence (Fig. 3E). One likely reason for this would be that the number of DE genes is over 20 times more numerous than either of the cell lines. Therefore, the gene regulatory network driving the conditional response is hidden behind more noise, likely due to the inherent increase in variability between patient samples compared with cell lines. This suggests that many of the DE genes in BALF samples may be more indicative of patient variability than of a response to SARS-CoV-2 infection. CE for weighted PageRank shows a similarly sparse and modular architecture to that observed by CE-Net (Suppl. Figure 8). However, the subgroups had some notable differences in the participating modules, which consisted of the orange module (S100P, PROS1, PTAFR), pink module (ICAM1, HBEGF, INHBA), purple module (CXCL5, ASS1, DTX3L, BIRC3, IFIH1), and black module (MAFF, HDGF, CSNK1E, SAMD9, ITGA5, MCFD2). Finally, the CE-PageRank network (Suppl. Figure 8) shows a variety of unique, modularized results including a purple module (CXCL5, ASS1, DTX3L, BIRC3, IFIH1)(which was similar to the red and blue modules observed in the consensus network (Fig. 4D–E)), a pink module (ICAM1, INHBA, HBEGF), an orange module (PROS1, PTAFR, S100P), and a black module (MAFF, HDGF, CSNK1E, SAMD9, ITGA5, MCFD2). Genes of potential import to highlight include ICAM1, which has been implicated in viral entry and survival [122], S100P, implicated in cal-

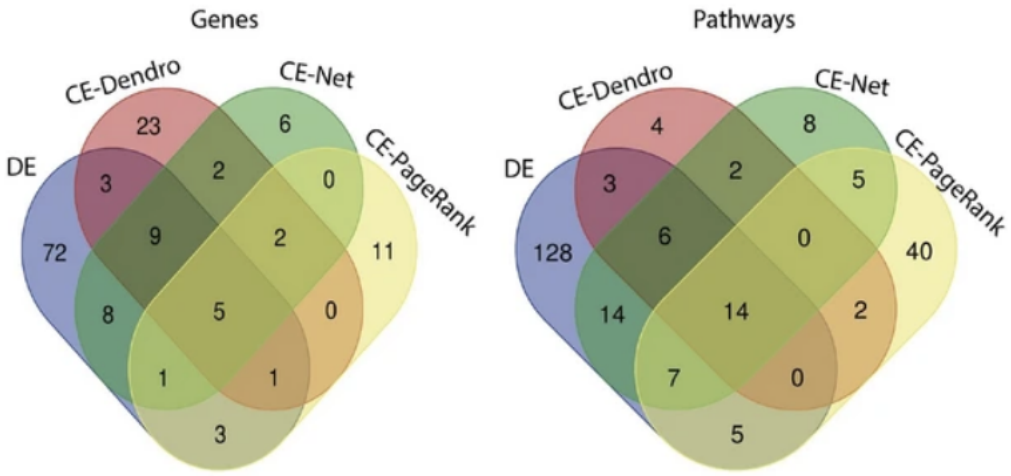


Figure 3.10: The overlap in gene and pathway results indicated by the DE comparison, Dendrograms, pearson correlation networks, and PageRank methods by venn diagram.

cium binding and human papillomavirus (HPV) [144], and SAMD9, shown to have antiviral properties [118].

Relative to DE, the CE views produce some common, but primarily distinct, genes and pathways (Fig. 4A). We collected 102 genes from the DE analysis that were implicated by two out of the three separate DE analyses run on BALF, NHBE or A549 (see "Methods" section; Suppl. Figure 9). The pathways in the core of the DE results were highly consistent and reminiscent of those described previously by the original authors [16]. We found that the candidate genes suggested by the three different CE approaches are distinct from the DE results (overlaps of 18, 23, and 10 with the CE-Dendro, CE-Nets, and CE-PageRank results, respectively). Of the 102 DE genes, a majority (72 genes) were only indicated through DE, suggesting that DE produces a large number of results that are exclusively a statistically significant conditional response and are not indicative of linear correlation to the source (CE-Nets), co-regulation with the source (CE-Dendrogram), or the most influential players in the geneset (CE-PageRank).

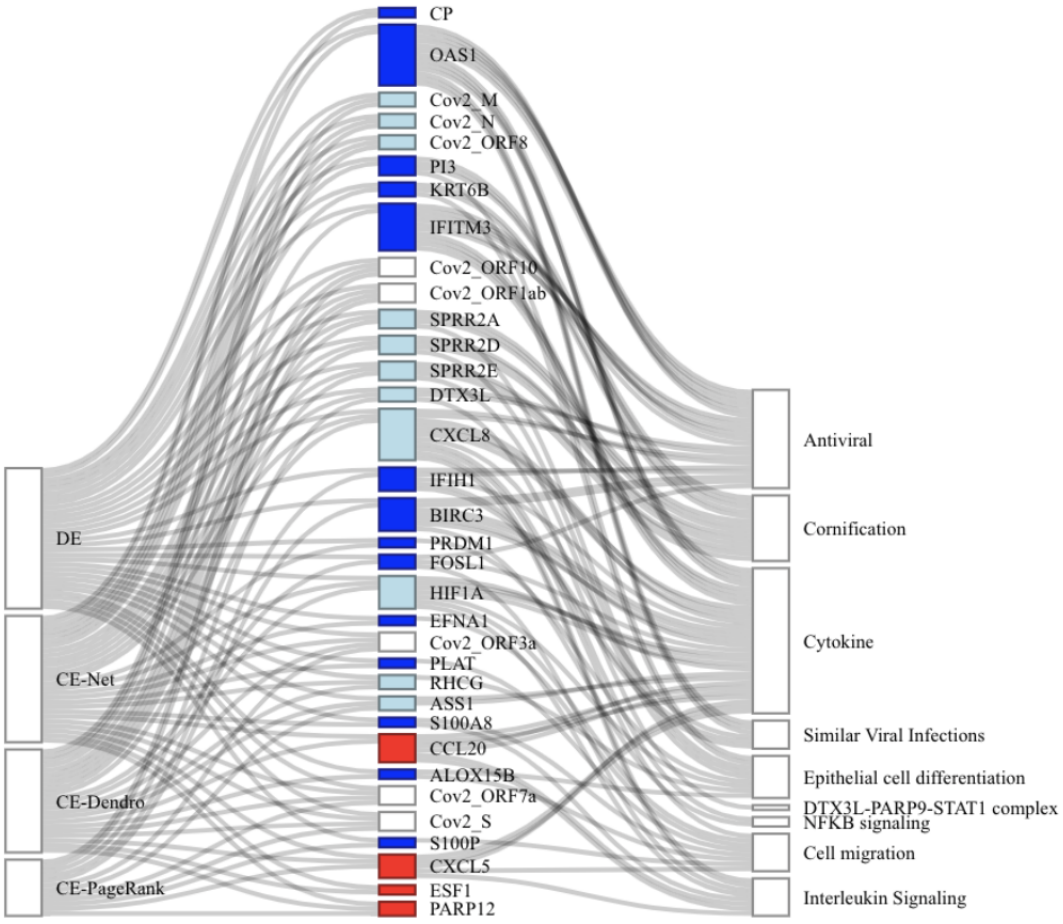


Figure 3.11: The sankey diagram shows genes that were found by at least 2 of the four views and the pathway themes that they participate in. This demonstrates that with a consensus of the multiple views the genes and pathway themes that are reported as concluding results and drivers of COVID-19 after extensive analysis in previous publications are immediately apparent with the HOUSE analysis. Genes are colored based on the views that they are found in, white, light blue, and dark blue indicate that DEA and 1, 2, or 3 other views, respectively, and red for a gene that is not found by DEA, but was by a CEA view. These genes are then connected to themes, which consists of a list of gProfiler pathways that they were found to be enriched in (Suppl. Table 1).

To reveal a core host regulatory response associated with SARS-COV-2 infection, we constructed a consensus from the results of DE, intersected with any of the three CE views and plotted the results as a Sankey diagram (Fig. 4B). Thematic biological processes were identified using the gProfiler tool31 on the resulting consensus to implicate pathways recorded in the gProfiler ontology database that are statistically enriched among the gene set (see "Methods" section). SARS-COV-2 genes themselves were not included in the consensus as gProfiler lacks annotations for this virus. The resulting consensus shares much in common with the CE-Net, which is expected given the common results revealed by the views as discussed above. Several genes and pathways were commonly implicated by DE and the CE views (e.g. 30 genes by DE and at least one CE; 16 genes by DE and two or more CEs). Five genes, all viral encoded—Cov2-ORF7a, Cov2-ORF1ab, Cov2-S, Cov2-ORF10, and Cov2-ORF3a, which were first quantified using dRAP—and 14 pathways indicative of a Immune and Defense response were found by all four views. Two genes—CXCL5, a chemokine, and PARP12, an interferon-stimulated gene involved in regulating inflammation—were found by the three CE views, but not by DE. The nine genes (Cov2-ORF8, Cov2-M, Cov2-N, SPRR2A, SPRR2D, SPRR2E, RHCG, HIF1A, CXCL8) found by DE, CE-Nets, and CE-Dendrogram, but not CE-PageRank could also be of interest for tight association with the virus albeit not central to the known pathway membership interconnections that influence the PageRank analysis.

Viewing the consensus as a network (Fig. 4C) highlights several sub-modules of densely interconnected genes and pathways (colored boxes, Fig. 4D–G). The first module (red box: CXCL5, CXCL8, CCL20, ASS1, HIF1A; Fig. 4D) consisted of genes belonging to several pathways expected to be implicated including Chemokine activity, IL-17 signaling pathway,

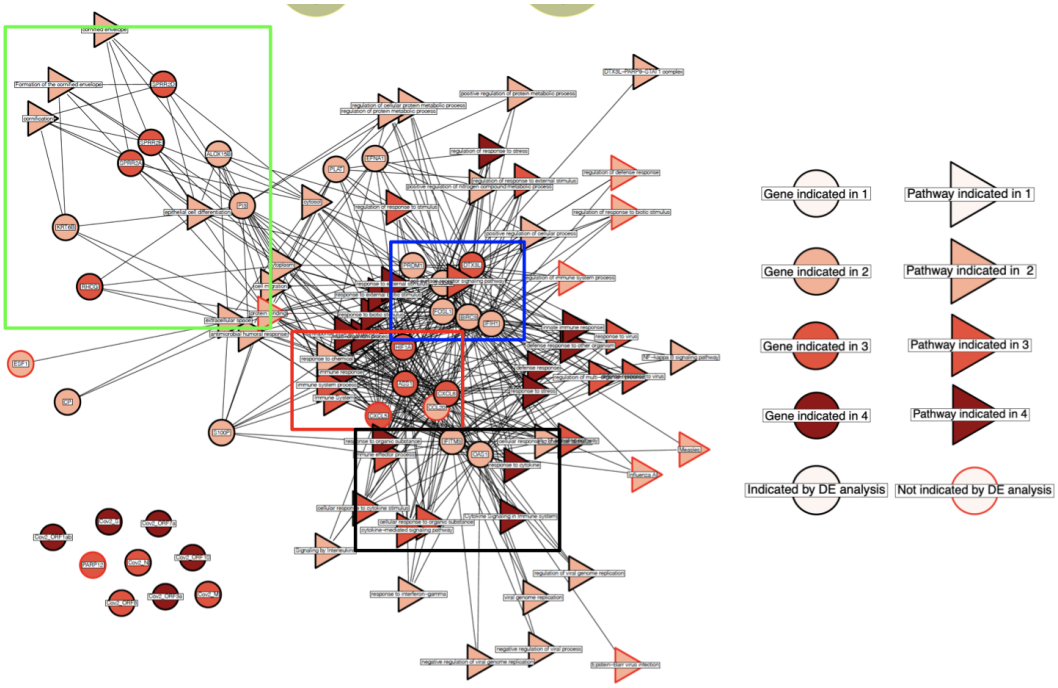


Figure 3.12: The consensus network for the cross-analysis overlaps between these four methods results in four gene modules.

viral protein interaction with cytokine and cytokine receptor, cellular response to lipopolysaccharide, and the 'ASATCAAAG' TCF-3 motif. The module is supported by previous findings that SARS-CoV-1 infection stimulates the lipopolysaccharide receptor, TLR4, shown to produce an immune response [38] and lead to disease pathogenesis [56]. The second module (blue box: DTX3L, IFIH1, BIRC3, S100A8, PRDM1, FOSL1; Fig. 4E) implicates roles for the innate immune system's defense response. DTX3L and PARP may be of particular interest based on previous findings that they are required for an interferon response to certain coronaviruses [194] and PARP12/14 are required to inhibit the replication of the macrodomain (a subunit of the transmembrane viral protein, nsp3) of coronaviruses and to produce the optimal IFN response [57]. The third module (lime box: SPRR2A, SPRR2D, SPRR2E, KRT6B, ALOX15B, PI3, RHCG; Fig. 4F) implicates the involvement of programmed cell death, epithelial cell differentiation and cornification and antiviral response through a DTX3L-PARP axis. This gene subgroup is characterized by Keratinization, Keratinocyte cell differentiation, cornification, formation of the cornified envelope, Epithelial cell differentiation, Epidermal differentiation, programmed cell death, and structural constituents of skin epidermis. Finally, a fourth module (black box: OAS1, IFITM3; Fig. 4G) indicates pathways of viral genome replication and cellular response to type 1 interferon. In summary, the consensus network reveals many expected aspects of response involving chemokines and inflammation but also some potentially new processes such as cornification that may implicate a specific apoptotic mechanism involving particular host cells (e.g. keratinocytes).

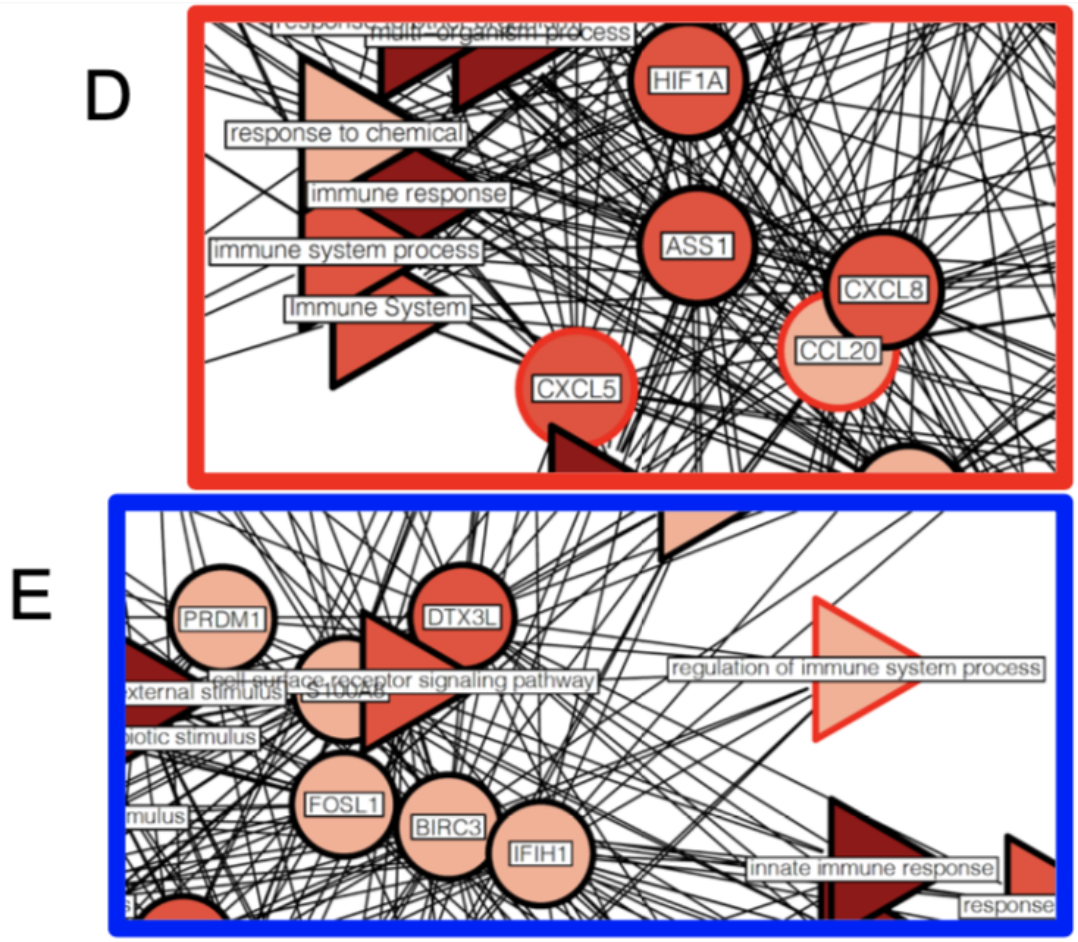


Figure 3.13: The red module (CXCL5, CXCL8, CCL20, ASS1, HIF1A) is indicative of chemokine activity, cytokines, and a lipopolysaccharide response. (E) The blue module (IFIH1, PRDM1, BIRC3, FOSL1, DTX3L, S100A8) indicates an innate immune response.

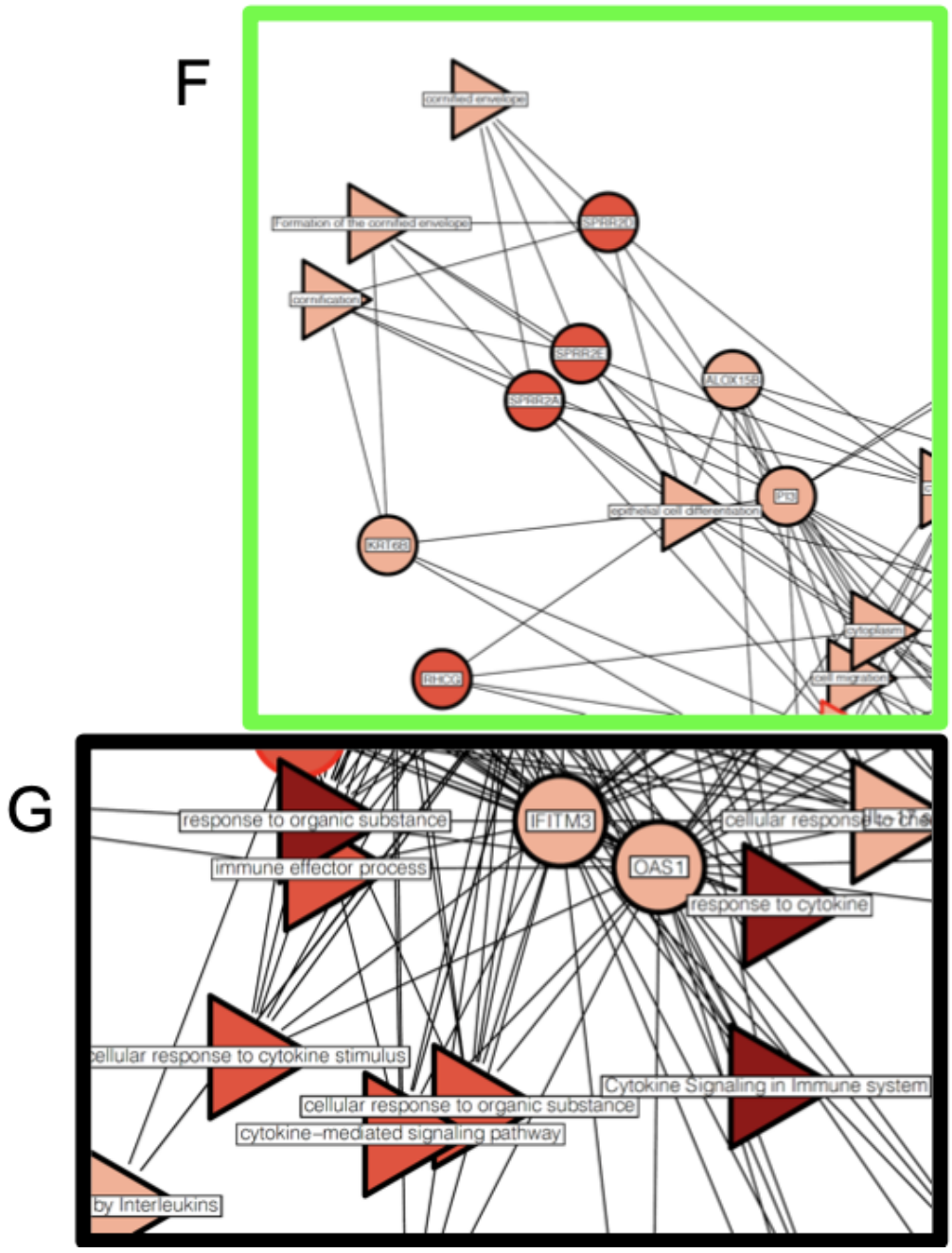


Figure 3.14: The lime module (SPRR2A, SPRR2D, SPRR2E, KRT6B, ALOX15B, PI3) indicates cornification/keratinization and epithelial cell differentiation changes. (G) The black module (OAS1, IFITM3) indicates regulation of viral genome replication.

3.3.5 Summary of dRAP findings

Using a dual RNA-seq analysis pipeline (dRAP) to detect both host and pathogen transcriptomic gene expression, we implicate genes involved in viral infection and response using both differential expression (DE) and coexpression (CE) analyses. To elucidate therapeutic implications, we applied dRAP to SARS-CoV-2-infected patient samples and cell lines. For the first time, to our knowledge, we quantify the levels of SARS-CoV-2 transcripts in human patients and cell lines. This new outlook revealed that the most strongly and consistently expressed transcripts were those that play essential roles in viral survival and propagation. The S, N, and M genes—which are key to viral replication (N gene), assembly (M gene), release (M gene), attachment (S gene), and entry (S gene) [5, 55]—had the highest levels of statistically significant expression.

dRAP suggested an appreciable difference in viral transcript expression between patient tissue types and found that BALF had the most robust levels. PBMCs exhibited low (or zero) levels of SARS-CoV-2 transcript expression, with no transcript detected as differentially expressed between control and infected conditions (DESeq2 using Cook's distance filtering; Table 1, Fig. 2A). Similarly, lung biopsies lacked robust expression of SARS-CoV-2 transcripts with all but one sample giving detectable levels (Table 1, Fig. 2A). In contrast, BALF samples were found to express every SARS-CoV-2 transcript at extremely high levels (at least 14 times higher in infected samples compared to controls; Table 1, Fig. 2A).

In addition to the robust viral response, the human transcriptional response of BALF also matched more closely with SARS-CoV-2-infected cell lines than the other tissues matched

to cell lines. dRAP identified that the NHBE cell line produced the highest magnitude of viral transcript expression (magnitudes higher than the A549 cell line, which also had some detectable viral transcript levels). The human transcriptional response, as measured by differential expression (DE) was found to be most similar between BALF and NHBE (Fig. 2C) with the other tissues having little in common with cell lines or BALF (Fig. 2E–F). Interestingly, the commonality between BALF and the NHBE cell line was even greater than when the cell lines were compared to each other, which helps increase the credibility of these findings given they are drawn from small sample sizes (Fig. 2B). These findings both underscore the relevance of using BALF as the tissue to quantify the gene expression programs of SARS-CoV-2 infection while suggesting some cell lines (e.g. NHBE) offer better laboratory models than others.

We investigated three different views to explore the roles of human genes and pathways relevant to viral infection implicated by coexpression (CE). DE measures the statistically significant conditional response to infection that may nominate genes with either direct or indirect viral associations. The integrated view of the consensus CE and DE together (Fig. 4A, Suppl. Fig.10) may provide a more enriched set of direct viral associations and host responses supported by literature observations of previous coronaviruses, viral response machinery, and symptoms observed in COVID-19 patients.

The modules of the consensus network suggest that SARS-CoV-2 may initiate a response via lipopolysaccharide through increases in chemokine and cytokine activity (red box, Fig. 4D), triggering an influx of intracellular calcium that induces the migration and programmed death of epithelial cells through a hardening of the membrane by a keratinization/cornification process (lime box, Fig. 4F) [38]. These findings are consistent with previous

observations that chemokines activate a response in multiple cell types and tissues (including epithelial and leukocyte) to produce high intracellular calcium and a migratory phenotype [2, 120, 119]. Imbalances and compensation among chemokines may predict response to infection [179]. Indeed, mutations in chemokine-associated genes have been associated with severe cases. Chemokine pathways may underlie symptoms that coincide with the various observations of COVID-19 symptoms including a respiratory mucosal immune response [179], inflammatory bowel disease (REF 44), interstitial lung disease [143], asthma [98], and Eosinophilic Pneumonia (EP) [77], which is associated with "progressive shortness of breath (dyspnea) of rapid onset and possibly acute respiratory failure, cough, fatigue, night sweats, fever, and unintended weight loss." Understanding the full scope of possible chemokine pathway engagement and how to inform therapeutic approaches is of great interest.

In summary, this work suggests that specific human patient tissues are implicated in robust SARS-CoV-2 expression, namely, Bronchoalveolar lavage fluid (BALF), while PBMC and lung biopsies showed little to no viral expression. NHBE and A549 infected cell lines had concordant gene expression changes compared to BALF, but PBMC and lung biopsy patient samples. Finally, the coexpression analysis enabled by dRAP predicts a possible mechanism by which COVID-19 may progress from the initial SARS-CoV-2 infection to patient symptoms that may provide additional clues into therapeutic targets.

3.4 Annotating and Profiling the SARS-CoV-2 Genome

In addition to the above, I entered into a collaborative effort and authorship of a reliable reference genome for SARS-CoV-2 during the pandemic [47, 114]. It is an interactive web-based tool, similar to the original UCSC Genome Browser, that collates the many international efforts to study the virus' evolution, mechanisms of action, and immunological characteristics. The many different research efforts were cataloged as annotated tracks that can be comparatively examined on a spatial map of the genome. These tracks highlight transcriptomic, genomic, and immunological features in a familiar and standardized data organization environment for thousands of researchers to make a one-stop shop for SARS-CoV-2 analysis.

Chapter 4

Chapter II: Associating transcription factors to single-cell trajectories with DREAMIT

4.1 Background

4.1.1 Gene regulation in single-cell trajectories

A cell's type and state are a product of gene regulatory mechanisms controlled by transcription factors. Transcription factors (TFs) play a central role in governing the transitions between different cellular states. These transitions are driven by the dynamic regulation of gene expression orchestrated by TFs. TFs bind to specific genomic regions, exerting precise control over the activation or repression of target genes. This regulation is critical for processes such as cellular differentiation, development, and responses to environmental cues. The identification of TFs responsible for orchestrating these transitions is a fundamental endeavor in understanding the molecular basis of cell state dynamics.

In recent years, advances in single-cell sequencing and transcriptomic methods have granted researchers the ability to scrutinize gene regulatory networks with great sensitivity and specificity. Moreover, the emergence of "cell trajectory" inference methods has enabled the identification of transitions between different cell states [141, 133]. Trajectory methods identify changes in development, maturation, or response to environmental queues using gene expression changes across cells having similar transcriptomes. The dynamic regulation of genes can then be inferred by following their relative expression across cells along a trajectory "branch" transitioning from one cell state to another (Figure 1A).

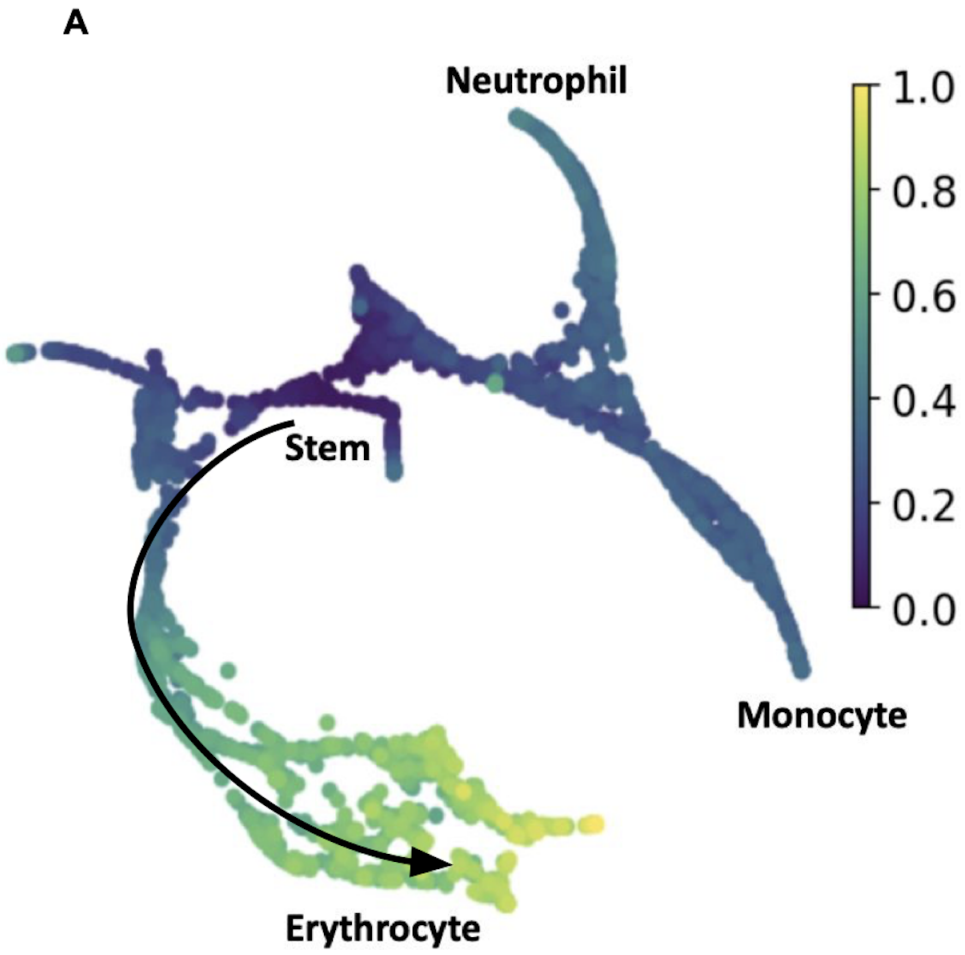


Figure 4.1: Expression across the cell transition in trajectory branches is used by DREAMIT to infer a dynamic view of TF-to-target gene regulation.

4.1.2 Motivation and existing methods

Unraveling dynamic transcription factor regulation of cell states from single-cell RNA sequencing (scRNAseq) data presents several challenges. scRNAseq data itself introduces analysis complications, mainly due to its vast scale and sparsity [7]. These include issues like gene expression dropouts, stochastic variations, delayed target responses, and disparities in chromatin accessibility. These factors can result in inconsistencies in known factor-to-target regulations [70, 29, 8].

To address these issues, the first approaches to infer Gene Regulatory Networks from scRNAseq data used "pseudobulked" transformations in which cell expression was grouped across related cells. Most recently, a growing number of methods are available to infer GRNs specifically from scRNAseq data (see [7, 105] for good reviews). To complement the information in single-cell RNAseq data, some approaches extend to include multi-omic data, for example the addition of ATACseq measured chromatin accessibility during the construction of GRNs [7]. Nevertheless, multi-omic data is both costly and not yet widely accessible. Thus, methods that extract TF-target relationships with strong support only from single-cell RNAseq data are still needed.

Importantly for this study, given a GRN, few methods exist to annotate which particular TFs are relevant for a particular context of the data. For example, even if a GRN is elucidated from a single cell dataset, it remains unclear which particular regulators to implicate as relevant for a cell state or cell transition uncovered in that dataset. As an analogy to bulk RNAseq analysis, GRNs can be inferred using methods like WGCNA [84] or ARACNe [102].

However, an additional step beyond GRN inference is needed to predict activities of genetic regulators using methods like MARINa [86], SPIA [159], and PARADIGM [171]. In the same way, approaches are needed to infer how genes within a GRN contribute to particular trends in a single-cell dataset. Current approaches use methods based on differential expression analysis, clustering, cell-type annotation, and dimensionality reduction [180, 127, 96]. Specific methods like TRADE-Seq [169] and PseudotimeDE [149] enable users to investigate differential gene expression as cells transition along a trajectory (see Figure 1B-C). Meanwhile, approaches like SINGE [39] employ Granger causality ensembles to infer potential regulatory interactions. Nevertheless, methods tailored for the specific inference of TF activity along trajectories remains poorly studied and the performance of methods remains a challenging task, often lacking objective criteria for evaluation [105].

4.1.3 DREAMIT algorithm overview

We introduce a novel method for implicating TFs to cell trajectories called DREAMIT – 'Dynamic Regulation of Expression Across Modules in Inferred Trajectories.' DREAMIT aims to analyze dynamic regulatory patterns along trajectory branches, implicating transcription factors (TFs) involved in cell state transitions within scRNAseq datasets (see Figure 1D-E). DREAMIT uses pseudotime ordering within a robust subrange of a trajectory branch (pseudotime focusing) to group individual cells into bins. It aggregates the cell-based expression data into a set of robust pseudobulk measurements containing gene expression averaged within bins of neighboring cells. It then smooths trends after searching for an optimal fitting spline across the bins. DREAMIT rejects further analyzing branches that produce highly variable smoothing

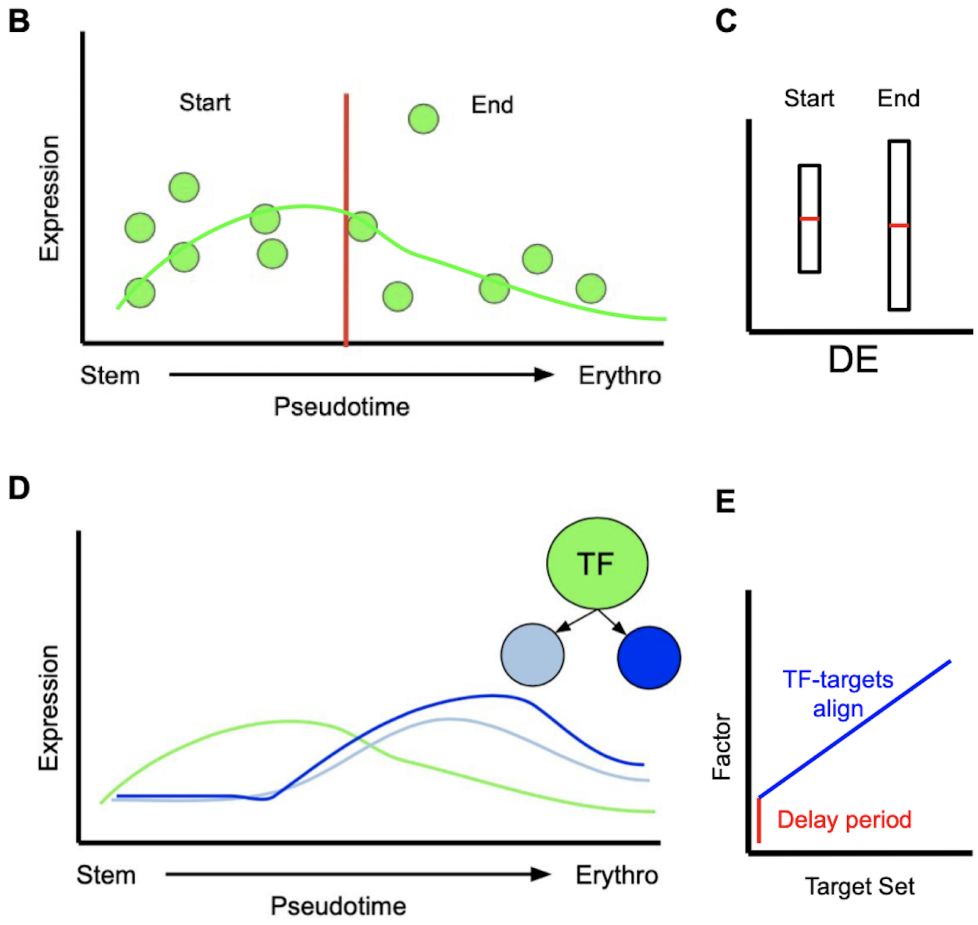


Figure 4.2: Associating transcription factors (TFs) to trajectory branches via identification of TF-to-target coexpression along pseudotime. B. Expression levels of a hypothetical gene in individual cells (y-axis) illustrating the division into arbitrary "start" and "end" states along the pseudotime of a theoretical differentiation process from stem cells to differentiated erythrocytes (x-axis). C. Differences in the expression level between "start" and "end" states may not exist which may cause Differential Expression ("DE") approaches to miss other patterns in the data (e.g. concordant fluctuations in the middle of pseudotime). D. DREAMIT models the entirety of the expression on the branch and assesses TF-to-target relationships that look for a consistent relationship between the expression levels (y-axis) of a TF (green line) and its target genes (blue lines) along pseudotime (x-axis). E. Alignment plot showing one TF (y-axis) aligned to a "typical" target from a target set (x-axis) illustrating how allowing for a lag or delay (red line) can help a metric pick up on an association between a TF and its targets over a subset of pseudotime (blue line) in which all the targets have the same lag in expression relative to the TF.

estimates (covariation in spline fitting parameters found to be greater than 1.0 across 80% subsampling; see Methods) as these branches may represent sparse or noisy parts of the data that could produce unreliable TF inferences.

Using the transformed smoothed data, it calculates the association between a TF and all of its predicted targets according to the TRRUST database assessed using multiple metrics (e.g., Pearson correlation, Mutual Information, Dynamic Time Warping distance, etc). Finally, DREAMIT uses a Relational Set Enrichment Analysis (RSEA) test to evaluate the significance of the TF-to-target associations and identify a core set of targets (target focusing) compared to a background model, which consists of arbitrarily selected targets.

4.1.4 Evaluating DREAMIT

We evaluated the performance of DREAMIT by measuring its ability to recover TFs with known relevance to several datasets. Our evaluation, although serving as a 'bronze standard,' is based on the lack of suitable reference datasets with known underlying regulations driving the major differences among cells, as previously noted [39]. To assess DREAMIT's performance, we employed a TF-Marker database, which allowed us to determine its effectiveness in identifying TFs known to play essential roles in specific tissues, previously established as high-confidence markers for those tissues [187]. While the evidence from gene expression alone should be viewed with caution as coexpression of transcription factors to targets are correlative and not causative, our findings revealed that DREAMIT outperformed traditional approaches, including differential expression analysis and GENIE3, in several instances.

4.2 The DREAMIT method

DREAMIT contains the following major steps: (1) pseudotime ordering to bin individual cells together, (2) averaging gene expression for each bin, (3) fitting a spline model to derive gene expression that smooths trends, (4) calculating the association between a TF and all of its predicted targets according to the TRRUST database using several different metrics, (5) identifying the significance of the set of targets using an empirically sampled null model and identifying a core set of the targets using an enrichment test and (6) retaining any TFs with a significant concordance after multiple hypothesis correction. These different steps are described in the following sections.

4.2.1 Datasets and Dependencies

DREAMIT relies on RNA expression data and the allocation of cells to specific branches, including their positions along these branches, which are represented as ”pseudotime.” These assignments are determined through the application of a cell trajectory inference method, as illustrated in Figure 1A. For this work, slingshot and PAGA were used to infer trajectories and pseudotime [152, 182]. PAGA was used for the Paul et al. dataset [128] since that study published a set of cell clusters that could be used as input to the PAGA method. Slingshot was used for the benchmark datasets (described next). The selection of Slingshot was based on the absence of pre-existing clustering assignments in these benchmark datasets and its well-documented performance in systematic evaluations [141]. The analysis is dependent on pseudotime assignments estimated by trajectory methods. Errors introduced by a trajec-

tory method can impact which gene regulatory relationships can be identified. DREAMIT is downstream of trajectory methods and is therefore subject to the same errors.

For benchmarking, we collected 7 datasets from EBI representing a diversity of tissues (including brain, heart, embryo, retina, bone marrow, testis) [101], a heart development dataset [52], and a dataset of stem to PBMC lineage [169] for testing DREAMIT. Each of these datasets have undergone preprocessing through the standard scanpy pipeline [181].

Traditional methods compare different states on the branch to each other in terms of differential expression, between a "start" and an "end" state (Figure 1B-C). DREAMIT takes a different approach in which the full set of expression values along pseudotime are used to infer a relationship between a Transcription Factor (TF) and its target. To do this, the method uses a set of predicted linkages between regulators and targets. For this study, we used both the human and mouse regulator-target predictions contained in the TRRUST database [63], which contains interactions mined from over 11,000 PubMed articles. To convert mouse regulogs to human, we used the orthology mapping published by the Mouse Genome Informatics (MGI) consortium [44]. These known regulatory datasets are provided with DREAMIT or the user can choose to include their own regulator-target interactions.

4.2.2 Trajectory Pre-Processing: Pseudotime Focusing, Expression Quantizing and Spline Smoothing

DREAMIT bases its analysis in first producing a smoothed representation of the data through psuedotime focusing and spline fitting. The use of splines to model scRNAseq data along a trajectory branch follows previous work [169, 149]. We selected spline models that

incorporated both goodness-of-fit and robustness.

The assignments of cells to locations along a trajectory, commonly referred to as pseudotimes, can reflect a fairly irregular distribution in which the density of cells can vary appreciably from one area to the next. This irregularity in cell number along pseudotime could result in only a few cells, or even a single cell, having a disproportionately large influence in correlation or distance calculations made by DREAMIT (described below). For this reason, DREAMIT applies a pseudotime focusing step in which it retains the cells that have been assigned contiguous pseudotime values falling within 1.5 times the interquartile range of all pseudotime values of the branch (Figure 4A-B). The outlier cells that could impact detecting robust trends across pseudotime are eliminated from all subsequent steps.

DREAMIT discretizes the assigned pseudotime values into discrete ordinal levels in order to buffer against further irregularities in the data. Discretization bins are chosen to have an equal number of pseudotime increments, where each bin also has a minimum of 10 cells. Expression levels are then averaged across all cells in the same bin (Figure 4C). Thus, the result of the quantization step produces a type of pseudobulk dataset in which gene expression levels are averaged across cells landing in the same ordinal bin. DREAMIT attempts different resolutions of discretization by varying the number of bins from 4 to 100 and chooses a satisfactory number using a hyperparameter search step (described below). It also provides the option of using the raw expression data for the next step of spline smoothing rather than the binned pseudobulk data.

The expression levels of the retained cells could have substantial noise due to technical and biological factors; for example, noise due to the well-documented zero-inflated bias

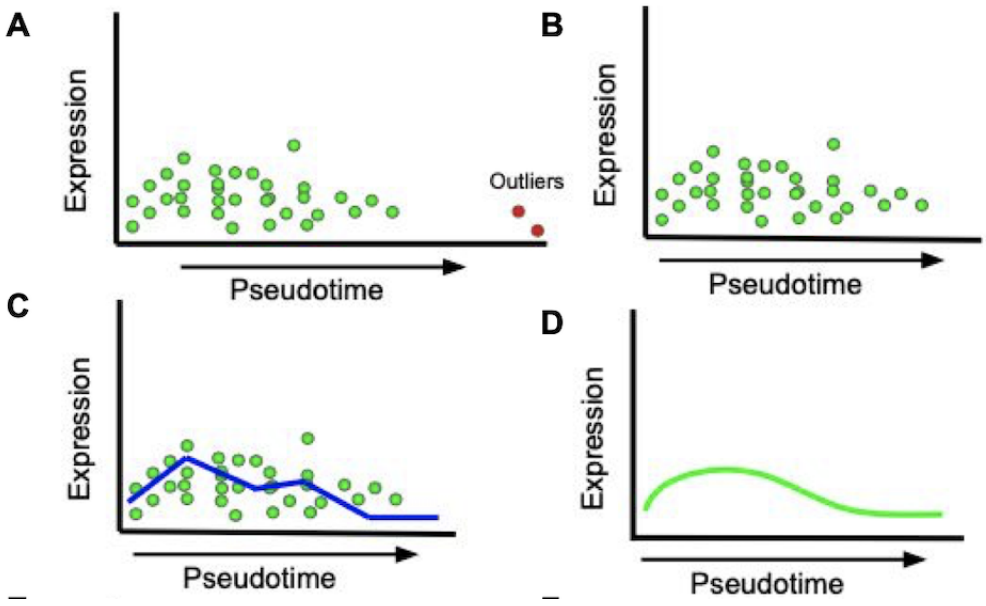


Figure 4.3: A. Cells with outlier pseudotime assignments (red dots) compared to the other cells (green dots) are shown. B. Pseudotime focusing removes outliers from the analysis and retains cells within 1.5 times the interquartile range of all pseudotime values of the branch (see Methods). C. Cells are grouped into bins containing at least 10 cells per bin. The average expression of a gene is calculated from all the cells in a bin (blue line) and this bin-averaged expression is used for all subsequent analysis. D. Spline smoothing incorporates information from cells in neighboring bins to further smooth out the expression changes in pseudotime (green curve).

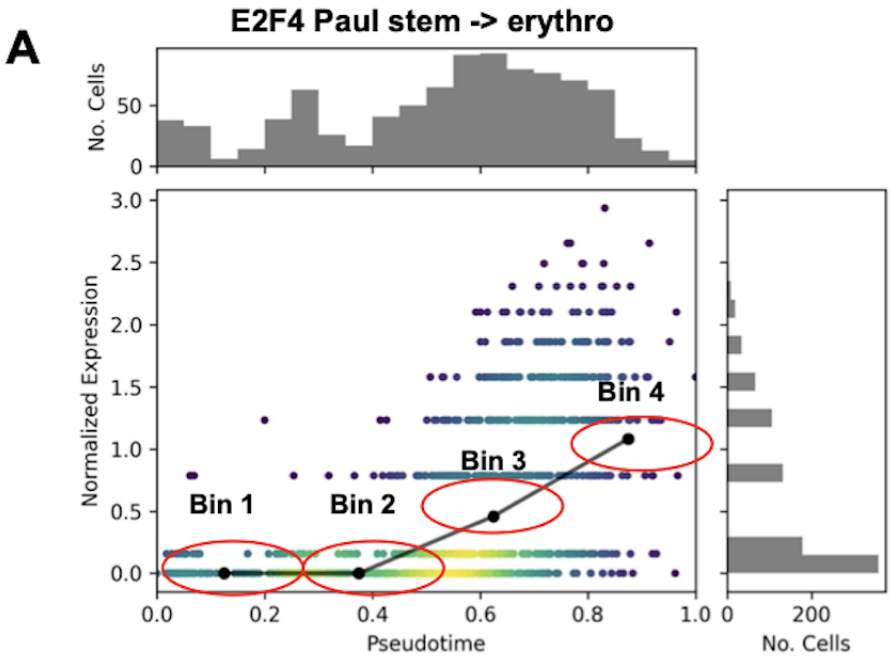


Figure 4.4: A. Illustration of E2F4 expression (y-axis) across pseudotime (x-axis) along the stem to erythrocyte trajectory branch from the PBMC dataset with normalized (Scanpy) expression (color indicates number of cells); spline-smoothed expression shown for each pseudtime bin (black line); red ellipses illustrate bin width. Histograms plot the distribution of cells at given expression increments (y-axis) and psuedotime increments (x-axis).

in single-cell RNAseq data [35] (see Figure 5A-B for two examples of raw expression levels). In addition to only averaging the expression levels within a bin, DREAMIT uses a spline smoothing step to incorporate information from cells in neighboring bins (Figure 4D). The discretized pseudotime units are treated as the independent variable and the average expression levels are fit with a zero-inflated negative binomial generalized additive model spline smoothing (NBGAMSS) of the expression values following the recent approaches of TRADE-Seq and PseudotimeDE [169, 149]. The number of cells in a bin is used as a "weight" for the fitted data point, placing more emphasis on areas with more support that are assumed to have more reliable estimates of average expression compared to areas with a fewer number of cells. A smoothing factor determines the number of resulting knots produced by the NBGAMSS (see Figure 5A-B for two examples of smoothed expression). Note that other smoothing operations are possible such as the kernel smoothing used by SINGE [40].

The hyperparameters for a spline are selected by searching for a combination that maximizes the goodness and robustness of the fit, measured by Akaike Information Criterion (AIC) and Coefficient of Variation (CV), respectively. Subsamples are used to compute the Coefficient of Variation (CV) to reflect robustness. We subsample a trajectory branch by choosing 80% of the cells, without replacement, repeated 30 times. The average Akaike Information Criteria (AIC) is calculated across these subsamples to reflect the accuracy of the fit to the quantized data. Likewise, the CV is calculated across these subsamples. We search for a set of tolerable spline parameters – the number of bins and the smoothing factor – that minimizes both the CV and the AIC. A perfect dataset would have an AIC and CV of 0. We used the distance to the origin of a spline's associated AIC-CV pair as a measure of goodness that combines both

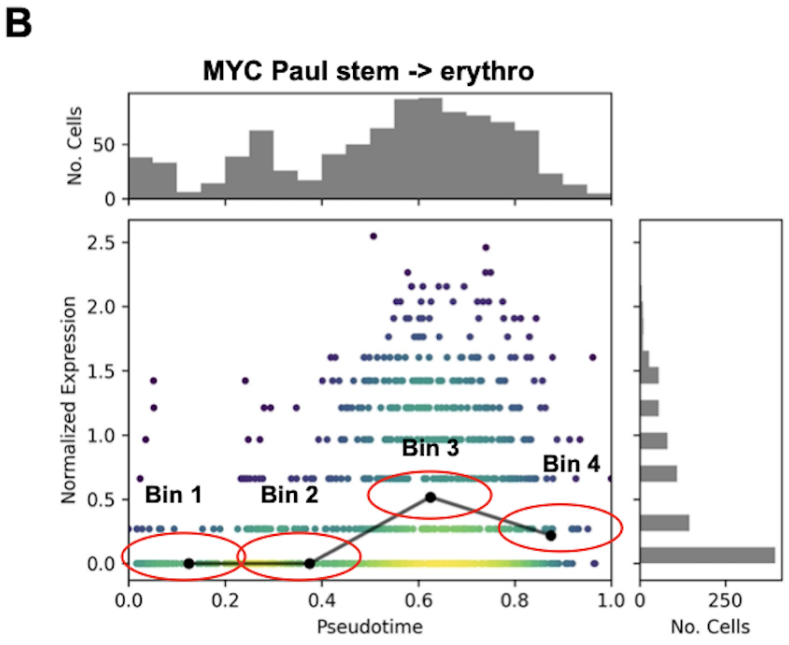


Figure 4.5: Same plot as in Figure 4.4 but for the MYC transcription factor.

the fit quality and robustness. To illustrate the spline selection process qualitatively, we plotted splines for ANKRD43 and FRY1 that had a good fit (low AIC and CV; suppl. Fig 11-12) as well as a poor fit (high AIC and CV; Suppl. Fig 13-14). The differences in a poor modeling are evident at high bin numbers, but it should be noted that most spline fits are not extremely off base in terms of modeling, but the best parameters enhance the smoothness. All of the tested spline representations of the PMBC dataset [169] were plotted to assess their AIC (goodness of fit) and CV (robustness of fit)(Figure 5C), with little to no trend being observed with the changing bin and smoothing factor parameters.

As RNAseq becomes more cost effective, datasets will increase in resolution, decreasing the distance between cell state transitions, and the discretization step may not be necessary

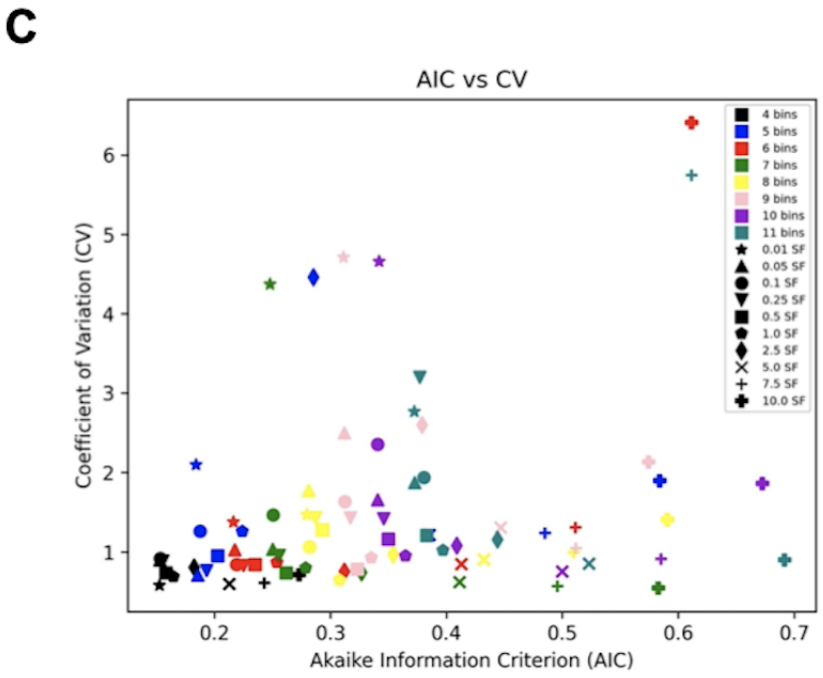


Figure 4.6: Each spline parameter choice produces a different fit to the data. Robustness of the fit was assessed by measuring the coefficient of variation (CV) across subsamples of the data (y-axis, see Methods). Goodness of the fit was assessed using Akaike information criterion (AIC) (x-axis). Parameter values that produce spline fits plotted toward the origin (bottom left) are preferred to those further away.

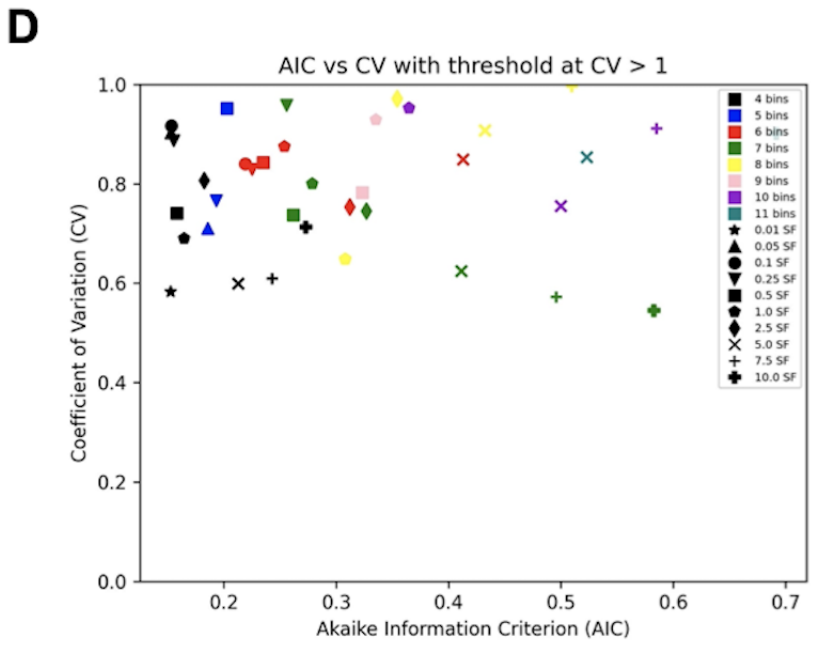


Figure 4.7: Same as in Figure 4.6 but only showing parameterizations that achieve tolerable levels of robustness (CV < 1).

for some datasets. Thus, DREAMIT provides the option of using the raw expression data for spline smoothing without first performing the quantization binning step. We tested the influence of binning on the Paul et al. dataset and found that it produced TF predictions with more tissue specificity compared to predictions obtained without the binning step (see Suppl. Fig 15a-b).

Expression data for each of the 100 most variable genes is used for the hyperparameter search. Once an optimal value for these parameters are found, the preprocessing is applied to all genes. In order to ensure that DREAMIT only ever reports on robust representations of the data the maximum threshold for Coefficient of Variation is set to 1 (Figure 5D) in this and all other data analyses. Only trajectory branches that meet this CV criterion will be assessed by DREAMIT.

4.2.3 Transcription factor to targets

To quantify relationships between a transcription factor (TF) and its targets, DREAMIT performs a series of pairwise tests (e.g. Pearson correlations or dynamic time warping alignments, see next section) between a particular TF and each of its predicted target genes. The targets of a particular TF are taken from the regulatory interactions recorded in the TRRUST database [63]. The number of targets for each TF in DREAMIT is a result of at least three factors: 1) which targets are linked to the TF in the TRRUST database, 2) are retained after scanpy filtering of the data to select for highly variable genes, and 3) are included after DREAMIT RSEA.

4.2.4 DREAMIT metrics to detect a variety of TF-target relationships

The connection between the expression levels of a transcription factor (TF) and those of its target genes can vary from being straightforward and linear to more nuanced. This variation depends on the particular regulon (i.e. the strength of association between a TF and a large or small subset of its target genes). For that reason, DREAMIT includes several metrics to pick up on TF-target associations. For each TF with an associated target set, DREAMIT calculates Pearson correlation, Spearman correlation, Dynamic Time Warping Cost, and Mutual Information content [10, 111, 109, 81]. We describe the details of the calculation of each of these methods in the following paragraphs.

Pearson correlation provides evidence regarding the strength of the linear relationship between the factor and its targets, while Spearman correlation reflects the strength of the

monotonic relationship. Because the factor and targets can have positive or negative correlation, the metric must be squared to create a distribution that is comparable to a random background distribution. Therefore, it is primarily the strength of the relationship that is considered in DREAMIT, but all relationships are included in the report for the user to investigate.

Dynamic Time Warping (DTW) measures how well the expression pattern of a factor can be aligned to the pattern of a target. Target genes may follow a similar pattern as a regulator but have delayed timing that may be detectable in a single cell dataset with enough resolution to reveal cellular processes. DTW finds an optimal match between the patterns that introduces the fewest delays in time [25]. Therefore, a factor with a small DTW distance from its targets implies that the targets are tightly controlled by the transcription factor with few delays in pseudotime.

Mutual Information (MI) measures the potentially non-linear dependency between the expression of a factor and that of a target gene. It measures how much a factor's expression distribution reduces the uncertainty about the target's expression distribution. Including MI among the metrics expands the relationships that can be detected between TFs and targets but can produce results in which the nature of the regulatory interaction is not clear (e.g. in the cases where correlation-based metrics fail to pick up a relation). [166].

In addition to the standard use of these metrics, a rolling metric calculation is also performed. In the rolling metric calculation, the possibility of a delay in targets responding to factor expression changes is considered by sliding the window in which the relationship is measured. In other words, a rolling window is applied to the target's expression where each window interval is a bin of smoothed expression (Figure 4E-F). Therefore, the rolling metric

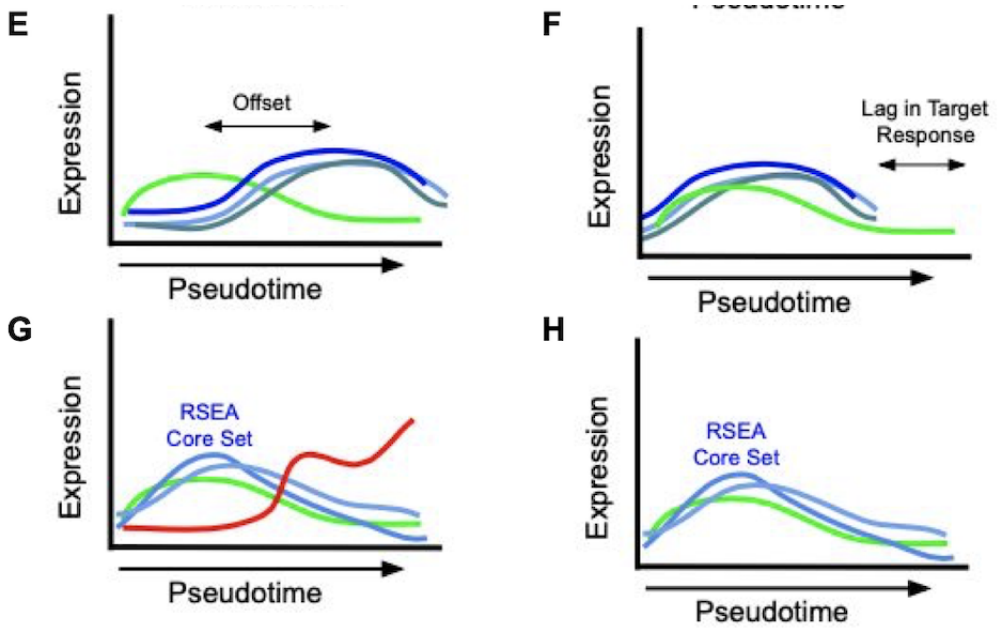


Figure 4.8: E. DREAMIT quantifies TF-target relationships through pairwise tests of the spline smoothed expression of the TF (green line) and its target genes (blue lines). F. Illustration of a "rolling" metric incorporating pseudotime lag. A significant lagged correlation will be detected when several targets share the same delay. G. Target focusing employs Relational Set Enrichment Analysis (RSEA, see Methods) to identify a "core" set of targets with high association to the target (blue lines) while excluding the targets with weak or poor association (red curve). H. 75% of the targets with the highest concordance to the factor are retained.

reports the bin increment delay where the peak strength of the delayed relationship occurs. DTW allows for delays in factor target relationships, but is flexible as to the synchronization. In contrast, the rolling metrics report significant factor target relationships with the same time delay, so that all targets respond to the factor expression at a similar pseudotime.

4.2.5 Relational Set Enrichment Analysis (RSEA): Detecting significant TF-target associations

In addition to a standard reporting of these prior known TF-target associations, DREAMIT uses a two-step Relational Set Enrichment Analysis (RSEA) to highlight the core set of relationships (Figure 4G-J). RSEA begins by performing target focusing to capture the enriched target response followed by a Kolmogorov–Smirnov test (KS test) to assess significance. A transcription factor may activate a subset of its targets in a particular tissue. DREAMIT attempts to identify the set of utilized targets by identifying a focus set. Target focusing uses the provided set of targets for a factor (e.g. in this study, [63]). DREAMIT performs two steps to identify targets with consistent expression with the factor. First, it checks if the direction of regulation is consistent with the activation/inhibition annotation available from the TRRUST database. If the database has no annotation about the directionality of regulation (46% of relations in the TRRUST database), then this consistency check is skipped. Second, DREAMIT performs a focusing step to retain a set of targets that mutually have the strongest relations to the TF; i.e. keeps only those targets with the highest scores to the TF using the current choice of metric. A range of percentile thresholds were tested from 10-35 with similar results for 20 and above, with 20 and 25 having the best early precision (Suppl Fig 15c-d). The target focusing threshold is a hyperparameter of DREAMIT; 25 was used in this analysis. Targets with an association to the TF above the 25th percentile of a DREAMIT metric are retained in a core "target focus" set. Conversely, targets with deviant expression are filtered out in an effort to reduce false positives.

To determine the significance of a DREAMIT relationship metric for a TF, the dis-

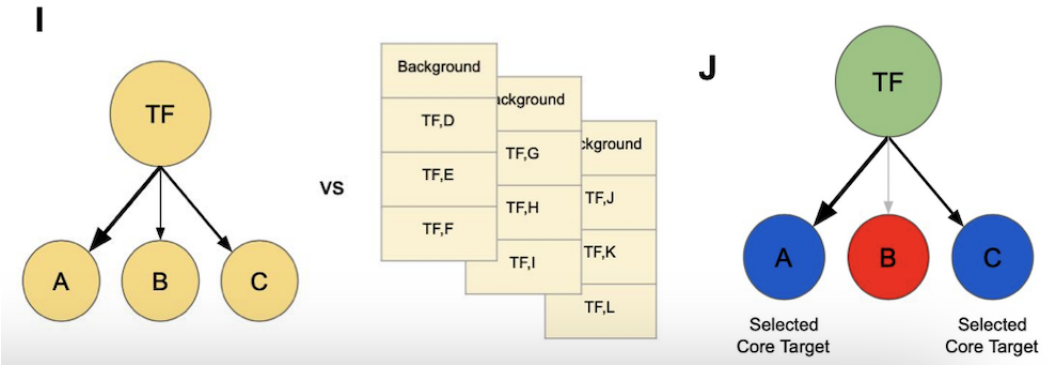


Figure 4.9: I. Significance is assessed by comparing TF-to-target metric scores to a random background in which random targets are chosen to be of the same size as the TFs original regulon (yellow lists) with a Kolmogorov-Smirnov test. J. The core set of targets (blue nodes) found by RSEA are used in the statistical analysis.

tribution of metric levels between the TF and its targets is compared to a random background distribution. The background set is chosen to be equal to the largest target set considered. After the random targets are selected, the same metrics are calculated for the random set and then compared to the true factor-to-target distribution using the Kolmogorov–Smirnov test (KS test)[11]. A Benjamini-Hochberg correction [160] is then performed on the p-values generated for each factor by the KS test to account for false discoveries. Transcription factors with adjusted p-values < 0.05 are considered significantly non-random for the association in question. Both RSEA results and the full set of TF-targets are evaluated using this KS test and p-value correction method.

4.2.6 Differential Expression (DE)

DREAMIT uses a differential expression (DE) t-test between the start and end of a cell trajectory using the original data values (“Raw DE”) as well as using the spline-inferred

data values ("Smooth DE"). The spline-inferred, Smooth DE approach uses a generalized additive model (GAM) as the basis for the DE test, similar to previous methods such as Monocle-2, TRADE-Seq and PseudotimeDE [169, 149, 134]. For the DE test (either raw or smooth), DREAMIT divides a trajectory branch into two equal segments, the "start" and the "end". For raw expression, outlier cells are pruned such that those with aberrant pseudotime values are ignored. If a gene shows a statistically significant change between these two conditions then it is considered differentially expressed. Statistics such as fold change, t-statistic, pvalue, and FDR are then reported for each gene on a branch for both the raw and smoothed data.

4.2.7 Evaluating the specificity of TF to target relationships

To determine the specificity of DREAMIT, ROC, Precision-Recall, and early precision were assessed. True "hits" are defined as transcription factors that are markers of the tissue under study. All other transcription factor findings are considered false hits. This assumption does not provide perfect accuracy to true biology as there are likely many factors that have yet to be classified in the TF-Marker database being used [187], and also every factor marker of a tissue need not be active at all times. However, this bronze standard metric does give an enhanced insight into DREAMIT's ability to highlight tissue specificity. ROC, Precision-recall, and early precision were assessed branch by branch and as an aggregate. These specificity curves were determined for findings by Pearson correlation, Spearman correlation, dynamic time warping, mutual information, rolling calculations, and by the most significant method for each factor. By examining the TPR, FPR, precision, and recall under various p-value thresholds, the specificity of these methods can be assessed comparatively. For early precision, we followed previous

works [40] that defined this metric as the precision at recall levels < 0.1.

In a similar manner, competing approaches, differential expression (DE) and GENIE3 [75], were comparatively assessed against these DREAMIT methods. For DE, if the transcription factor (TF) is a listed marker of the tissue under study then it is considered a true hit, while all other TFs are considered false hits. This is the same method used for DREAMIT. Additionally, the DE targets were assessed by comparing the T-statistics of the targets with a background set using the KS test. For the TF in question, it is considered significant if its targets score better than the random background. Both the scanpy processed expression (rawDE) and the spline smoothed expression from DREAMIT (smoothDE) were assessed for both TF DE and DE targets. For GENIE3, a distribution of weights is determined for TFs to known targets from Trrust [63]. This is then statistically compared to a distribution of weights for TFs to randomly selected genes (representing the background) with a KS test. This is done for both scanpy processed expression (rawGENIE3) and the spline smoothed expression from DREAMIT (smoothGENIE3). Specificity is then assessed in the same way as DREAMIT where tissue markers are true hits and all other TFs are false hits.

We note that rawDE has some processing done (IQR pruning of cells with outlier pseudotime assignments). This processing was done, because often dividing a branch strictly on the original pseudotime assignments results in severe imbalances in the number of cells in the start and end segments. Therefore, some processing was still useful in this analysis above the traditional approach, even before results are considered.

4.2.8 Evaluating the overlaps between DREAMIT and other methods

In order to compare the findings of DREAMIT, DE, and GENIE3 methods, a set of TF-branch pairs were recorded for each method's predictions. For example, the pair (MEK,Branch-3) would record that MEK was associated with the third trajectory branch of the dataset. The TF-branch pair sets of each method were compared and visualized using an Upset plot. This was done for both the tissue-specific markers of the respective branches being assessed and for the non-markers for DREAMIT, rawDE, smoothDE, rawDEtargets, smoothDEtargets, rawGENIE, and smoothGENIE. Likewise, the DREAMIT subcomponent methods Pearson, Spearman, DTW, MI, and Rolling were also assessed via Upset plot for both tissue-specific TF marker and non-marker findings. We note that different upstream methods for manifold learning and trajectory inference could be used in conjunction with DREAMIT (such as scGNN [173], DESC [87], and scMGCA [192]).

4.2.9 Evaluating specificity in reporting TF-markers in a high-fidelity PBMC dataset

To assess the specificity and biology of DREAMIT in a "silver standard", a PBMC dataset was used [128]. Because this dataset was derived from mice, a set of regulogs [44] was used in this analysis so that downstream specificity and tissue markers can be assessed. Due to a higher proportion of TFs being categorized as blood-specific markers or not, the specificity for DREAMIT can more accurately be determined. ROC, precision-recall, and early precision are determined in the same way as above, using the TF-marker database [187]. Individual

factors were then investigated further as to their status in the marker database and in the overall literature.

4.3 Results

4.3.1 DREAMIT Identifies Distinct PBMC markers

To evaluate DREAMIT's specificity in a highly curated setting in which confident tissue-specific transcription factor (TF) regulation is well known, we chose the blood marrow dataset from Paul et. al. [128]. This dataset contains stem cells transitioning to various blood cell types (erythrocytes, monocytes, neutrophil) suitable for trajectory branch inference and contains a well-characterized set of marker genes. We estimated the accuracy of the methods using an average of 16.3 markers per branch by calculating precision-recall, and early precision (see Methods). DREAMIT was found to have the highest average precision and early precision (AUC=0.57, E= 1.00) while the other approaches had lower estimates – smoothGENIE (AUC=0.46, E=0.38), rawGENIE (AUC=0.43, E=0.27), rawDE (AUC=0.33, E=0.29), smoothDE (AUC=0.33, E=0.17), rawDEtargets (AUC=0.49, E=0.40), and smoothDEtargets (AUC=0.32, E=0.17) (Figure 2A). These estimates demonstrate that DREAMIT on average achieves levels of precision moderately higher than chance expectation as 35% of the transcription factors were tissue-related for this analysis. For example, the method achieves both a precision and recall of 0.55 which is significantly different than chance guessing at the 0.05 level ($P<0.026$, Hypergeometric test). In addition to tissue specificity, we also compared DREAMIT to Perturb-Seq results in hematopoiesis from Lara-Astiaso et. al. [85]. The TFs annotated to

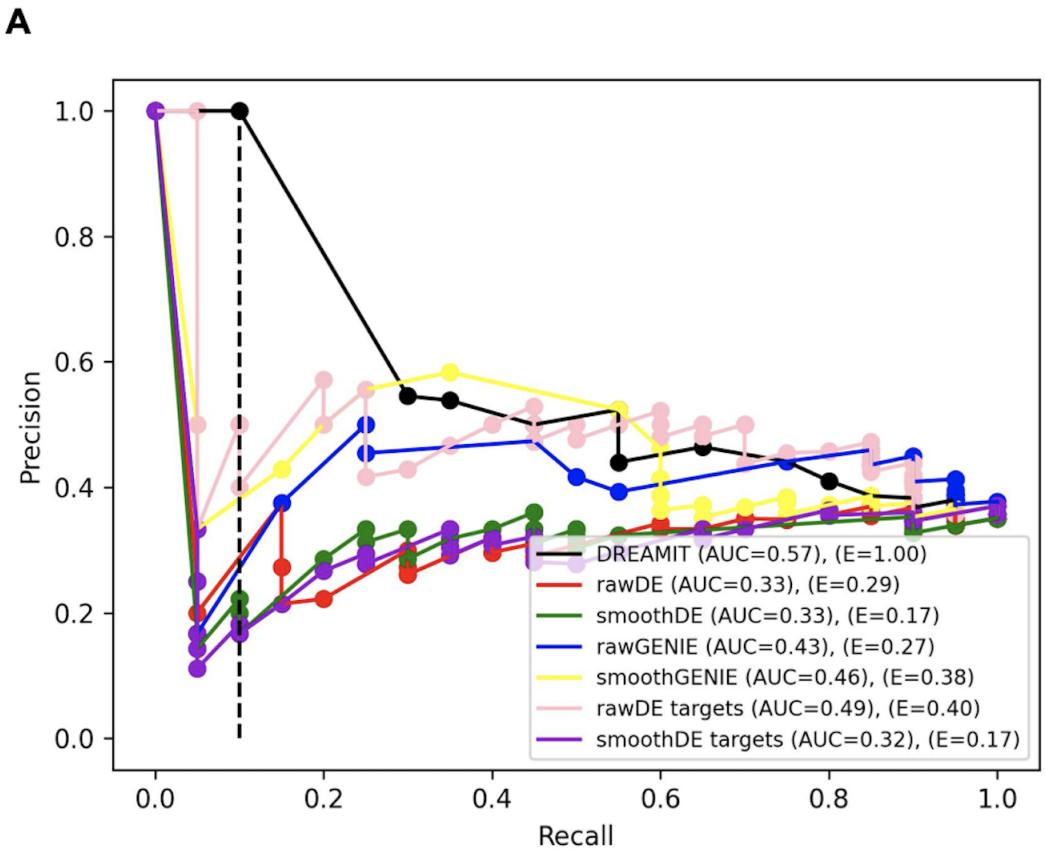


Figure 4.10: Performance of inferring blood differentiation TFs on the PBMC dataset. A. Precision-recall curve measuring rate of identifying blood-related transcription factors annotated in TF-Marker DB from the analysis of the PBMC dataset. A dotted line is plotted on the precision-recall to denote where early-precision is considered.

either the monocyte or erythrocyte branches had a high degree of overlap with Perturb-Seq data that report on factors that modulate these lineages (6 out of 8 TFs) and the downstream differential expression observed in marker genes had a higher concordance with those TFs selected by DREAMIT compared to the two that were not selected (Suppl. Fig 16).

4.3.2 DREAMIT inference of gene regulatory logic for PBMC fate specification

The majority of transcription factors (TFs) identified by DREAMIT were annotated as markers of stem cells or blood cells, which aligns with the input data from the Paul et al. dataset. This dataset represents the transition from stem cells to differentiated blood cell types. For instance, on the trajectory from stem cells to erythrocytes, DREAMIT found 16 significant TFs. Among these, five were well-known blood and peripheral blood markers based on the TF-marker database (RUNX1, GATA1, EGR1, STAT1, ETS1). Furthermore, three other TFs were stem cell markers (YBX1, MYC, RELA). Out of the remaining eight TFs identified by DREAMIT, six had established literature support connecting them to roles in stem cell and peripheral blood mononucleocyte (PBMC) development, such as DNMT1 [3], EZH2 [76, 69], E2F4 [73, 72], KLF6 [24, 46], NFE2L2 [115], and TP53 [80, 125]. The last two TFs, MYB and MYCN, had a less clear relationship.

On the trajectory from stem cells to monocytes, DREAMIT identified a total of 13 significant TFs. Six of them (CEBPA, ETS1, IRF1, ATF4, RUNX1, STAT3) were blood and peripheral blood markers, while three were recognized as stem cell markers (IRF8, KLF4, NFKB1). Among the remaining four TFs on this trajectory (ZBTB16, MYCN, ELF1, VDR), only the vitamin D receptor (VDR) had established literature supporting its involvement in monocytes [155, 22].

To further investigate how DREAMIT findings provide insight, a TF-to-TF network was created alongside their temporal order of activation. For the significant DREAMIT findings from the high fidelity Paul et. al. PBMC dataset [128], TF-to-TF relationships were plot-

ted for the stem-to-monocyte trajectory (Figure 2B). TFs are connected in the network if they were recorded as linked in the TRRUST database and significant by DREAMIT. Analyzing the (smoothed) expression patterns of TFs in the network revealed a correlation with the temporal changes in expression and the connectivity in the TF-to-TF network, as illustrated in Figure 2C. For example, the factors STAT3 and RUNX1 are upregulated at the initial time point and their targeted TFs are upregulated in a later time point, consistent with the network inferred using the TRRUST TF-to-TF relationships [53, 19, 124]. In another example, a concurrent increase in ZBTB16 and CEBPA (regulated by ZBTB16) was observed followed by increases in their downstream target TFs – STAT3 (regulated by CEBPA) and VDR (regulated by STAT3). On the stem-to-erythrocyte trajectory branch, DREAMIT identified an analogous TF-to-TF network (see Suppl Fig 17). RUNX1, EGR1, MYCN, and ETS1 increased in the early stages of the trajectory branch with widespread increased expression across the TF network in later stages. GATA1 and DNMT1, both regulated by a single distinct TF, show increases in expression following increases in their single regulator (MYC -¿ GATA1, DNMT1 -¿ TP53).

4.3.3 DREAMIT Identifies tissue-relevant TFs at a higher rate than standard approaches.

Systematic assessment of a method's accuracy in identifying TFs for a specific trajectory branch requires access to a diverse range of tissues with well-documented TF roles. Unfortunately, for many tissues, comprehensive single-cell analyses have not yet been conducted to establish a reliable set of TFs. Nonetheless, some relevant information has been gathered and stored in repositories like TF-Marker DB [187]. To address this challenge, we gathered

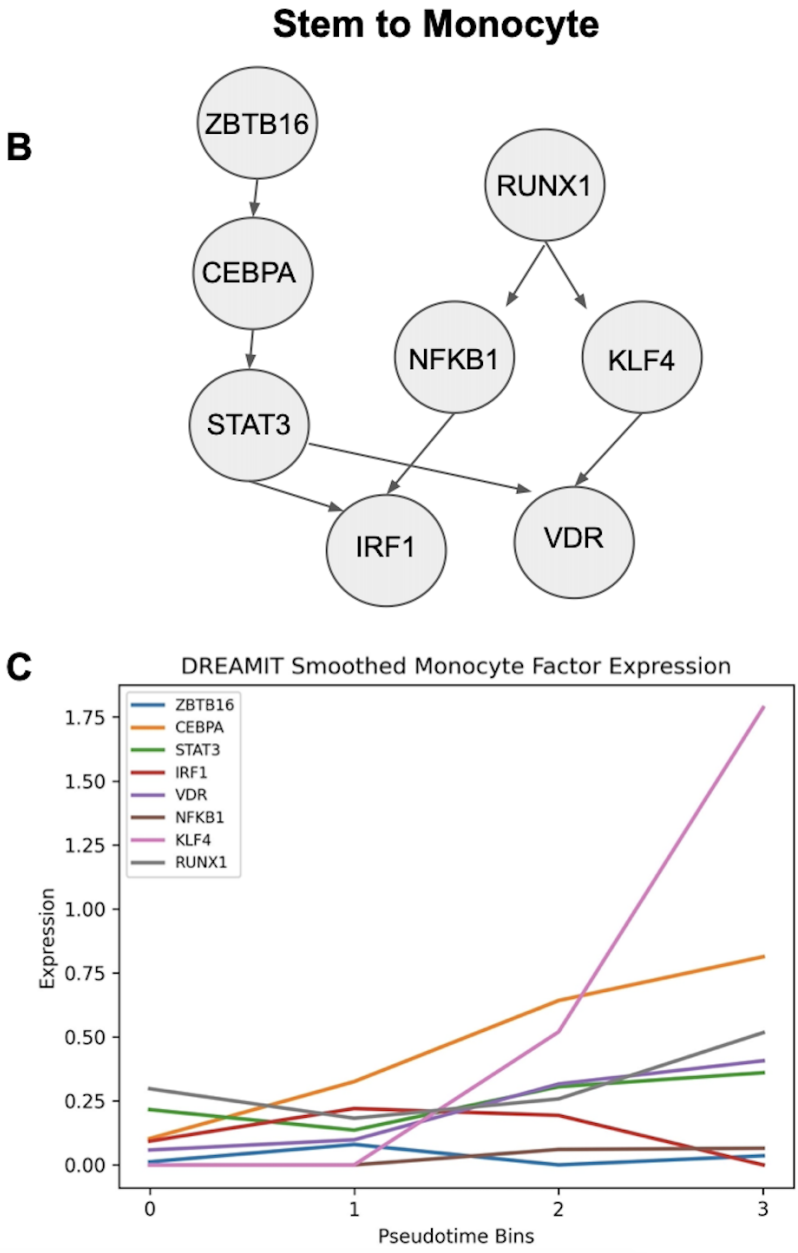


Figure 4.11: B. TF-TF relationships found by DREAMIT depicted for the stem to monocyte branch from this PBMC dataset. C. The expression (y-axis) of the TFs from part B plotted across pseudotime (x-axis).

seven datasets in which at least one TF was found in both TRRUST and TF-Marker DB, and was annotated as a regulator in tissues examined in the experiments. In total, this ”bronze standard” benchmark encompassed 84 TFs identified as pertinent, resulting in 207 instances of TF associations with various datasets. This benchmark covered 15 trajectory branches across six different tissues, including the brain, heart, embryo, retina, bone marrow, and testis.

To assess the methods' capacity to identify tissue-specific TFs in the benchmark datasets, we employed a precision-recall analysis. This approach is suitable for situations where we anticipate a far larger number of negatives than positives, primarily because we treat all unknown markers in the benchmark as negatives. To facilitate a robust comparison, we evaluated the overall performance of DREAMIT, Differential Expression (DE), and GENIE3 across all trajectory branches and tissues. To gauge the performance of these methods, we used the area under the curve (AUC) to evaluate their general performance across all recall levels. Additionally, we used ”early precision” (E) [39] to assess performance under a strict confidence threshold, where only the top-ranked relationships are taken into account for calculation.

DREAMIT had the highest average precision (AUC=0.20) surpassing its competitors, rawDE (AUC=0.13), smoothDE (AUC=0.13), rawDEtargets (AUC=0.16), smoothDEtargets (AUC=0.16), rawGENIE (AUC=0.16), and smoothGENIE (AUC=0.18) (Figure 3A). Additionally, DREAMIT demonstrates a much stronger early precision (E=0.42) compared to rawDE (E=0.09), smoothDE (E=0.08), rawDEtargets (E=0.19), smoothDEtargets (E=0.13), rawGENIE (E=0.33), and smoothGENIE (E=0.21).

The individual metrics of DREAMIT (excluded from Figure 3A for clarity) maintained their respective specificity rankings when assessed by precision-recall with the exception

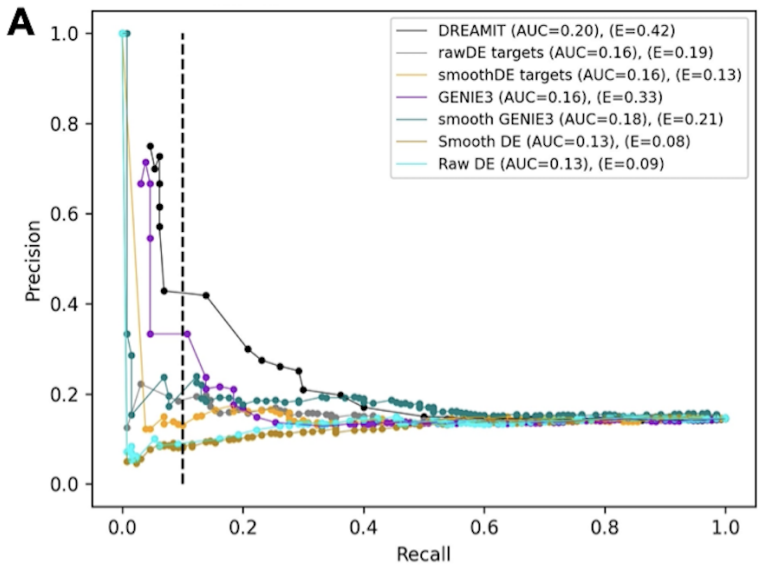


Figure 4.12: Precision (y-axis) versus recall (x-axis) measuring each method's ability to detect TFs from one of the 15 benchmark trajectory branches in which true positive TFs were assumed to be those annotated by TF-Marker DB as previously associated with the tissue assayed by the experiment.

of MI (AUC=0.17) and Rolling (AUC=0.18), which both fall behind Pearson (AUC=0.22), Spearman (AUC=0.21), and DTW (AUC=0.19). On the other hand, the early precision of these component methods is ranked quite differently with Rolling (E=0.43) scoring the best (even slightly superior to the DREAMIT ensemble, E=0.42) followed by MI (E=0.38), DTW (E=0.35), Spearman (E=0.28), and Pearson (E=0.28), respectively. Both the component methods and the ENSEMBLE outperformed DE and GENIE3 in average and early precision.

To further compare DREAMIT, DE, and GENIE3, we created an upset plot to view the distinct and common TF-to-branch association pairs found across all of the 15 branches in the benchmark (Figure 3B). The tissue-specific TF-to-branch predictions found at FDR¡0.05 in each method are shown. DREAMIT finds the most number of TF-to-branch associations (109;

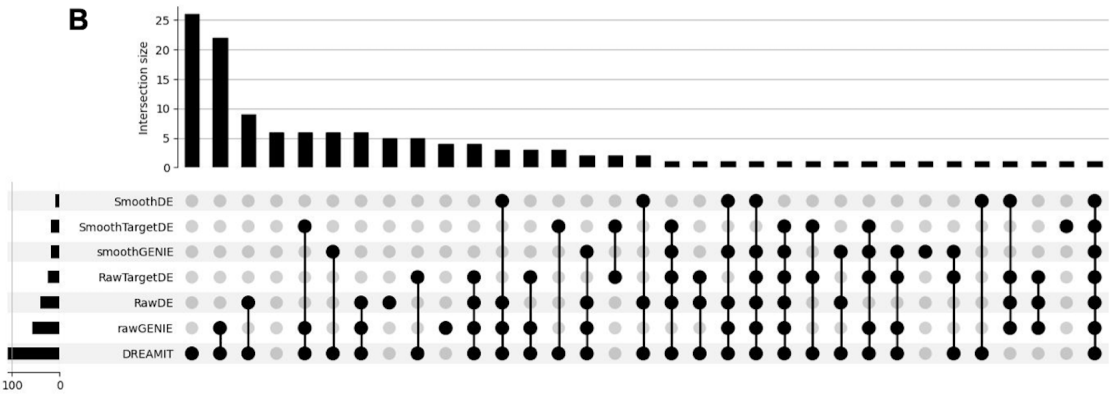


Figure 4.13: Upset plot illustrating the set intersections between all methods of the TFs inferred for each branch; i.e. set members compared are TF-branch pairs. DREAMIT had the highest number of TF-branch pairs (109) with 26 uniquely predicted.

7.3 TFs per branch) followed by rawGENIE (58; 3.9 TFs per branch), rawDE (40; 2.7 TFs per branch), rawDEtargets (25; 1.7 TFs per branch), smoothGENIE (18; 1.2 TFs per branch), smoothDEtargets (17; 1.1 TFs per branch) and smoothDE (10; 0.7 TFs per branch). DREAMIT produced the largest number of TF predictions per trajectory branch making it the most sensitive method. In addition, DREAMIT shares the most overlap with rawGENIE, 52 associations (89.6%), and with rawDE, 33 associations (82.5%). We also investigated the overlaps of TF-to-branch associations found by the components of DREAMIT (data not shown). DTW found the most associations (73) followed by MI (71), Rolling (66), Pearson (64), and Spearman (39). There was a high degree of overlap amongst all of the methods. Spearman had the least overlapping associations, but 100% of its findings were also reported in one of the other 4 methods. Mutual information and DTW had the most exclusive TF-to-branch associations. Overall 16.5% of the associations found by DREAMIT were found by all component methods, and 66.2% were found by at least two.

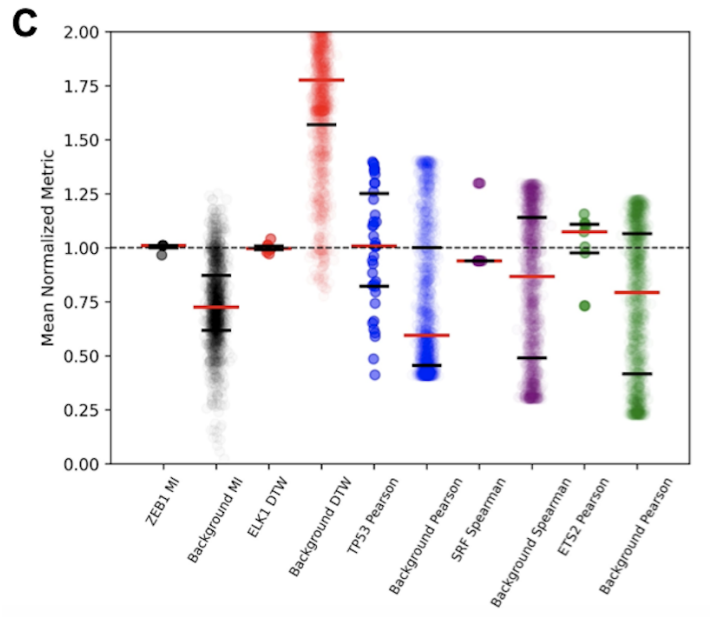


Figure 4.14: DREAMIT compares the TF-target relationship distribution to a background for each constituent metric and reports significant TFs through a Kolmogorov-Smirnov test. Examples of five different TFs with different metrics plotted along the x-axis (colors indicate TFs); each TF plotted as a pair with left side showing the observed metric for the targets of the factor and the right side showing the distribution of randomly selected targets.

Taken together, the vast majority of tissue-specific TF-to-branch associations found by the two competing methods were also found by DREAMIT, and in addition DREAMIT found more than 50 associations missed by these methods. The total number of associations that could have been reported in this analysis was 130. This means that DREAMIT found 83.8% of tissue specific associations, while rawGENIE and rawDE found 44.6% and 30.8%, respectively. DREAMIT had the highest degree of specificity and the highest percentage of tissue-specific markers.

DREAMIT found several cases, missed by other methods, in which the TF-to-target distribution was distinct from the background and found to be significant with a KS test (Figure

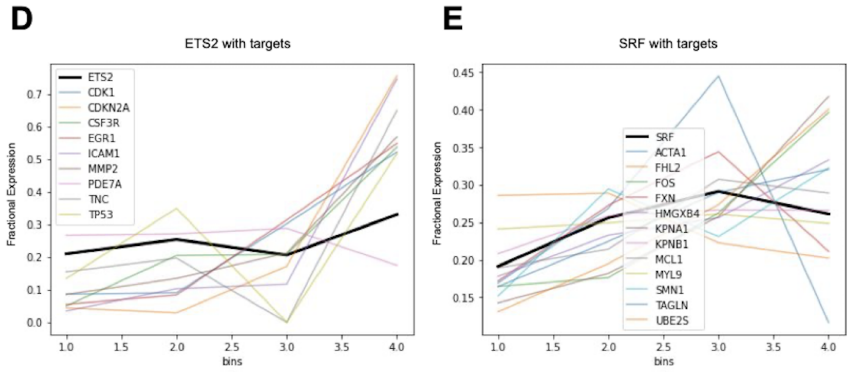


Figure 4.15: D. Illustration of ETS2 factor with its targets found to be significant by Pearson. To visualize genes of different expression scales in one plot, each gene's expression from one bin was divided by that gene's expression summed across all bins (fractional expression; y-axis). E. Illustration of SRF found to be significantly correlated with its targets using Spearman correlation. Fractional expression (y-axis) was used for visualization purposes.

3C). First, the marker ETS2 (Figure 3D) was found to be significant according to the DREAMIT Pearson method (Rsq=0.77, FDR<0.001), as well as the marker SRF (Figure 3E) on the same cardiac trajectory branch by the DREAMIT Spearman method (Ssq=0.70, FDR<0.005) [51]. DREAMIT captured these two annotated, biological markers in the trajectory data, while all other methods overlooked this as significant with the exception of rawDEtargets for ETS2 (pval<0.005). Second, the marker TP53 (data not shown) was observed to be significant by DREAMIT Pearson (Rsq=0.77, FDR<1e-7) [100], despite the large number of targets (n=78) introducing potential noise into the calculation of the Pearson correlation. This demonstrates that DREAMIT is able to find strong TF-to-target relationships in both small and large target sets. This finding was missed by rawDE, rawDEtargets, smoothDE, and smoothGENIE, but was reported to be significant by rawGENIE and smoothDEtargets.

The DREAMIT DTW method also detected a significant association for ELK1 (D=0.24, FDR<0.01) (Figure 3F), an association that was missed by all other methods [51]. Finally, the

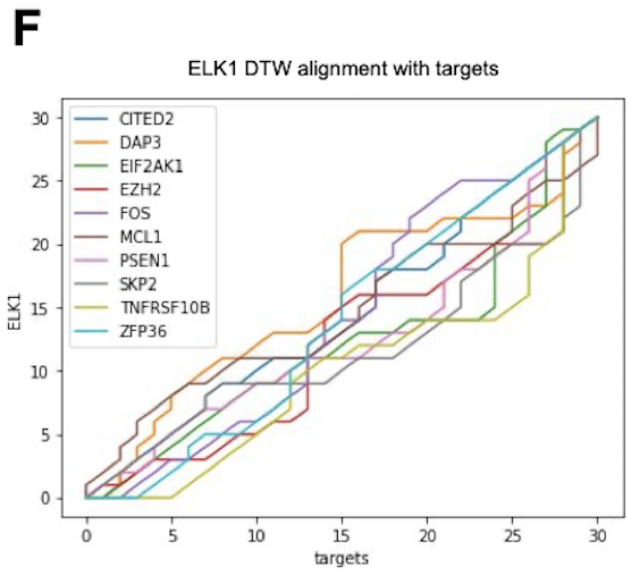


Figure 4.16: Dynamic time warping picked up on a significant relationship between ELK1 (y-axis) and its targets (x-axis); the alignment graph illustrates that the targets all maintain a relationship with the factor but there is variability from one target to the next in terms of the exact nature of the relationship.

DREAMIT MI method found significance in the ZEB1 marker (MI=2.35, FDR$<$0.005), which was missed by rawDE, smoothDE, rawDEtargets, smoothDEtargets, and rawGENIE, but was found by smoothGENIE, suggesting that the smoothing spline implemented provides benefit to other methods of analysis.

To demonstrate the significance and minimal variability of the above results, their distributions were plotted as a swarm plot. The markers ZEB1, TP53, SRF, and ETS2 were all significantly above the background indicating a strong relationship found through either MI content, Pearson correlation, or Spearman correlation. The marker ELK1 was significantly below the background DTW distance, which indicates a stronger DTW alignment for ELK1. Altogether, these findings demonstrate that DREAMIT associates TFs to trajectories consistent

with their known tissue specificity and that these findings are missed by DE and GENIE methods in many cases.

4.4 Discussion

In this work, we presented DREAMIT, Dynamic Regulation of Expression Across Modules in Inferred Trajectories, a novel framework for testing and identifying dynamic gene regulation in TF to target relationships. The method can be used downstream of any method that projects single cell data into a lower dimensional manifolds and derives a cell trajectory solution (e.g. Monocle instead of Slingshot used in this study).The method was developed to aid researchers in identifying the gene-gene regulatory relationships governing the state transitions cells undergo detectable in associations between their transcriptomes. Previous methods either ignore gene regulation, such as TRADE-seq or PseudotimeDE [169, 149], or were not designed specifically for single-cell trajectories, such as GENIE3 [75]. As such, DREAMIT provides a complementary perspective on interpreting gene regulatory mechanisms from scRNAseq data.

TF-target sets, whether taken from databases or user-supplied, will likely contain targets that are irrelevant for the analysis of a particular tissue and may not reliably detect an association of a TF to a trajectory branch. For this reason, we introduced the use of several association metrics (e.g. Pearson, Mutual Information, Dynamic Time Warp, etc) as well as the use of a Relational Set Enrichment Analysis (RSEA) to detect the significance of a target set's association relative to a random background that can tolerate such noise in the target set. Furthermore, we found that using a target focusing step, much like an enrichment analysis can

identify a "leading edge" of contributing pathway genes to a differential expression signature, helps boost the RSEA detection signal.

In a benchmark set of data encompassing six tissues, we found that DREAMIT exceeded the performance of Differential Expression (DE) and GENIE3. DREAMIT had both a higher sensitivity overall and captured more tissue-specific markers.In summary, DREAMIT was shown to have higher sensitivity, finding over 80% of the tested markers, with higher ROC, precision-recall, and early precision compared to DE and GENIE3 based methods. In the PBMC dataset, DREAMIT also had the highest specificity in terms of ROC (AUC=0.66), precision-recall (AUC=0.57), and early precision (E=1.00). On the erythrocyte branch, DREAMIT found 16 significant TFs, 5 of which were known blood markers and 3 were known stem cell markers. On the monocyte branch, there were 13 significant TFs, 6 were known blood markers and 3 were stem cell markers. Of the significant TFs found by DREAMIT that were not established markers in the TF-marker database [19], 9 of them had established or emerging roles in stem and PBMC development in the literature (see Suppl. Table 1). For example, the finding association of VDR with monocytes was not documented in the TF marker database but has been demonstrated in recent literature [155, 22]. Thus, among DREAMIT's predictions are potential new examples of tissue-specific regulation by novel factors.

Additionally, DREAMIT was able to find 26 TF-to-branch associations missed by any other competing method (Figure 3B). For example, significant marker findings of ETS2, SRF, TP53, ELK1, and ZEB1 were captured by DREAMIT, but missed by the others. DREAMIT presumably can pick up on these overlooked cases presumably because it considers multiple relationship modalities between TFs and their targets (e.g. Pearson, DTW, MI, etc).

DREAMIT uses only scRNAseq data to implicate TFs to particular cell transitions. As such, the association of TFs to branches based on the gene coexpression of targets should be viewed as suggestive. In addition, since the metrics are based on coexpression relations and since the directionality of pseudotime may not reflect "real time" in a cell, the TF could promote or repress activity along a branch. As new multi-modal datasets are increasingly published (e.g. ATACseq and RNAseq on the same cells), an obvious question becomes whether incorporating additional datasets would improve the performance of a TF association method like DREAMIT. One such approach called Dictys was published recently [174], which uses a probabilistic model to incorporate ATACseq and RNAseq for TF activity inference. To quantify DREAMIT's limitation in using only the RNAseq data, we compared its performance to Dictys on the human hematopoietic dataset analyzed by the Dictys authors with their published trajectory solution (using the STREAM method [29]). Surprisingly, we found the performance of DREAMIT to be comparable and often superior to Dictys on the B-cell, the Erythroid, and the Monocyte branches (Suppl Fig. 18). Notably, DREAMIT's precision was higher than Dictys over most recall levels considering all 10 different temporal GRNs produced by Dictys for each of the three trajectory branches. For this comparison, DREAMIT was run using the author-provided smoothed data as well as the raw expression data. We found that the results from smoothed data had the best early precision. Thus, even using only the RNAseq data and ignoring the ATACseq data that Dictys incorporated, DREAMIT identified as many tissue-specific TFs (and sometimes slightly more at the early precision levels) compared to Dictys and that many of the TFs were those reported by the authors. The recent PerturbSeq study in mouse by Lara-Astiaso et al. [85] in which TFs were systematically knocked out and then specific marker gene ex-

pression assayed, also showed a high concordance with DREAMIT findings from the Paul et al. dataset [128]. We found that, out of the 8 murine TFs reported for the HSC-to-monocyte or HSC-to-erythrocyte lineages, six of the human orthologs had been associated by DREAMIT to either one of these branches. Furthermore, the two that had not been associated with either branch showed much lower levels of marker gene differential expression in the monocyte and erythrocyte branches when the TFs were knocked-out (see Supplemental Fig. 16). These further demonstrates the advantage of DREAMIT to compensate for the lack of multi-omic data through the use of TRRUST relationships and RSEA.

One limitation of DREAMIT is that it considers one TF at a time even though it is known that TFs work together in combination. Even so, several significant associations are found by considering individual TFs in isolation. Extensions of the work are possible that could test combinations of TFs. For example, starting with pairwise TF combination detection, one could test the association between two TFs as well as all pairwise associations between the distinct members of their target sets using the same metrics and statistical tests defined in this work. The drawback of the approach would be the difficulty currently in evaluating its success as there is limited availability of datasets in which TF combinations have been annotated as relevant.

While DREAMIT exceeded competitors in our evaluations, all of the methods had low precision. This is likely due to the limited availability of relevant TFs associated to a particular tissue and to specific branches produced by trajectory inference. For example, using TFs from the TF Marker database allowed us to consider multiple datasets for the evaluation, but it assumed TFs annotated to a tissue were relevant for any/all branches in a dataset that assayed

a specific tissue. It is certainly possible that other TFs or other biological differences underlie the variation in the observed transcriptomes of individual cells. There is clearly a need for well annotated datasets that can be used as benchmarks for gene regulatory network inference in single cell analyses [40]. As more datasets with multiple data modalities become available (e.g. ATACseq and RNAseq), it will be possible to develop sets of TFs relevant for branches in a more unbiased and systematic fashion.

DREAMIT code is available on Github https://github.com/nathanmaulding/DREAMIT.git. The grid search that DREAMIT uses to find good parameters for spline fitting contributes the biggest impact to running time and grows linearly with the number of cells along a trajectory. For example, the smallest dataset in this study took 14.8 seconds and the longest took 189,704 seconds. Parallelization of the grid search to utilize nodes on a large compute cluster could greatly improve the running time as each parameter combination is independent.

4.5 Conclusions

In conclusion, we developed and evaluated DREAMIT, a novel framework for investigating dynamic gene regulation in TF-to-target relationships gleaned from cell trajectories inferred in single-cell RNAseq data. DREAMIT was found to outperform baseline approaches in 15 different trajectory branches in a benchmark dataset and the well-characterized PBMC dataset. DREAMIT detected the association of TFs to tissue-specific trajectories in several instances where the association was missed by all other methods, demonstrating its variety of metrics may help it detect some of the dynamic interdependencies preserved in the pseudotime

inference provided by the cell trajectory analysis. It would be interesting to apply DREAMIT to single-cell time-series datasets like Klein et al [79] as they become more widely available. In conclusion, DREAMIT offers a complementary approach for shedding light on TF-to-TF networks that govern the temporal regulation assayed by emerging single cell datasets.

Chapter 5

Chapter III: Transformer-based modeling of Clonal Selection and Expression Dynamics (TraCSED)

5.1 Background

While cancer is traditionally considered a disease of genetic alterations, studies have shown adaptive resistance to treatment often involves molecular differences in genetically identical cells [146, 135, 145, 140, 148, 60, 147, 42, 157]. The plastic nature of the response to treatments and its heterogeneity may even be a necessity to enable the emergence or selection of resistant clones [37]. Characterizing and understanding tumor heterogeneity and how it affects treatment response is critical to the development of new therapeutic strategies in oncology. Advances in single-cell technologies, such as the tracking of molecular states over time, can be used to address those questions [42, 13, 15, 176, 61, 121, 50, 168, 163, 139, 130]. In particular, transcriptional heterogeneity can be captured through clonal barcoding methods such as TraCe-Seq [28]. Studies have also shown that "twin" clones (i.e. sister cells with the same barcodes) show a higher degree of transcriptional similarity than other clones [54], suggesting that the transcriptional state of barcoded clones persists after multiple divisions. Leveraging this assumption, TraCe-Seq measures the fitness and transcriptional trajectory of clones in parallel pools of cells undergoing different treatments. To uncover differences associated with the selection process, the fitness of clones at the end of treatment can be mapped back to their transcriptional states pre-treatment or at intermediate time points.

While experimental advances have been substantial, methodologies to interpret such multidimensional single-cell transcriptional data sets are still in their infancy. For example, clustering approaches can fail to distinguish resistant from sensitive cell states as recent clonal barcoding systems have shown that a few genes, or even a single gene, can determine a cell's

fate [138]. However, incorporating clonal information through a tunable parameter into the dimensionality reduction has been shown to benefit the identification of features associated with cell outcomes [138]. Innate resistance may play a major role as intrinsic cell states conferring resistance could be encoded prior to external stimuli [54]. Additional forms of non-genetic response may also contribute to resistance. For example, altering epigenetics or rewiring signaling networks could underlie adaptive resistance that emerges after treatment has begun [27]. Because traditional methods rarely integrate longitudinal data as a continuum and associate them with an outcome, they yield limited understanding of adaptive resistance, its underlying gene programs, and their time at which such changes occur.

5.1.1 Understanding innate and adaptive resistance with TraCSED

To address those shortcomings, we first developed a semi-supervised approach with partial least squares regression (PLSR) to identify pre-treatment markers of innate resistance in TraCe-seq datasets (Figure 1A-D). This approach incorporates both the transcriptional similarities of cells together with a phenotypic response to reveal signatures that might otherwise be hidden from traditional methods. PLSR uncovers single gene markers of innate resistance in breast cancer, such as SNHG25 and CLDN1, which were missed by traditional clustering and differential expression methods.

Next, to identify gene programs associated with adaptive resistance, we introduce TraCSED (Transformer-based modeling of Clonal Selection and Expression Dynamics) that learns a dynamical process of clonal selection using a clone trajectory's inferred pseudotime from TraCe-Seq single-cell RNA-seq data as the time series input for model fitting (Figure 1E-

G). TraCSED uncovers interpretable gene programs that are directly related to the selection process and estimates time periods at which these may be critical for resistance.

We applied PLSR and TraCSED to infer resistance mechanisms for two treatments used in the clinic to treat breast cancer, palbociclib and giredestrant, which target CDK4/6 and estrogen receptor (ER), respectively. We compared the results of single-agent treatments to combination therapy in which we observed different factors associated with resistance. In particular, all clones that were resistant to single-agent treatments became sensitive to the combination except one that was still resistant to the combination. This demonstrates the value in TraCSED in finding important pathways associated with a phenotypic variable, in this case clone fitness, and how this information can guide our biological understanding of drug response.

5.2 Results

5.2.1 Treatments induce different clonal selection and transcriptional responses

We performed a TraCE-seq experiment with T-47D cells treated with giredestrant, a selective estrogen receptor (ER) antagonist and degrader, or palbociclib, a CDK4/6 inhibitor, for 1, 4, 8, or 26 days (Figure 1A). The acute response from T-47D cells to either drugs is cytostatic as shown by the normalized growth rate (GR) values close to zero (Figure S1A) [62]. Both treatments showed a decrease of clone diversity, quantified by a reduction in the number of unique clonal barcodes in the cell population (Figure 2A). We classified clones as either negatively selected (NS) or positively selected (PS) based on the ratio of clone frequency in the overall population at day 26 versus day 0, which we termed the “endpoint selection value”,

Figure 1

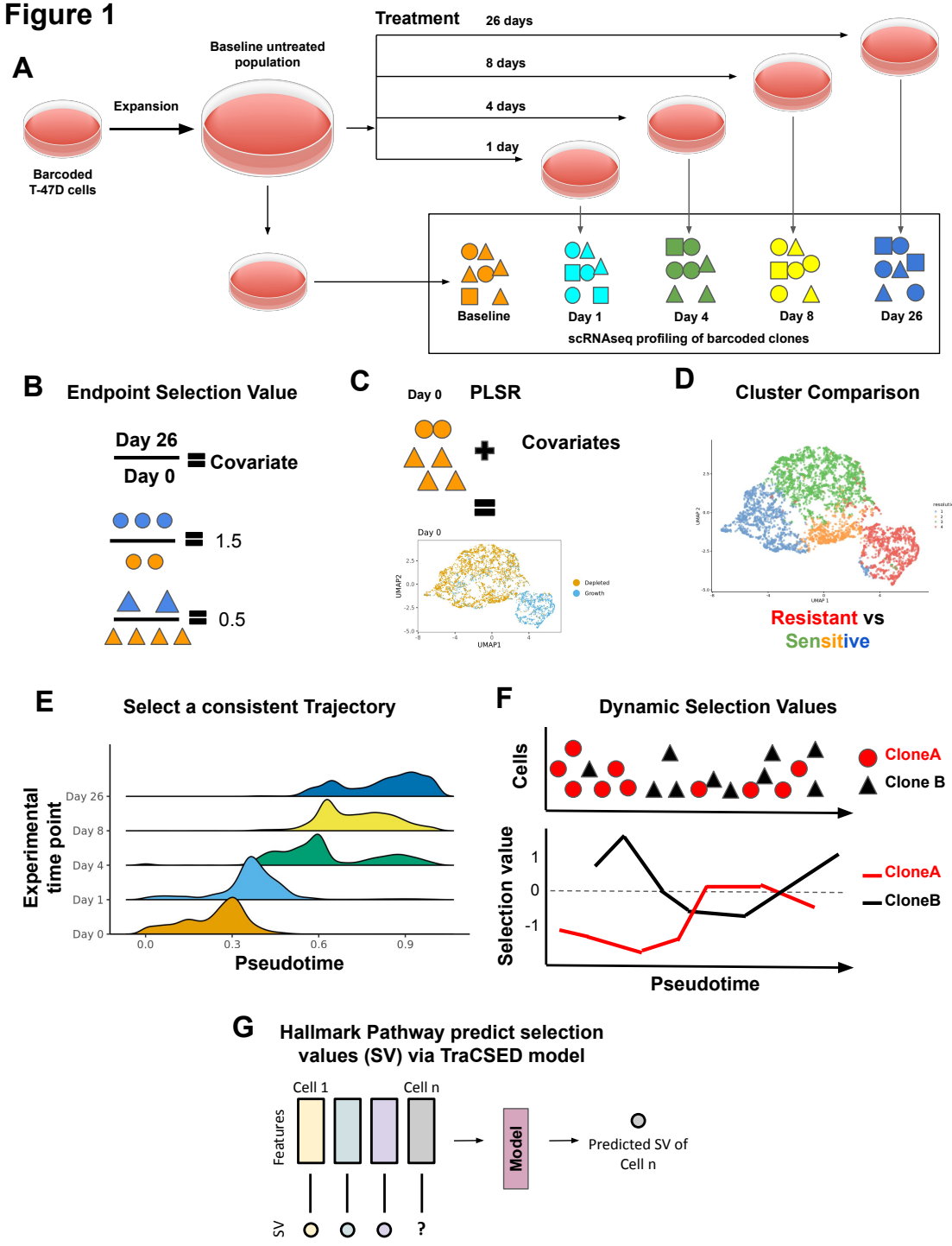

A
Baseline untreated population
Expansion
Barcoded T-47D cells
Treatment
26 days
8 days
4 days
1 day
Baseline
Day 1
Day 4
Day 8
Day 26
scRNAseq profiling of barcoded clones
B
Endpoint Selection Value
Day 26 / Day 0 = Covariate
1.5
0.5
C
Day 0
PLSR
Covariates
Depleted
Growth
UMAP1
D
Cluster Comparison
Resistant vs Sensitive
E
Select a consistent Trajectory
Experimental time point
Day 26
Day 8
Day 4
Day 1
Day 0
0.0
0.3
0.6
0.9
Pseudotime
F
Dynamic Selection Values
Cells
CloneA
Clone B
Selection value
1
0
-1
CloneA
CloneB
Pseudotime
G
Hallmark Pathway predict selection values (SV) via TraCSED model
Cell 1
Cell n
Features
SV
?
Model
Predicted SV of Cell n

Figure 5.1: (A) TraCe-seq experiment includes single-cell RNAseq data from a series of pre-treatment and post-treatment time points that originate from the same pool of barcoded cells. In our case, the treatments include either giredestrant alone (0.1μM), palbociclib alone (0.2μM), or their combination (same concentration). (B-D) Baseline investigations begin by determining the "endpoint selection value" for each clone (circles or triangles) by dividing the fraction of the clone 26 days after treatment by the fraction of clones present pre-treatment relative to the population (B). These covariates are assigned to clones at baseline and combined with the single-cell RNAseq data to create a partial least squares regression (PLSR) dimension reduction (C). A majority of resistant cells are specific to a cluster, which is then compared to clusters with a majority of sensitive cells (D). (E-G) To model adaptive resistance the selection values of clones across the time of treatment must first be determined. This is initiated by selecting a pseudotime cell-by-cell ordering inferred by slingshot which is consistent with the experimental time ordering (E). Clone selection values along the selected pseudotime are determined by dividing the prevalence of the cells ahead of a given cell in the trajectory by the prevalence behind it (F). Finally, clone selection values are paired with gene or pathway features for the single cell to model the prediction of single-cell selection value (G). This prediction considers the temporal context as well as the feature values, such that downstream model interpretability analysis yields dynamic features important in resistance.

a surrogate for clone fitness (Figure 1B). Clones with a greater frequency after 26 days of treatment would have an endpoint selection value greater than 1 and would be considered PS, whereas clones with less frequency would be less than 1 and NS. As an illustration, when we selected the 11 clones with at least 200 cells across all time points, we observed differences as to which clones were PS for giredestrant compared to palbociclib treatment (Figure 2B).

UMAP based on PCA showed that both treatments induced a strong transcriptional response with clear separation of the cells in each treatment group, suggesting the dynamics of response are characteristic of each treatment (Figure 2C, S1B). By day 26, the separation in transcriptomes between treatments was clear with cells from the palbociclib-treated sample overlapping in UMAP space with baseline cells. In contrast, the giredestrant-treated cells formed a separate UMAP cluster, reflecting a more pronounced shift in transcriptional state. In terms of pathway scores, reduction of ER activity is sustained throughout the time points of

the giredestrant-treated samples (Figure 2E), but the fraction of cells with E2F positive score returns to baseline levels in both giredestrant- and palbociclib-treated samples (Figure 2D). Giredestrant's inhibition of ER activity was maximal around 4-8 days aligned with its effect on cell cycle (Figure 2E). Consistent with this result, we observed by Western Blot a downregulation of ER and its target, the progesterone receptor (PR) in the giredestrant-treated sample (Figure S1C) and a regain of pRb in both PS populations. The PS populations for each treatment were, as expected, more resistant than the parental population to the treatment used for selection as shown by the positive GR values reflecting partial growth inhibition (Figure S1A, C). PS cells can proliferate, albeit slower when rechallenged with palbociclib treatment, whereas the PS cells in the giredestrant treatment became completely insensitive to giredestrant. We also observed a partial cross-resistance between the two drugs (Figure S1A): the palbociclib-PS population has much GR values for the response to giredestrant than the parental population and the reverse is also true.

5.2.2 PLSR identifies baseline signatures associated with selection of clones

The observed "rebound" in the E2F proliferation signature score could be due either to an innate fitness advantage of some clones prior to treatment, which allow those clones to overcome other ones, or it could be due to an adaptive response induced by the treatment itself in the PS clones. In either case, we hypothesized that the initial transcriptional state influences the likelihood of a clone being positively selected under treatment. However, initial attempts using pseudobulk samples in which cells from clones of similar endpoint selection values were grouped revealed few differentially expressed genes for either treatment condition (Figure S2A).

Figure 2

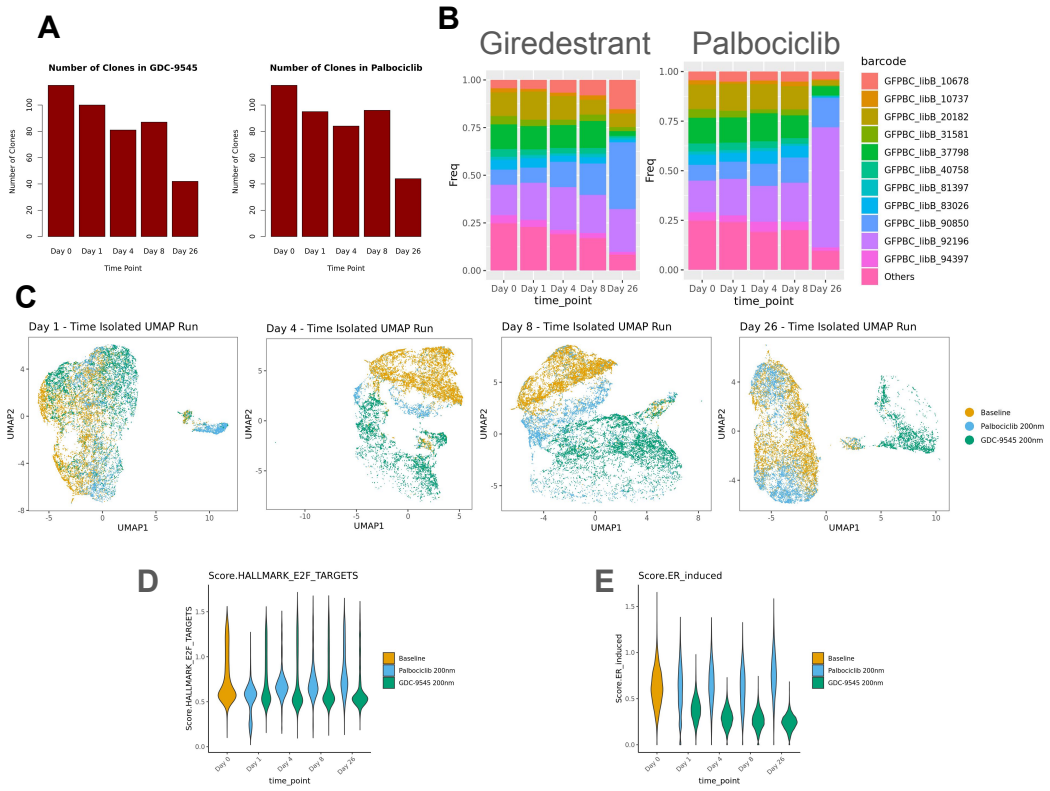


Figure 5.2: Giredestrant and palbociclib induce unique clonal selection and transcriptional responses over time. (A) Overall number of clones detected in samples of T-47D cells treated with either giredestrant (left) or palbociclib (right) at different time points. (B) Fraction of cells for clones with a minimum of 200 cells for giredestrant (left) or palbociclib (right) treatments at different time points. (C) UMAP of cells from the baseline sample and each time point individually for both giredestrant or palbociclib treatments. (D-E) Pathway scores from scuttle for the E2F targets pathway (D) and ER activity (E) for baseline cells and cells treated with either palbociclib or giredestrant at different time points.

Alternatively, unsupervised clustering via PCA did not separate PS and NS clones for either treatment due to high gene expression variability across cells of one clone compared to the separation achieved through PCA clustering (Figure S2B). Thus, both DE and PCA clusters failed to identify informative resistance pathways.

We therefore undertook a more sensitive approach to detect resistance pathways using semi-supervised partial least-squares regression (PLSR) dimensionality reduction (see Methods; Figure 1C). PLSR incorporates the transcriptional state as independent variables and the endpoint selection value as a dependent variable to search for a data projection in which PS and NS clones are well separated in the lower dimensional space. Using the PLSR-corrected UMAP, cells from PS clones were better separated from cells from NS clones (Figure 3A,B). In this projection, we were easily able to define clusters that were strongly enriched clones that were positively selected across time (Figure S2C, S2D) revealing baseline markers of selection (Figure S2E).

PS cells were aggregated into a single PLSR cluster for giredestrant treatment (Figure 3A), which enabled a pseudobulk differential gene expression analysis comparing expression observed in cells of the PS cluster to cells outside the cluster. Among the differentially expressed genes (DEGs), high SNHG25 and low CLDN1 expression were identified (Figure 3C, 3E). Similarly, a PS cluster was also identified for the palbociclib-treated cells (Figure 3B), with SNHG25 and CLDN1 identified as DEGs for the palbociclib PS cluster (Figure 3D, 3F). CLDN1 is a known marker used for classifying subtypes of breast cancer, and low CLDN1 is a marker for aggressive triple-negative BRCA and predictive of recurrence [196, 107, 95, 49]. SNHG25 has been shown to be associated with tumor activity, but is poorly

Figure 3

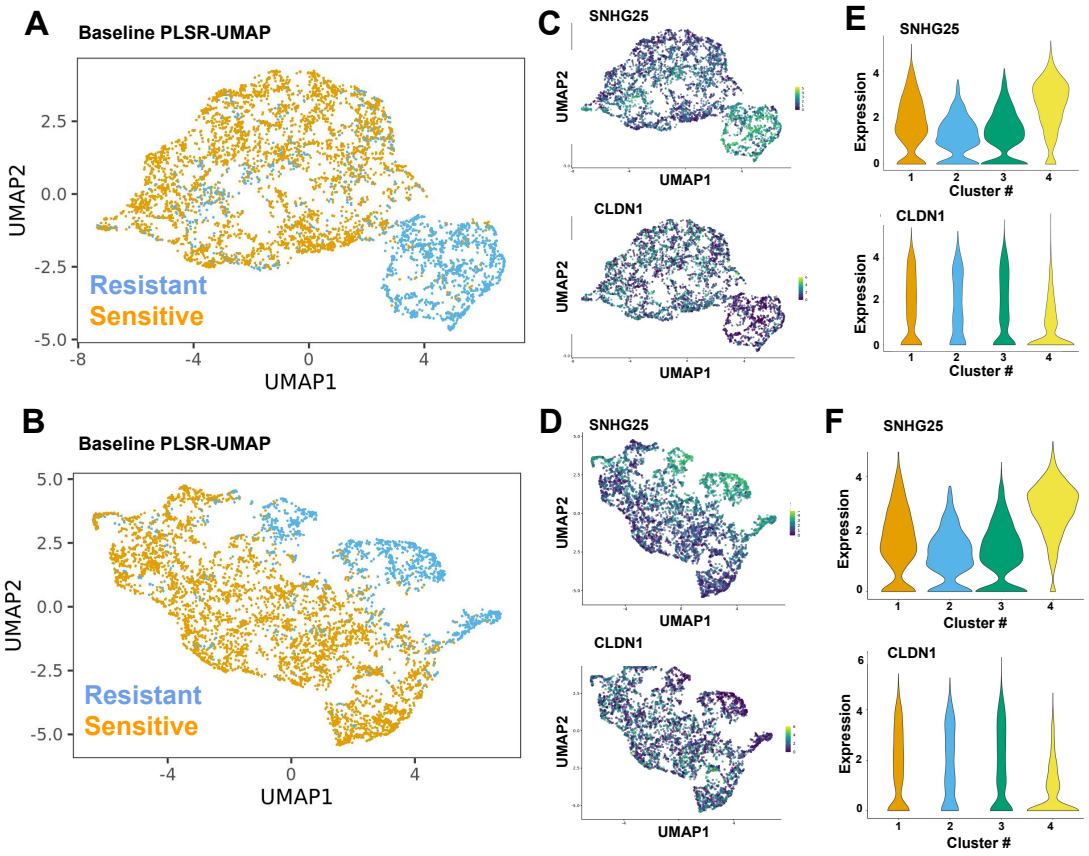


Figure 5.3: Partial least squares regression (PLSR) reveals baseline signatures hidden by traditional approaches like differential expression. (A-B) PLSR-based UMAP for baseline cells associated with outcome of either giredestrant (A) or palbociclib (B) treatment. Positively selected (PS) cells are in blue and negatively selected (NS) in orange. (C-F) SNHG25 and CLDN1 expression levels are displayed in the UMAP (C-D) and violin plots (E-F) for giredestrant (C,E) or palbociclib (D,F). Cluster 4 contains most PS cells for both giredestrant and palbociclib (see Figure S1D).

understood [90, 195, 65, 193, 14].

5.2.3 Fitting a generative model for drug resistance

As both positively and negatively selected clones had a strong transcriptional response as illustrated in the UMAP (Fig S3C-E), we extended our modeling framework to leverage the longitudinal transcriptional data and explore potential acquired resistance mechanisms. We therefore took a dynamical modeling approach to identify the transcriptional programs associated with the clonal selection process revealed by the TraCE-seq data. We assumed that a population of cells ordered along a continuous cell trajectory captured the progression from sensitivity to resistance of individual clones. We used the Slingshot method to estimate the cell trajectory and project the cells along a response curve, establishing a pseudotime ordering of the cells. We calculated selection values across the Slingshot pseudotime for each clone (Figure 4A-B, see Methods). To associate transcriptional changes with treatment response, we developed an interpretable generative model that predicts selection values based on a set of pathway features, called TraCSED for Transformer-based modeling of Clonal Selection and Expression Dynamics. After the Slingshot pseudotime is established, we calculated the selection value along the trajectory for every cell belonging to a clone (Figure 1E-F, Figure 4A-B). The selection value for a cell belonging to clone c at time t was calculated as the fraction of clone c's cells present after time t in the pseudotime ordering. The transformer was then trained to predict the selection value of the cell at time t using its own features as well as the features and selection values of clone c cells in the context of the cell of interest (Figure 1G). Testing of model accuracy occurred over four pseudotime intervals (Figure 4C). In this way, the model

learned a conditional relationship between the features and selection values based on where the cell resided in pseudotime.

We required a minimum number of cells (set to n=200) to be present across the time course to model the selection process of a specific clone as selection values were too noisy otherwise (see Methods). 11 clones in total fit this criteria for at least one treatment (Palbociclib - 9, Giredestrant - 10, Combination treatment - 11). To predict the selection value across time (Figure 4D, see Methods), we used the individual cell Hallmark pathways scores [88] as features. Across the four testing periods, the quality of predictions varied, but the general observation was that the prediction was good when the selection values remained in the range to the preceding training data. However, the selection value can get noisy in the last testing period due to its dependence on the number of cells, which decreased at the end of the trajectory (Figure 4E). When assessing the contribution of the attention and convolutional layers, we found that the combination of both sets of layers yielded the best compromise of fit quality and prediction accuracy (Figure S4). While the performance of the model may not be enough to extrapolate beyond the experimental time, TraCSED provides interpretability of the features associated with the positive or negative selection values of specific clones.

5.2.4 Model pathway features reveal adaptive resistance mechanisms

Permutation of TraCSED's input layer enables the identification of features that induce changes to the predicted selection value, a procedure called permutation importance [6]. In addition, by altering the inputs as a function of pseudotime, a dynamic pathway importance can be calculated, enabling the interpretability across time for individual clones. Using this

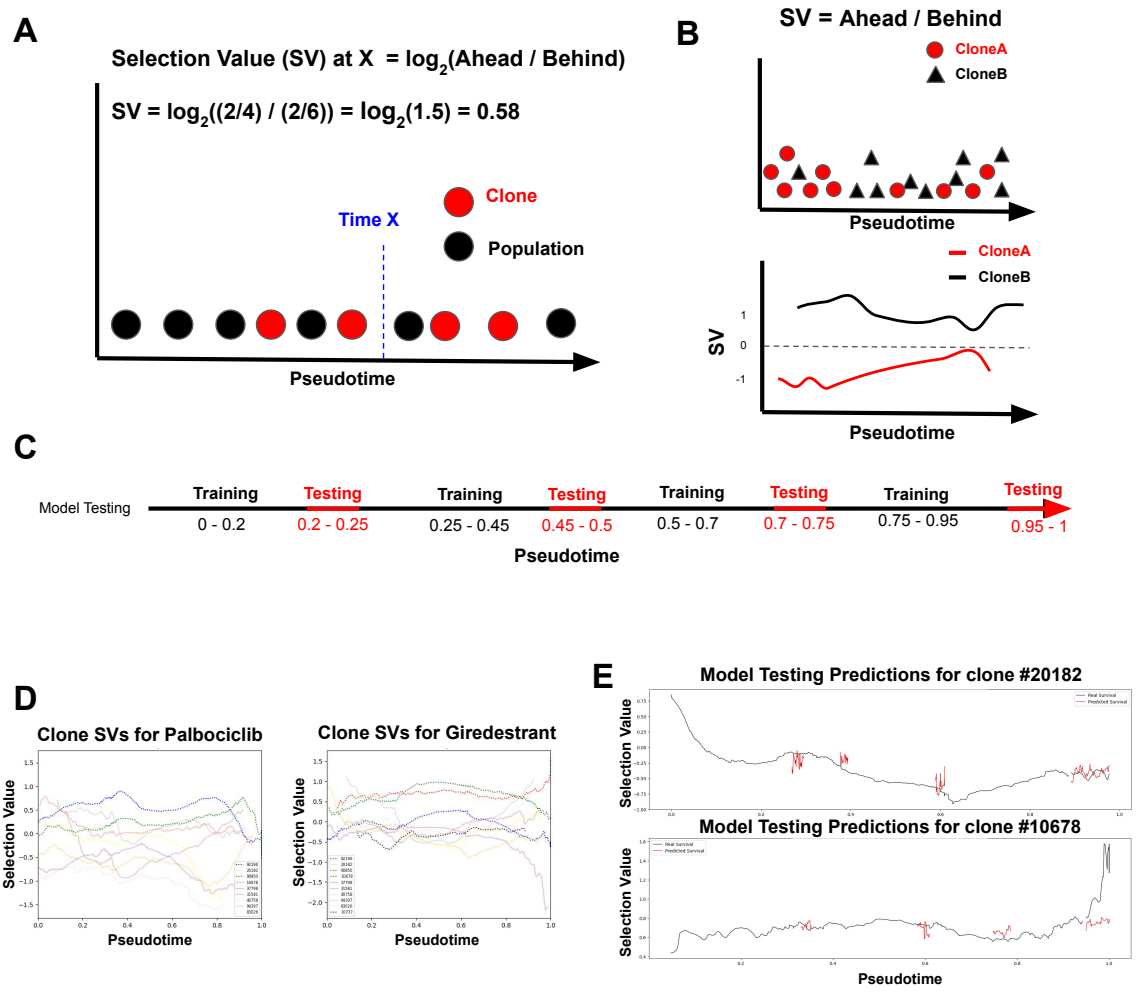


Figure 5.4: Transformer-based modeling of Clonal Selection and Expression Dynamics (TraCSED) predicts selection values across pseudotime. (A-C) An example clone (red) exists as part of a total population of cells (black) across a pseudotime progression. To determine the selection value at any given time (blue) for the clone in question the fraction of cells ahead and behind a given pseudotime is used (A, see Methods). This results in a continuous selection value curve for each clone across pseudotime (B). This model of dynamic selection is then tested in four different increments (red) along pseudotime to ensure a quality model fit at all time points (C). (D) The resulting transformation from raw data along pseudotime to selection value curves for each clone is shown for each clone in palbociclib (left) and giredestrant (right) treatments. (E) Example of the model testing (red) for two clones under giredestrant treatment is displayed.

procedure we found, for example, permuting scores from "Epithelial Mesenchymal Transition" (EMT) resulted in a deviation from the true selection value (upper panel) and a spike in feature importance (middle panel) within a pseudotime range that aligns with the 4 to 8 day time points of giredestrant treatment (lower panel) (Figure 5A). This exemplifies how the model is made interpretable for a single clone. In order to understand whether these selection mechanisms observed were more universal, we plotted pathway scores and feature importance for all modeled clones together. Scores for the EMT pathway in PS clones were decreased compared to NS clones between 0.5-0.9 pseudotime for giredestrant treatment (Figure 5B).

Similarly, we examined the feature importance for clones undergoing palbociclib treatment. For example, the E2F target score was important between 8 to 26 days (Figure 5C). When looking at all clones together, the "Estrogen Response Late" scores for PS clones were important and increased above NS clones in pseudotime ranges of 0.35-0.5 corresponding to the early time points (Figure 5D). This Estrogen Response signature in PS clones preceded another signature where E2F target scores increased within pseudotime ranges of 0.6-0.8 (Figure 5E). Other pathways were similarly noted as important in selection including "G2M checkpoint" in palbociclib and the AKT pathway in giredestrant (Figure S5). Taken together, these findings suggested that single-agent treatment was not sufficient to overcome adaptive resistance but the combination of palbociclib and giredestrant might be effective because the giredestrant may prevent the positive selection of clones in the palbociclib treatment that had higher ER activity.

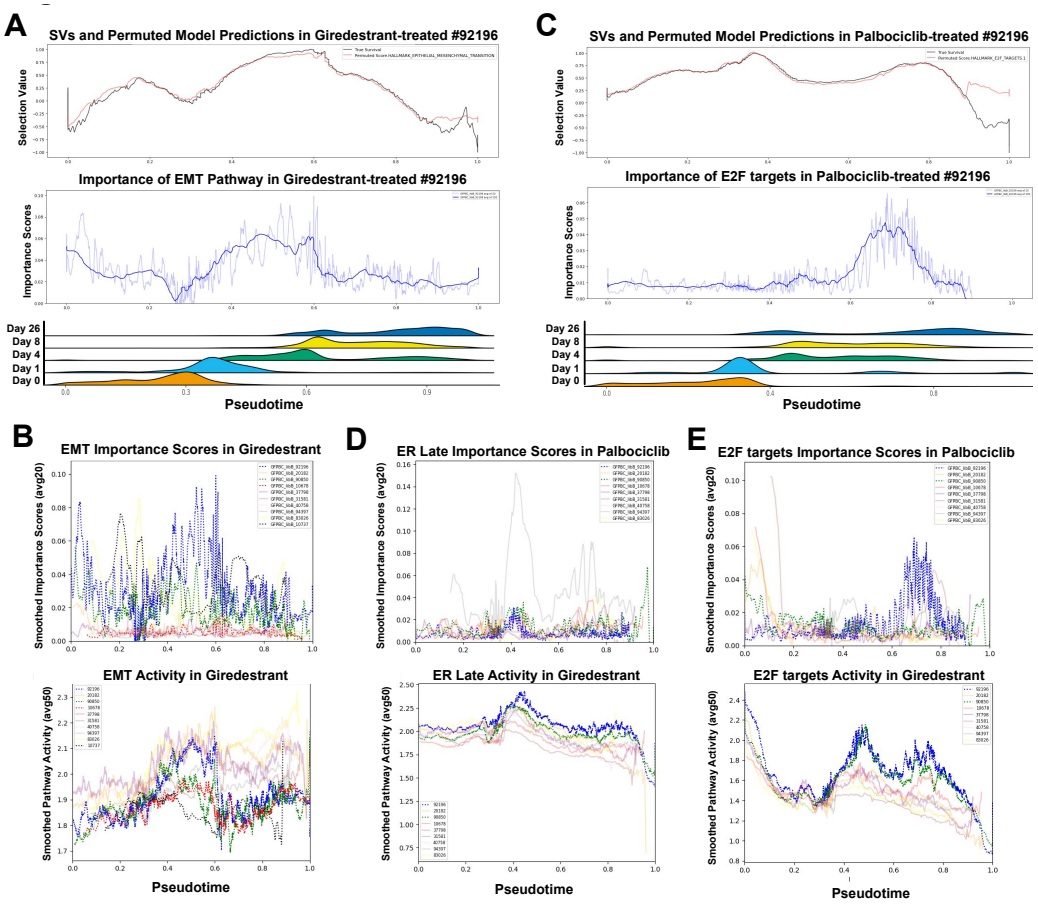


Figure 5.5: (A) Predictions by a clone model with permuted EMT activity (red) were plotted against the true selection values (black) under giredestrant treatment (A, upper panel). The models reported feature importance for the EMT pathway with averaging smoothing of 20 (light blue) or 150 (dark blue) adjacent cells is shown along pseudotime (middle panel). The ridge plot of pseudotime distributions for each time point in giredestrant treatment is aligned with the above plots (lower panel). (B) Smoothed model importance (averaging 20 adjacent cells) was then plotted for PS clones (bold, dashed lines) and NS clones (faint, solid lines) in the giredestant treatment for the EMT pathway (upper panel). Smoothed pathway activity (averaging 50 adjacent cells) was also plotted (lower panel). (C-E) Similar to A-B for palbociclib treatment: true selection values (black) and a clone model with permuted E2F targets (red) were shown (C, upper panel) with smoothed model importance scores for the pathway (C, middle panel) and ridge plots of pseudotime distributions for palbociclib treatment (C, lower panel). Smoothed model importance (upper; adjacent 20 cells) and smoothed pathway activity (lower; adjacent 50 cells) were then plotted for Estrogen Response Late (D) and E2F targets (E) in palbociclib treatment.

5.2.5 Adaptive resistance is reduced with combination treatment

Next we treated cells with palbociclib and giredestrant as a combination to investigate our earlier finding. Combination therapy proved to be effective in suppressing proliferation and ER activity across the whole population, which was not previously achieved by palbociclib or giredestrant alone (Figure S6A-B). While most clones that were PS in only one of the single agent treatments became NS in the presence of combination, one clone was PS in both single-agent treatments and remained PS in the combination treatment after six months (Figure 6A). The selected population is indeed much more resistant to treatment than individual resistant populations (Figure S1B). Similar to the giredestrant-PS population, the surviving population lost ER and PR expression compared to the parental population suggesting independence on ER activity (Figure S6C). In addition, surviving cells have lower cyclin D expression, while retaining high pRB, showing less dependency on CDK4/6-CyclinD activity to proliferate (Figure S6C). Our PLSR method identified two different clusters associated with PS (Figure 6B and S6D). Similarly to individual treatments, high SNHG25 was also associated with resistance in combination, but in contrast, CLDN1 no longer showed enrichment in the PS clusters. Additionally, SNCG was specifically associated with resistance in combination therapy (Figure 6C-E). SNGC is a synuclein protein known as a marker for late-stage breast cancer [184, 162, 197], implying that cells resistant to combination treatment could be reflecting a cellular state found in tumors from late-stage patients.

We then used TraCSED to characterize the dynamic selection process occuring in the combination treatment. Because the Estrogen Response signature was important in the

selection process for palbociclib alone and preceding higher E2F targets signature scores, we were curious to see the effect of suppressing ER signaling with giredestrant. We observed that both ER and proliferative signatures had lower activity and showed reduced importance in predicting the selection values of PS clones (Figure S6E-F) compared to what was observed for the palbociclib treatment. Similarly, the EMT pathway had low activity in PS clones under giredestrant treatment, but in the combination treatment all clones had similar levels of EMT pathway activity (Figure S6G).

While the combination treatment is overall more potent, one PS clone (92196) still emerged after 6 months. Scores for key gene signatures such as "Hallmark Estrogen Response Late" (Figure 6F), "ER induced" (Figure 6G), or "EMT" (Figure 6H) pathways were not substantially different than a clone (90850) that is outcompeted at 6 months. The lack of differentiation is consistent with our model identified that the features associated with positive selection are substantially less important in the combination treatment compared to the single agent ones. Therefore, baseline expression of the SNHG25 and SNCG genes found through PLSR are the primary markers of positive selection for combination treatment and differentiate clone 92196 from the other ones (Figure S7). In conclusion, TraCSED is complementary to PLSR for identifying the different factors associated with adaptive and innate resistance. PLSR highlighted markers of innate resistance that are consistent with late-stage cancer phenotypes, whereas TraCSED allowed us to identify vulnerabilities in adaptive resistance to single agent treatments that were addressed through combination treatment.

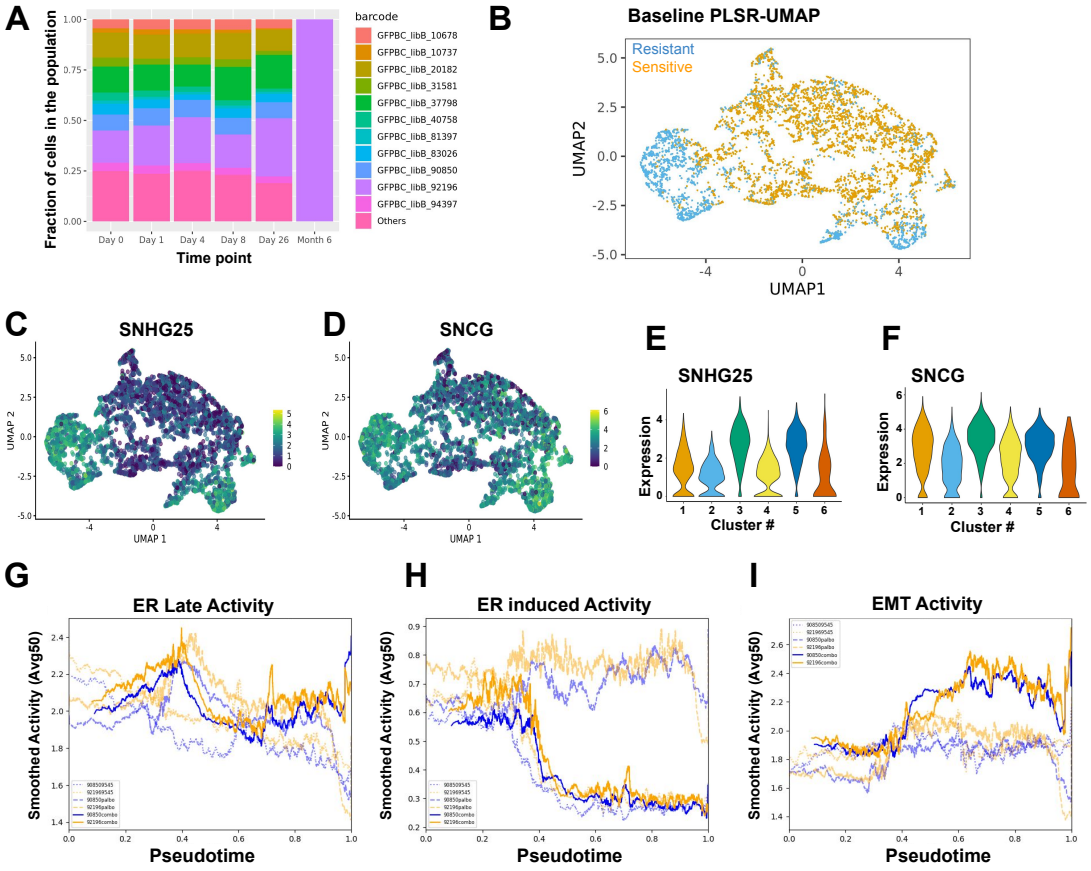


Figure 5.6: Combination of palbociclib and giredestrant reduces the importance of adaptive resistance mechanisms. (A) Fraction of cells for clones with a minimum of 200 cells in the combination of giredestrant and palbociclib at different time points. (B) PLSR-based UMAP for baseline cells associated with outcome combination treatment. Positively selected (PS) cells are in blue and negatively selected (NS) in orange. (C-D) Gene expression in baseline cells was shown via UMAP from (B) for SNHG25 (C) and SNCG (D). (E-F) Violin plot of SNHG25 (E) and SNCG expression (F) in baseline cells. PS cells are in clusters 3 and 5 (see SFig 6C). (G-I) Pathway activity for a PS clone (orange) and a stable clone (blue) plotted for palbociclib (faint, dashed lines), giredestrant (faint, dotted lines), and their combination (bold, solid lines) using gene sets for Estrogen Response Late (G), ER induced (H), and Epithelial Mesenchymal Transition activity (I).

5.3 Discussion

Giredestrant and palbociclib were effective in blocking proliferation of T-47D cells and resulted in a selective bottleneck as reflected by the reduction of the number of clones over time. This process was associated with changes in the transcriptomes of cells, in particular down-regulation of ER and, respectively, proliferative signaling. However, some clones were positively selected and differences were observed in the selection process between the two single-agent treatments (Figure 2A-E). This positive selection can occur due to an innate resistance detectable in pre-treatment conditions, or it could be a result of an emerging adaptive resistance. To investigate these mechanisms, we developed a PLSR approach for finding markers of innate resistance (Figure 1A-D), and a transformer-based model, called TraCSED, for finding markers of dynamic selection in adaptive resistance (Figure 1E-G).

We found that both unsupervised clustering and differential expression analysis comparing PS and NS clones of pseudobulked samples to study innate resistance was not effective in this study (Fig S2A-B). On the other hand, supervised clustering based on PLSR balances intrinsic variability of the clone with the outcome and enabled us to isolate PS clones into a single cluster for giredestrant (Figure 3A) and palbociclib (Figure 3B), which revealed markers unique to the PS cluster (Figure S2E). Two examples of these PS markers were SNHG25 and CLDN1. In giredestrant and palbociclib treatments, SNHG25 is increased in the PS cluster and CLDN1 is decreased (Figure 3C-F). SNHG25 is associated with a variety of cancers [90, 195, 65, 193], and is part of a family of long-noncoding RNAs (SNHGs) with multi-functional roles in cancer progression, particularly in the EMT pathway [14]. In addition, CLDN1 levels are well-

documented for their role in EMT and breast cancer, have been used previously for subtyping different types of cancer such as the claudin-low basal subtype in breast cancer [68], and correlate with breast cancer recurrence [196, 107, 95, 49]. PCA-based clustering was not able to identify those genes (Figure S2F), and neither of these PS markers were found with pseudobulk differential expression (Figure S2F), demonstrating the value of PLSR in finding mechanisms of innate resistance.

In addition to innate mechanisms of resistance, we observed transcriptional changes in both PS and NS clones after treatment (Figure S3C-E). Because existing methods do not associate outcomes like selection with temporal transitions, we developed TraCSED, a context-aware and interpretable model composed of attention and convolution layers, to investigate mechanisms of adaptive resistance. The dynamic selection for each clone is modeled across pseudotime (Figure 1E-G; Figure 4A-B) and then tested at four pseudotime intervals (Figure 4C, E). The TraCSED model facilitates interpretability by highlighting features important for the selection process of individual clones (see Methods). The EMT pathway had high model importance and low activity for PS clones in giredestrant treatment (Figure 5B), suggesting this pathway may be associated with resistance. In palbociclib treatment, there was high model importance and increased activity in the Estrogen Response Late pathway (Figure 5D) which preceded a subsequent increase in the E2F targets pathway (Figure 5E). Thus in both cases, the TraCSED model captured an adaptive resistance mechanism to single-agent treatments.

When combining giredestrant and palbociclib, only one PS clone survived 6 months of combination treatment (Figure 6A). Consistent with our hypothesis, the adaptive resistance pathways we identified with TraCSED in the single agent treatments were no longer associated

with dynamic selection in the combination (Figure S6C-E). Additionally, the single surviving PS clone did not demonstrate activity in the ER and EMT pathways differing from NS clones (Figure 6F-H). Thus, only the high SNHG25 and high SNCG activity (Figure 6C-E and S7) identified by the PLSR approach in the two PS clusters differentiated the PS clone from the NS ones suggesting that innate resistance mechanisms were driving resistance to giredestrant and palbociclib combination treatment. Increased SNCG is a known marker of resistance and late-stage Breast cancer [184, 162, 197], suggesting that the innate cellular state of these PS clusters may resemble a state found in late-stage tumors. While we elected to use pathway features for our modeling for compute efficiency, increased interpretability, and reduced noise, our modeling approach can also be applied at the gene level. In cases where data quality allows for gene-level analyses, TraCSED may be able to identify individual genes associated with resistance that may not be captured in pathway-level analyses.

Determining how a cell will respond to treatment and how it might be resistant is an ongoing field of study. Here we developed a PLSR approach and the TraCSED model for determining markers of innate and adaptive resistance, respectively. With PLSR, we were able to identify SNHG25 and CLDN1 markers of positive selection in single-agent treatment, which were not found by traditional methods of pseudobulk and unsupervised clustering. Through TraCSED dynamic modeling we found pathways associated with adaptive resistance, including increased ER activity in palbociclib treatment. We were then able to show that the adaptive resistance is reduced when palbociclib is combined with the ER suppressing activity of giredestrant. Consistently with our hypothesis, our model identified that no feature was associated with an adaptive resistance in the combination treatment, leaving the innate resistance

marker, SNHG25 and SNCG, found with the PLSR model, as the main features associated with positively selected cells. Altogether, our two complementary modeling approaches contribute to study and characterize non-genetic sources of therapeutic resistance.

5.4 Methods

5.4.1 Datasets and Dependencies

T-47D breast cancer cells were labeled with TraCe-seq library as previously described [28]. A subculture was established from 300 cells to establish the input cell population. These 400 cells were expanded to 2 million cells total, and randomly distributed to the following treatment groups in replicates: 100 nM giredestrant treatment, 200 nM palbociclib treatment, and 100 nM giredestrant + 200 nM palbociclib combination treatment. Cells were trypsinized after 1, 4, 8, or 26 days of treatment and profiled by single cell RNA-sequencing using Chromium Single Cell 3' Reagent Kits (10x Genomics). Baseline scRNA-seq profiling was obtained after 1 day of treatment in DMSO. In addition, for the combination treatment, cells were harvested and profiled after 6 months following exposure to both compounds.

5.4.2 Drug response

T-47D (female, adenocarcinoma from pleura) ER+ breast cancer cell lines were cultured using standard aseptic tissue culture techniques at 37°C in RPMI medium supplemented with 10% FBS, 2mM L-Glutamine (Catalog 10440, Sigma), 1X Minimum Essential Media, Non-Essential Amino Acids (MEM NEAA, Catalog No. 11140-0500, Thermo Fisher) and

1X Antibiotic-Antimycotic (Catalog No.15240-112, Thermo Fisher). Cell line ancestry was determined using Short Tandem Repeat (STR) profiling using the Promega PowerPlex 16 System. Cells were treated for 7 days with either 0.2μM of giredestrant, 0.2μM of palbociclib, or their combination and the relative number of viable cells was determined by CyQUANT (ThermoFisher, C7026). Growth rate normalization was performed [62].

5.4.3 Single-cell analysis

For pathway features, we used Hallmark pathways [88] and a select set of curated pathways. Single-cells were scored on their pathway activity using the scuttle package in R [106]. Activity across time and treatments is shown via violin plot. Additionally, PCA dimensionality reduction to 5 principal components, followed by UMAP plotting was used to view how cells transcriptomic features change throughout time in the different treatments.

5.4.4 Quality control

The number of clones present across time and treatment was plotted. Clones with 200 cells or more were displayed with a stacked barplot showing the various clones' relative proportion across time and treatment (Figure 2B). Clones were considered positively selected if the proportion of the cells increased relative to the population. All clones that did not have a total of 200 cells across time were combined and displayed as "other". No bias in read counts between PS and NS clones or across treatments and time points was observed.

5.4.5 Partial least squares regression (PLSR)

Phenotypic comparison based on pseudobulk was performed with DESeq2 differential expression (DE) [94] on clonal response to treatment in which clones that increased over their baseline percentage by 26 days of treatment (resistant) were compared to those that were absent after 26 days (sensitive). Additionally, expression based PCA clustering was performed to group cells by similar transcriptomes.

For the partial least squares regression (PLSR) method [1], the endpoint selection value is included as a covariate for the dimensionality reduction of expression that we termed the selection value. The endpoint selection value is a ratio of the cells present after 26 days of treatment over the cells present at baseline. The first 5 components are then used to plot a UMAP representation of the data. Leiden clustering was then done at various resolutions. Pairwise distance of PS and NS cells was performed for both the PLSR and PCA methods to compare which is better at distinguishing the two states. To understand how the clones are distributed across the PLSR clusters, a heatmap displaying the fraction of the clone population in each cluster was plotted. Alongside the heatmap is included the increase (red) or decrease (blue) for each clone compared to pretreatment conditions for Days 1, 4, 8, and 26 after treatment (Figure S2D). Resistant clones were concentrated in a single cluster which was then compared to sensitive clusters via the findMarkers package by scran [99]. PS (blue) and NS cells (orange) were plotted as a UMAP. PS markers found through findMarkers were plotted via UMAP and violin plots.

5.4.6 Overview of the dynamic generative model

The fundamental steps of the generative modeling approach are 1) selecting an appropriate pseudotime trajectory representation of the data, 2) determining the selection values for each clone across pseudotime, 3) modeling the selection values of each clone, 4) evaluating the model for overfitting, and 5) assessing feature importance across time through permutations of the model. This method allows for the prediction of selection values, but its primary focus is to avoid overfitting and to create feature interpretability across time.

5.4.7 Selecting an appropriate trajectory representation

In order to model the single-cell selection value of a particular clone, cells must be ordered along a continuous trajectory with start and end points consistent with the experimental design. In this case the trajectory continuum is made using Slingshot pseudotime inference [152]. Selecting a representation that closely resembles the experiment is done through taking the TRACE-Seq data and 1) performing dimensionality reduction through PCA or PLSR, 2) producing Leiden clustered representations at various resolutions, 3) performing Slingshot inference with assigned start and end clusters and normalizing pseudotime, 4) selecting an appropriate pseudotime representation through a summation of KS statistics between TRACE-Seq time points. This results in a pseudotime representation that can be used for modeling clonal selection values.

The TRACE-Seq data is represented in both an unsupervised (PCA) and semi-supervised (PLSR) fashion for comparison. Leiden clustering is then performed on a series of resolutions

between 0.2 and 1. Start and end clusters are then selected (based on consistency with the experiment and maximizing distance between start and end points) as input for each and used as input for Slingshot. The pseudotime inference produced by slingshot is then normalized on a 0-1 scale, where the first percentile of pseudotime is set as the minimum value, 0, and the ninety-ninth percentile of pseudotime is set as the maximum value, 1. Each of these normalized pseudotime series is then assessed via a KS statistic summation [103], where the pseudotime distribution of Day 0 is compared to Day 1, Day 1 to Day 4, etc. The trajectory representation with the largest KS statistic summation is then selected and used for modeling clonal selection values.

5.4.8 Determining clonal survival for modeling

Modeling resistance in an interpretable way begins with clonal selection values. With four different testing periods and a quality model fit being the utmost priority for interpretability, clones must have a minimum of 200 cells across time to be modeled. This left 11 clones across treatments which have sufficient cells. To make selection value a dynamic metric, we map it across the pseudotime continuum. This is done by considering the number of cells of the clone of interest compared to the total population. At each pseudotime t that a cell exists, a calculation is made where the cell population "ahead" of the cell of interest in pseudotime is divided by the cell population "behind" it (Figure 4A) using the following equation:

$$log2((\#cloneA_{T_{after}}/\#population_{T_{after}})/(\#cloneA_{T_{before}}/\#population_{T_{before}}))$$

To avoid the strong fluctuations occurring at the start and end points of the selection

value curves due to a smaller number of cells present at the extremes (Figure 4B), we fit a smoothing process through iteratively averaging 5 adjacent cells (Figure 4D). Then, a boundary threshold is placed to preserve the regions of pseudotime with less than 0.15 max-normalized absolute change in selection value.

5.4.9 Framework and modeling design choices

Our primary objective is an interpretable model across pseudotime that avoids overfitting, whereas extrapolation beyond the experimental time is not of interest. Since the modeling is sequential and we want to evaluate the fit across the whole pseudotime, we decided to test our model on multiple intervals instead of only testing the end of the pseudotime. We withheld four intervals that represented 20% of the pseudotime for testing: [0.2-0.25], [0.45-0.5], [0.7-0.75], and [0.95-1] (Figure 4C). Additionally, clones are trained and tested separate from each other to create individual models. This facilitates interpretability and keeps resistance mechanisms distinctive between clones.

5.4.10 Evaluation of the model

In order to evaluate the contributions of different aspects of the model, we employed three "control" approaches: regression, a transformer without convolutional layers, and a transformer without attention layers. Regression provides a simple baseline by which to assess whether the temporal context adds value to the model fit. Subtracting the convolution or attention aspects of the model, reveals the influence of these aspects' importance to the architecture for generalizing an accurate learning rule from the data. These models are then assessed via

mean squared error (MSE) on the testing data in the four time periods described above (Figure 4E).

5.4.11 Temporal importance of features through permutation

To assess the influence of features on the prediction of the selection value we used permutation feature importance [6]. In this approach we randomized a feature of interest 30 times and used the model with the altered feature to make a prediction. Changes in an important feature result in increased prediction error from baseline. For the 30 permutations of a feature, we compute the error at each pseudotime position to assess which features are important in predicting the selection value at specific pseudotimes. The mean error at each position is used in downstream analysis.

5.4.12 Interpretability of the model

As an example of how feature permutation provides interpretability to the model of dynamic selection in a particular clone, the true selection value curve was plotted (black) with the permuted model's selection value prediction (Figures 5A and 5C). Below this was the model importance scores for the permuted pathway with various smoothing factors (averaging the adjacent 20 or 150 scores). Ridge plots of the pseudotime distributions of real time were plotted to understand how the pathways found to be important can be attributed to periods within the experiment.

Mean error scores from each of the clone models were smoothed by averaging the adjacent 20 scores and plotted together. The activity for the feature is also plotted along pseu-

dotime with a similar smoothing process of averaging the adjacent 50 scores. In each of these plots, PS clones are shown in bold, dashed lines and NS clones in faint, solid lines (Figures 5B and 5D-E). This was done for the Epithelial Mesenchymal Transition, Estrogen Response Late, and E2F targets pathways. To compare palbociclib, giredestrant, and combination treatment the activity scores for a PS clone and a stable clone were plotted for EMT and ER pathways. In these plots the pathway activity of the PS clone (orange) is plotted for each of the three treatments across time with the activity of the stable clone (blue) so that the two clones and treatments can be assessed (Figure 6G-I).

5.4.13 Software availability

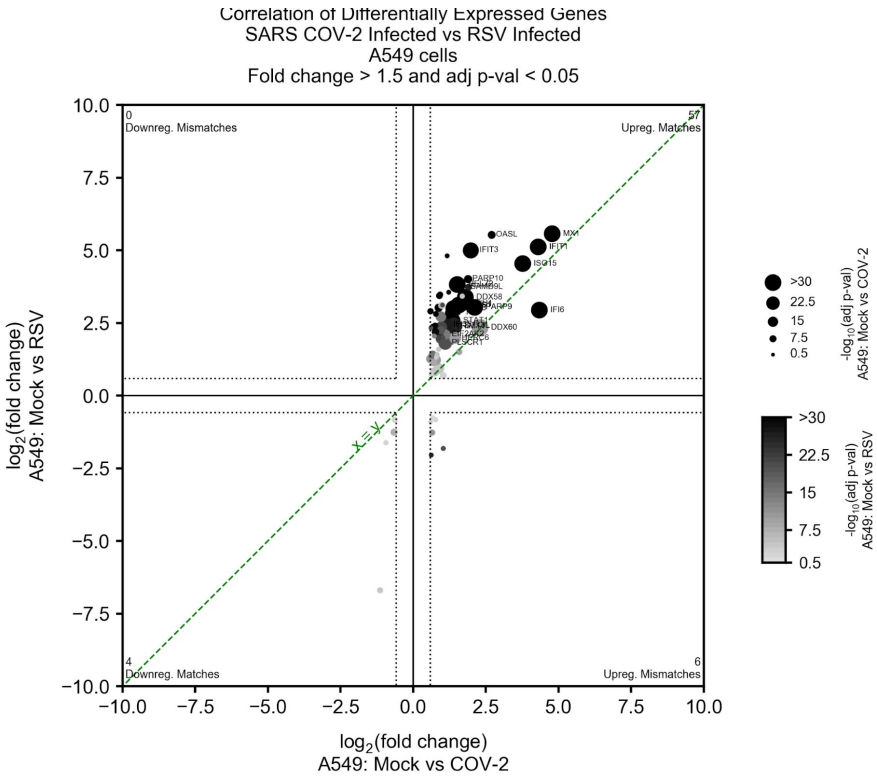
The code used for the TraCSED model and PLSR investigation of innate resistance in this paper can be found at https://github.com/Genentech/TraCSED.

5.4.14 Data availability

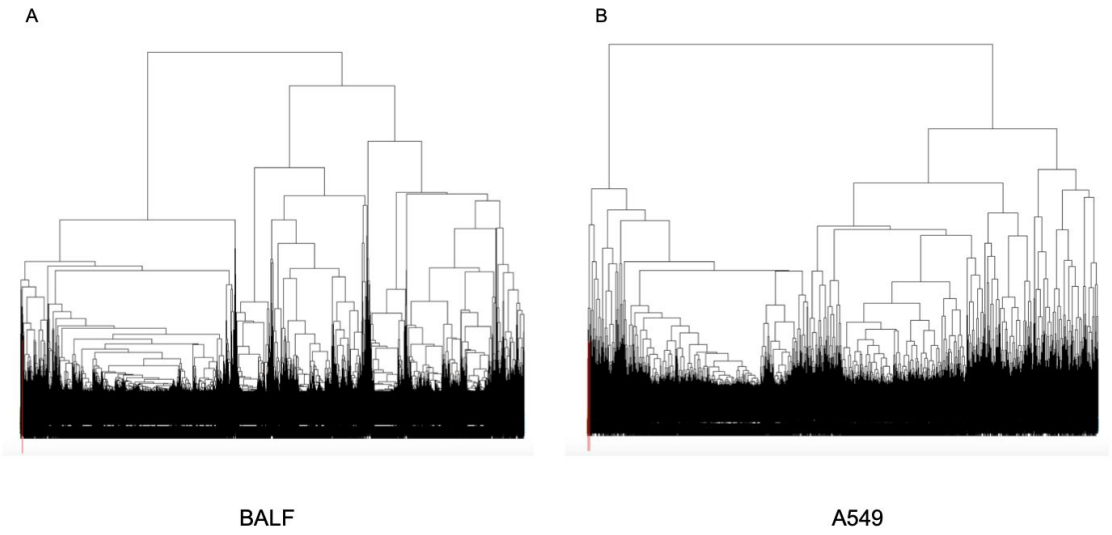
Single-cell data can be accessed in GEO under the series GSE260703 (reviewer access token: yxkxcwikpzobhcr).

Chapter 6

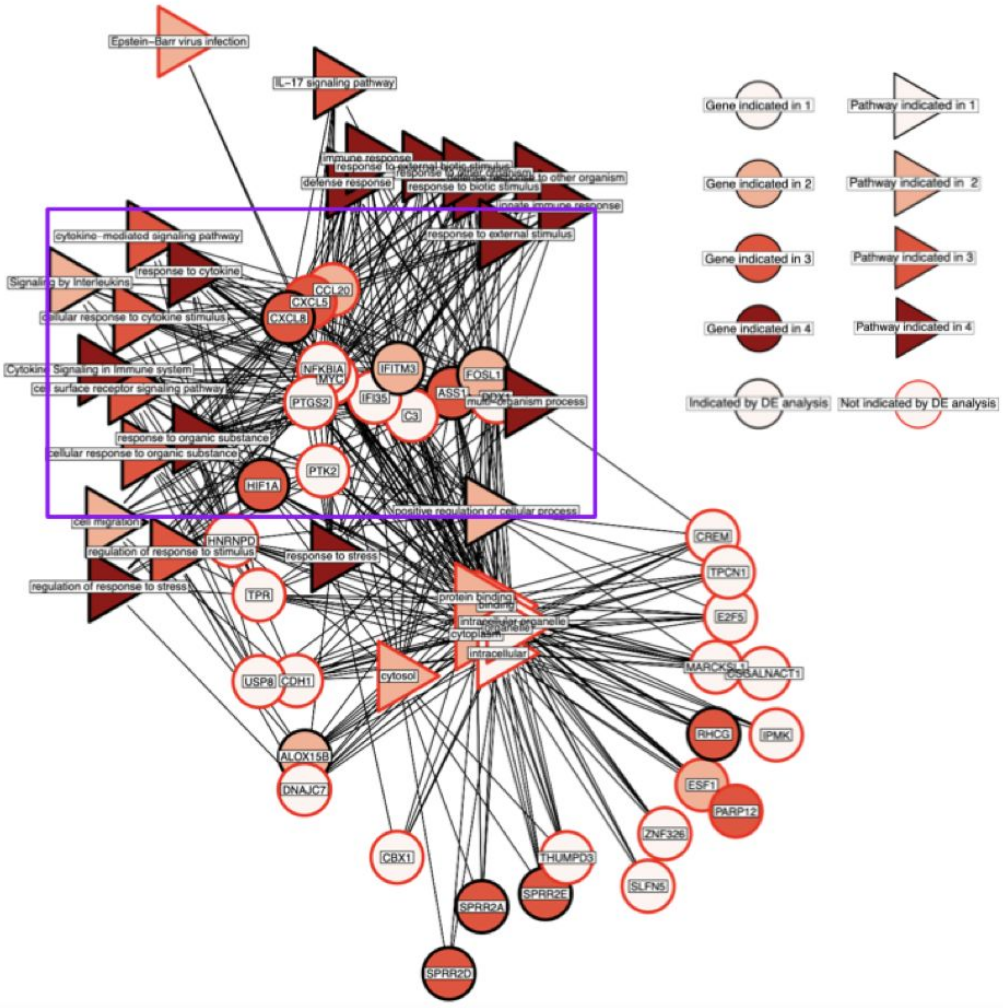
Supplementary Figures



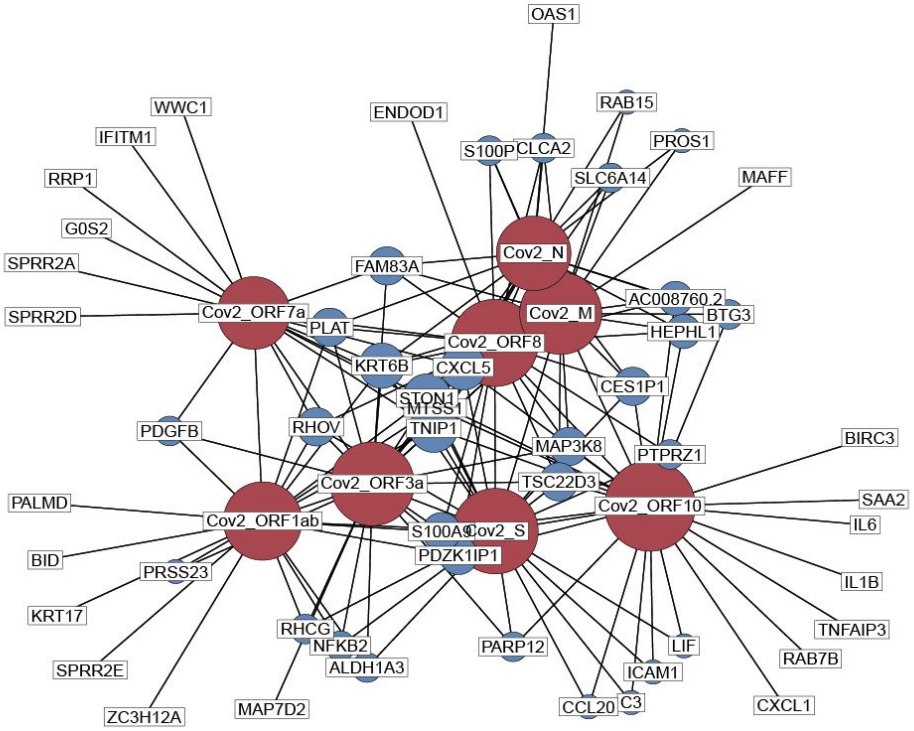
Supplemental Figure 1. Comparison differential expression programs in SARS-CoV-2 (COV-2) infected versus respiratory syncytial virus (RSV) infected cells. Differential expression (DE) of infected versus control in A549 cell lines produced a set of genes (circles, symbols displayed for a selected few) passing a log2 fold change (1.5) and significance cutoff (p-val$<$0.05, dashed vertical and horizontal lines) in both RSV (y-axis) and SARS-CoV-2 (x-axis) with high concordance (line of perfect correspondence shown; green line) as indicated by the number of genes with matching expression in the upper right quadrant of the plot (57 matches) or lower left (4 matches) compared to off-diagonal quadrants (0 and 6 genes). Significance levels of individual genes are depicted with size (COV-2) and intensity (RSV) from the two viruses.



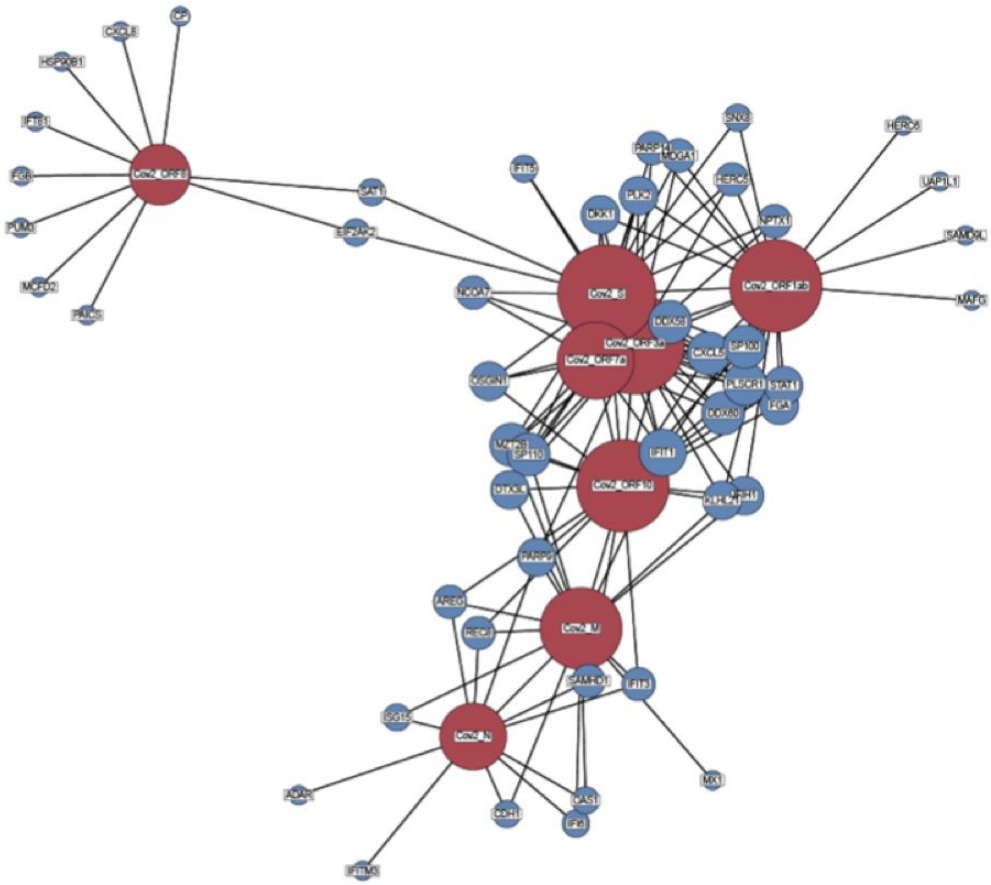
Supplemental Figure 2. Normalized gene counts are clustered by average linkage distance Pearson correlation and visualized by dendrograms. For both BALF tissue (A) and A549 cell lines (B), SARS-CoV-2 transcripts were located within a single clade (boxed in red to the left of the dendrogram) after 200 clades were created across the entirety of the hierarchical structure.



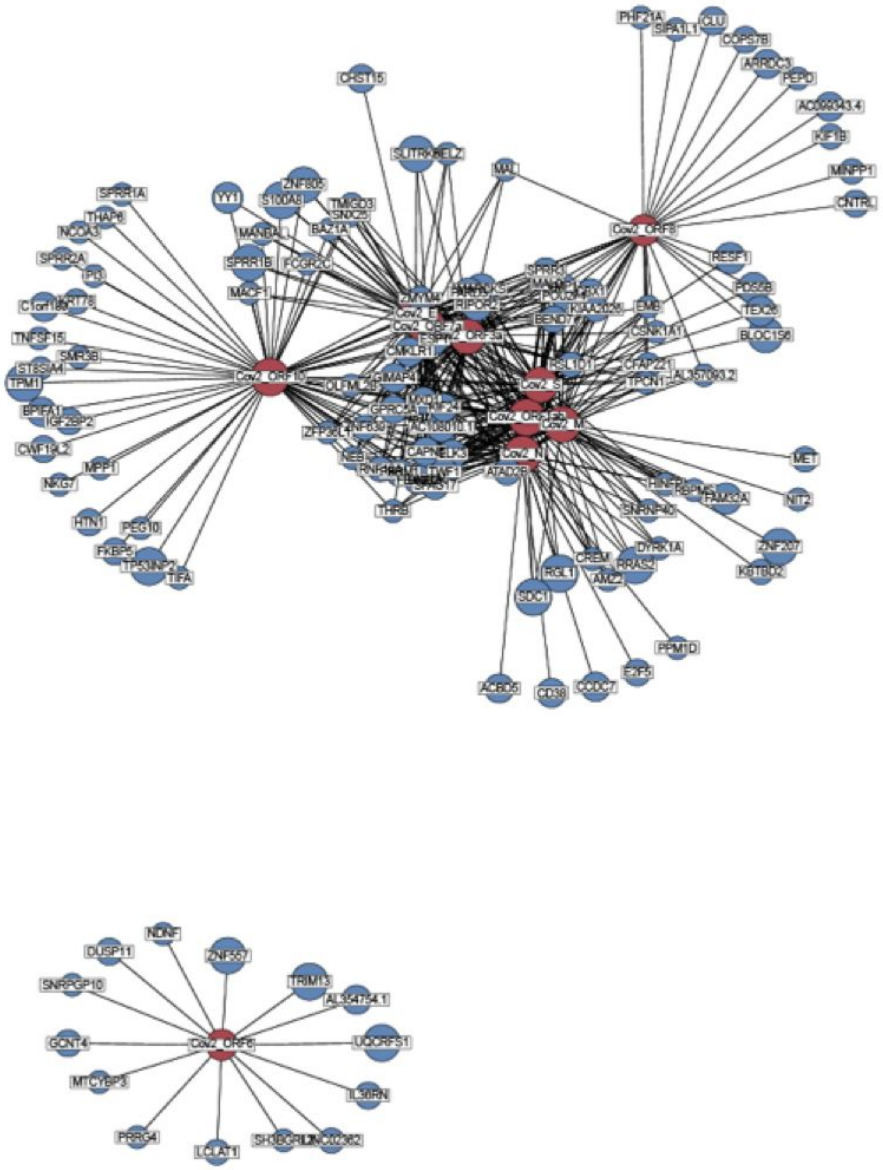
Supplemental Figure 3. The CE-Dendro gene and gProfiler pathways are visualized as a network where an edge is drawn between genes and the pathways for which they are members. A purple box is shown to highlight a particular group of genes with functional significance that were largely results not found by the DE view (CXCL5, CCL20, NFKBIA, MYC, PTGS2, IFI35, C3, DDX1, PTK2). These genes are important for cytokine and chemokine activity, antiviral response, and the innate immune system.



Supplemental Figure 4. Gene results from CE-Net in NHBE cells that were correlated with SARS-CoV-2 transcripts at $|R| > 0.98$ are visualized as a network where edges are drawn between human genes and the SARS-CoV-2 genes which they are correlated with. SARS-CoV-2 genes are shown in red and human genes in blue. Increasing node size indicates that the gene is connected to more nodes within the network. The S, ORF10, and ORF8 genes were found to be connected with the most human genes at this $|R|$ threshold, suggesting that they may be functionally important for SARS-CoV-2 in its influence on the human transcriptome.

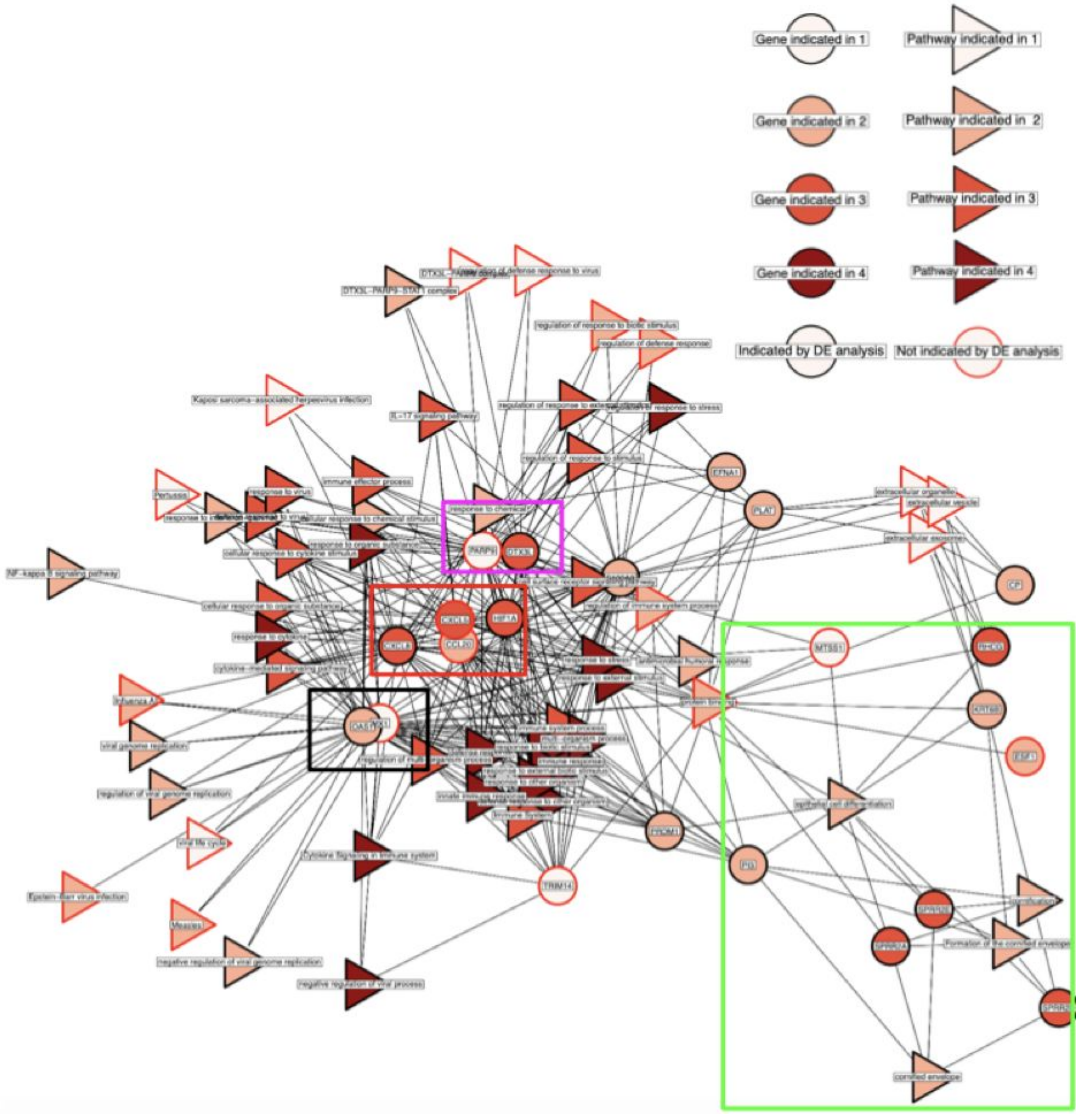


Supplemental Figure 5. Gene results from CE-Net in A549 cells that were correlated with SARS-CoV-2 transcripts at R $>$ 0.98 are visualized as a network where edges are drawn between human genes and the SARS-CoV-2 genes which they are correlated with. SARS-CoV-2 genes are shown in red and human genes in blue. Increasing node size indicates that the gene is connected to more nodes within the network. The S, ORF10, ORF8, and ORF1ab genes were found to be connected with the most human genes at this $|R|$ threshold, suggesting that they may be functionally important for SARS-CoV-2 in its influence on the human transcriptome.



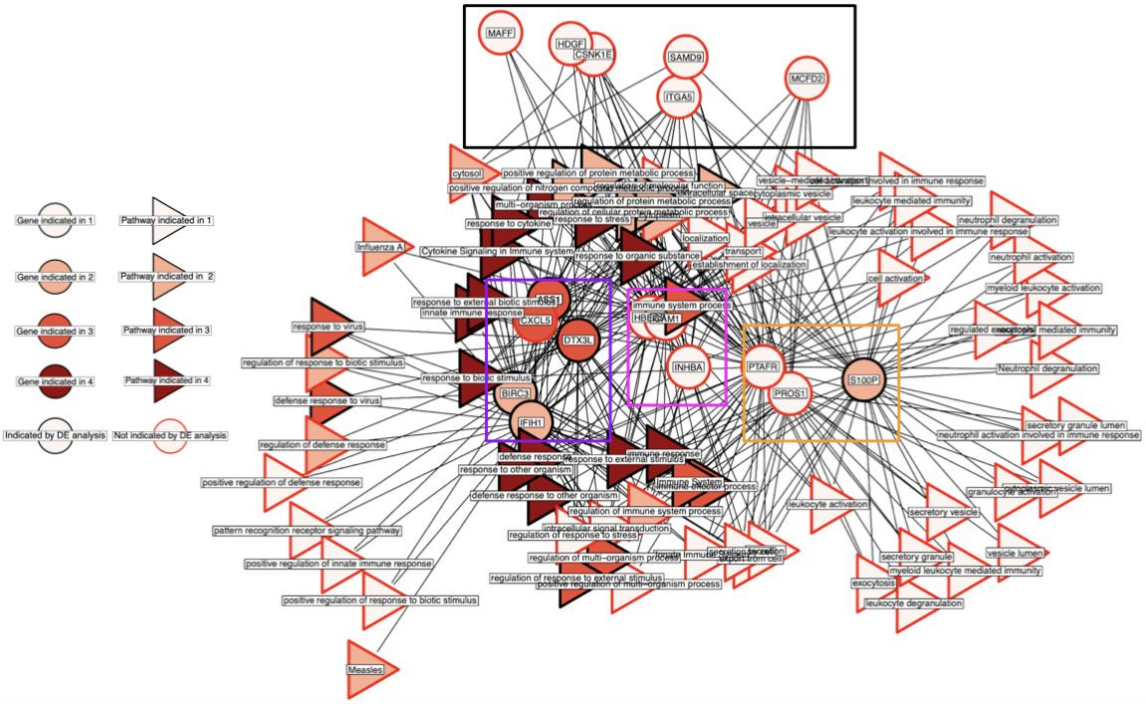
Supplemental Figure 6. Gene results from CE-Net in BALF tissue that were correlated with SARS-CoV-2 transcripts at R $>$ 0.98 are visualized as a network where edges are drawn between human genes and the SARS-CoV-2 genes which they are correlated with. SARS-CoV-2 genes are shown in red and human genes in blue. Increasing node size indicates that the gene

is connected to more nodes within the network. Most of the SARS-CoV-2 genes depicted in this network were similarly connected to many human genes. In contrast, ORF6 was not co-correlated to human genes with any other SARS-CoV-2 gene.



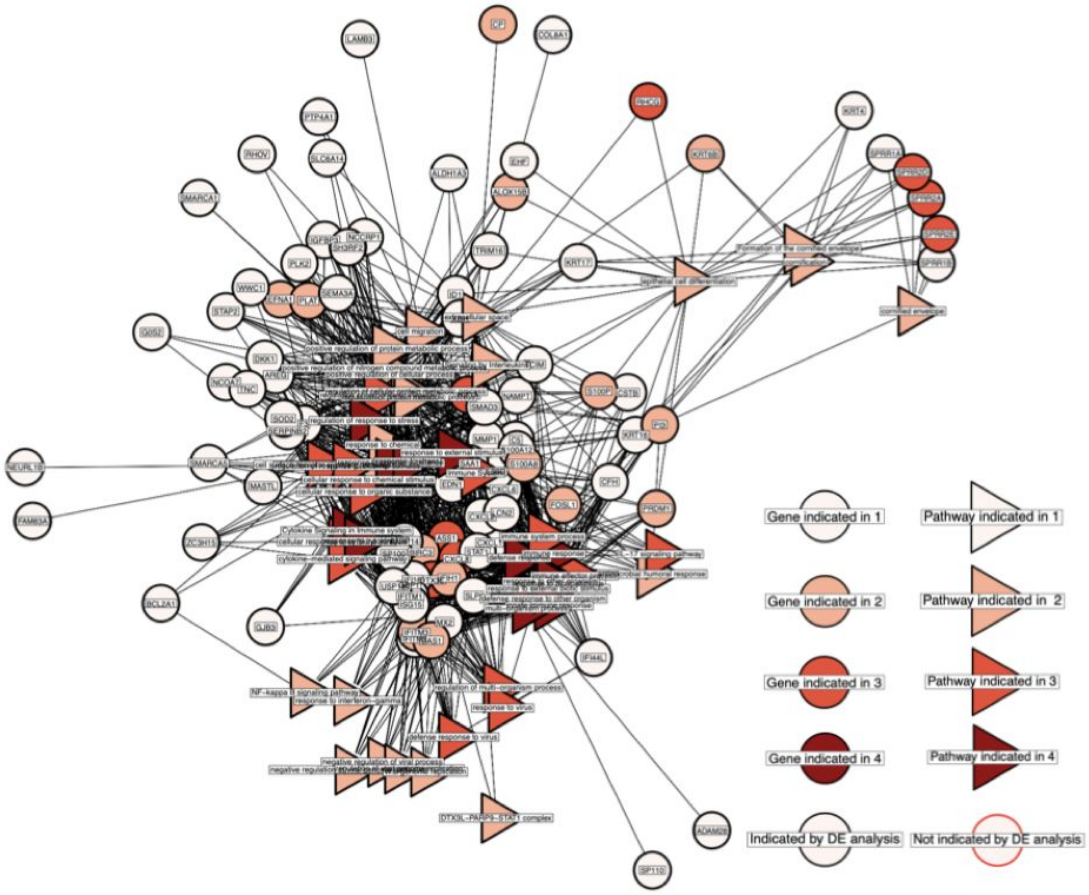
Supplemental Figure 7. The CE-Net gene and gProfiler pathways are visualized as a network

where an edge is drawn between genes and the pathways for which they are members. The black box shows genes (OAS1 and MX1) with antiviral properties. The red box shows genes indicative of chemokine and cytokine activity (CXCL5, CXCL8, CCL20, HIF1A). The pink box highlights the PARP9-DTX3L gene group. Finally, the green box shows a group of genes involved in cornification and cell death processes (MTSS1, RHCG, KRT6B, ESF1, PI3, SPRR2A, SPRR2D, SPRR2E).



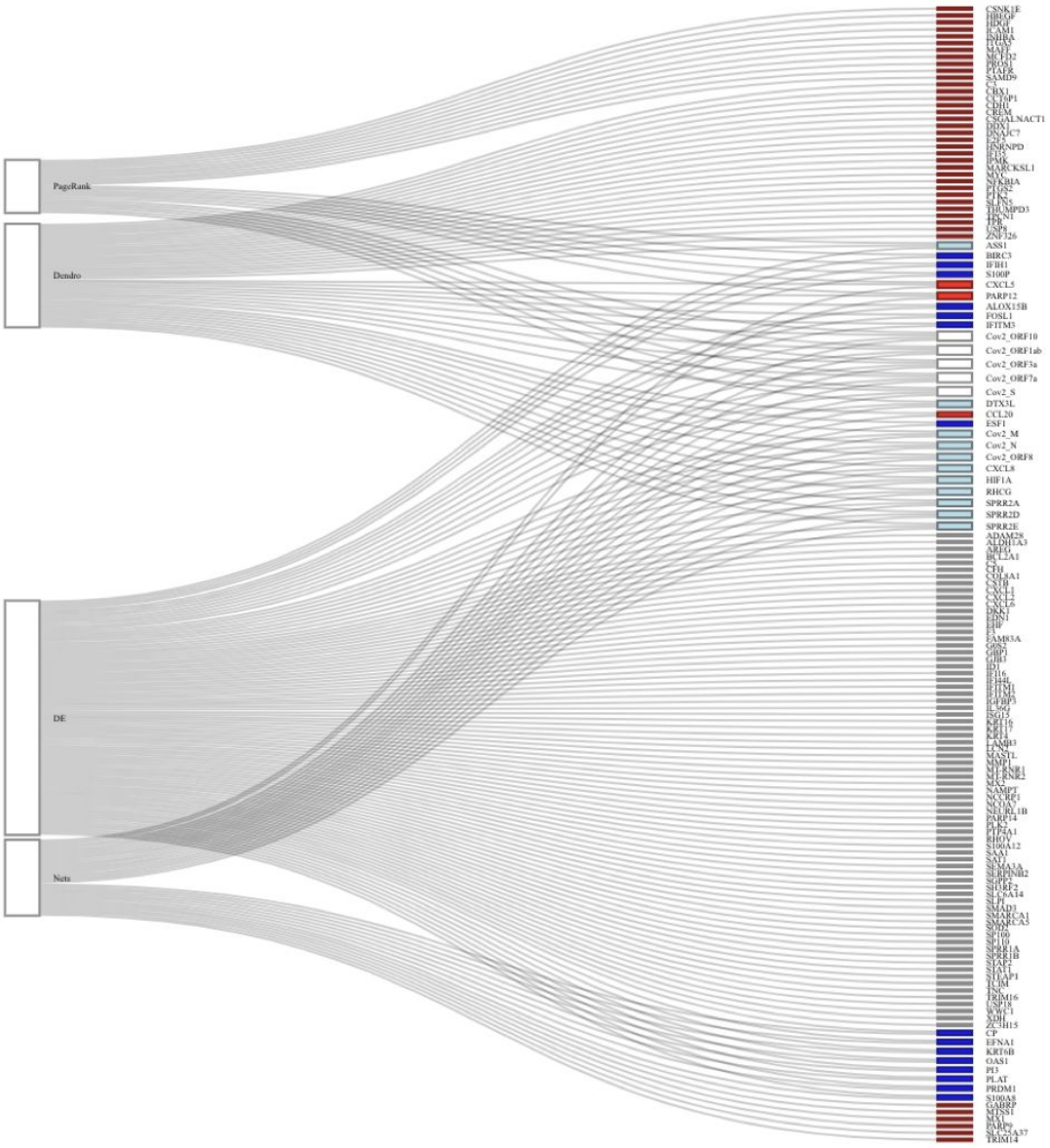
Supplemental Figure 8. The CE-PageRank gene and gProfiler pathways are visualized as a network where an edge is drawn between genes and the pathways for which they are members. The purple box highlights a gene group (ASS1, CXCL5, DTX3L, BIRC3, IFIH1) indicative of lipopolysaccharide response and cytokine, chemokine, and immune activity. The pink box

shows genes (ICAM1, HBEGF, INHBA) involved in various processes including viral entry and survival. The yellow box shows genes involved in calcium binding and secretory activity (S100P, PROS1, PTAFR). Lastly, the black box highlights gene results that were not found by DE (MAFF, HDGF, CSNK1E, ITGA5, SAMD9, MCFD2).



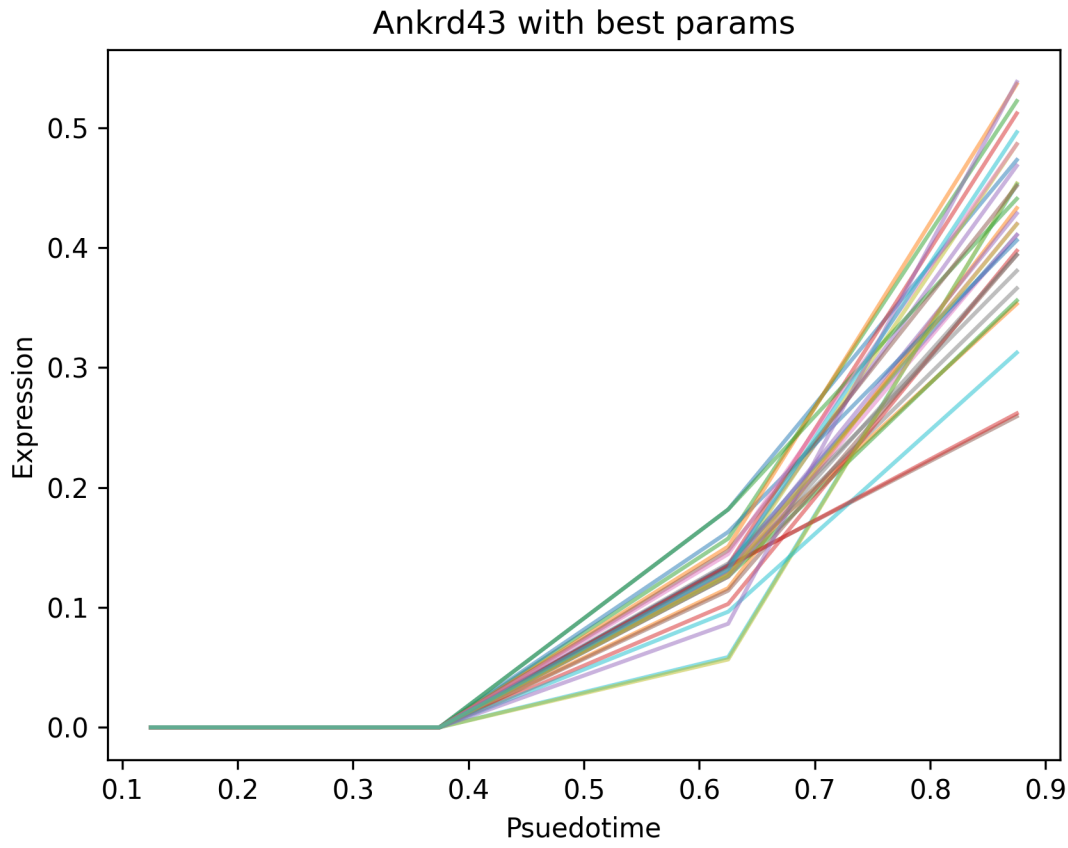
Supplemental Figure 9. The DE gene and gProfiler pathways are visualized as a network where an edge is drawn between genes and the pathways for which they are members. White nodes (indicating genes or pathways that were only found by DE) were a dominant feature (72 genes) among the gene results for DE, suggesting that DE highlights the conditional response of human

genes to SARS-CoV-2 infection.

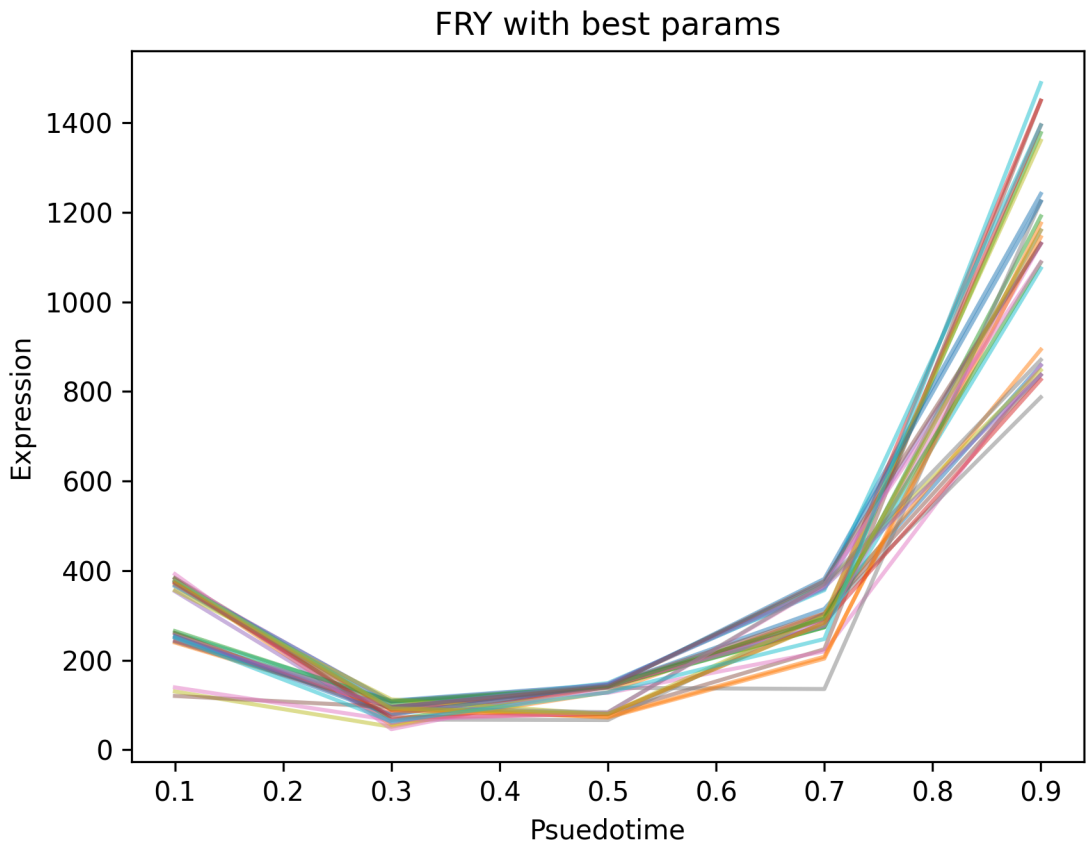


Supplemental Figure 10. The sankey diagram shows genes that were found by any of the four views and the pathway themes that they participate in. Genes are colored based on the views

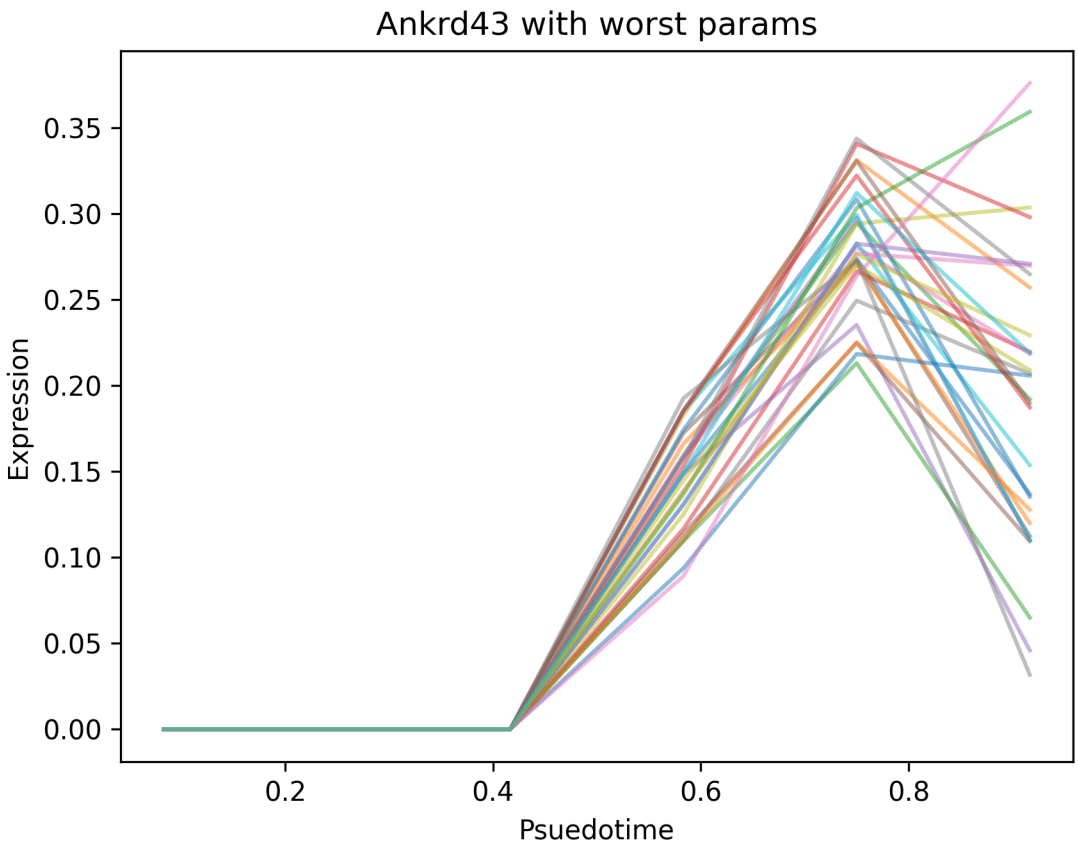
that they are found in, white, light blue, and dark blue indicate that DEA and 1, 2, or 3 other views, respectively, and red for a gene that is not found by DEA, but was by a CEA view.



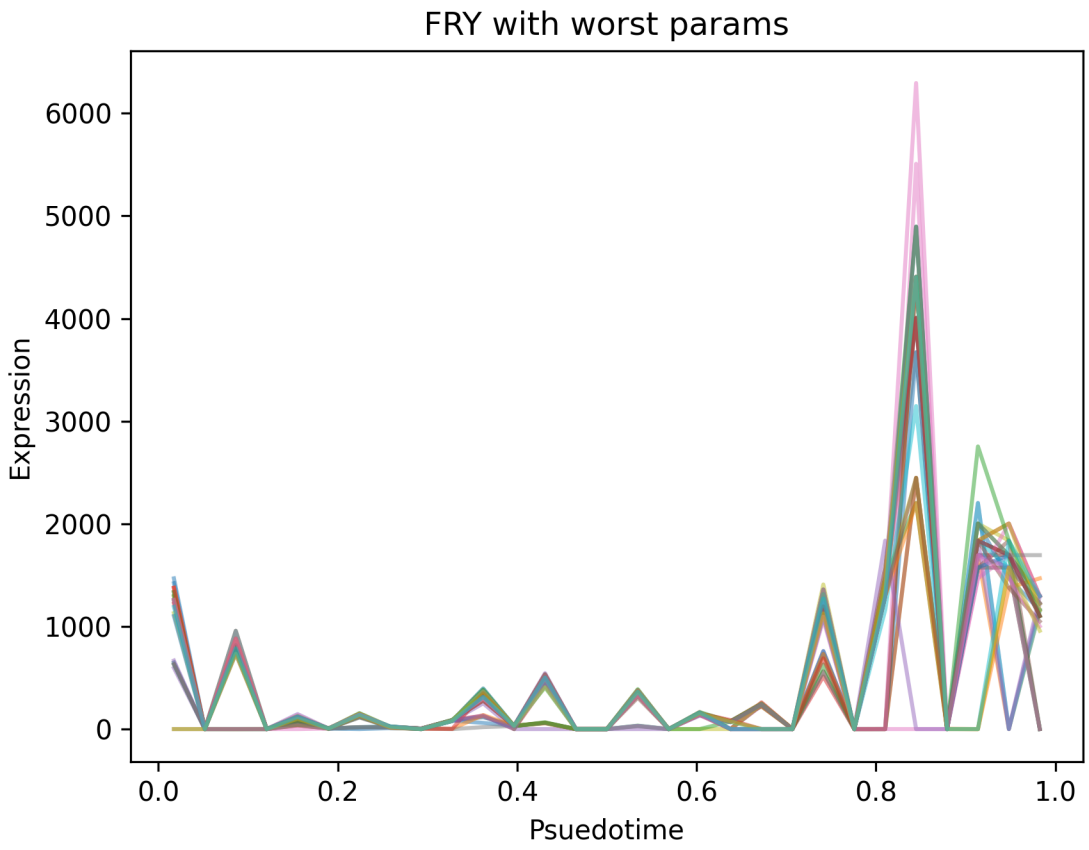
Suppl Figure 11. The 30 spline subsampled smooth data representations for Ankrd43 are plotted with expression plotted on the y-axis and pseudotime on the x-axis. This was using the best set of params (lowest AIC and CV) determined through the AIC and CV balancing depicted in Figure 5 and described in the methods.



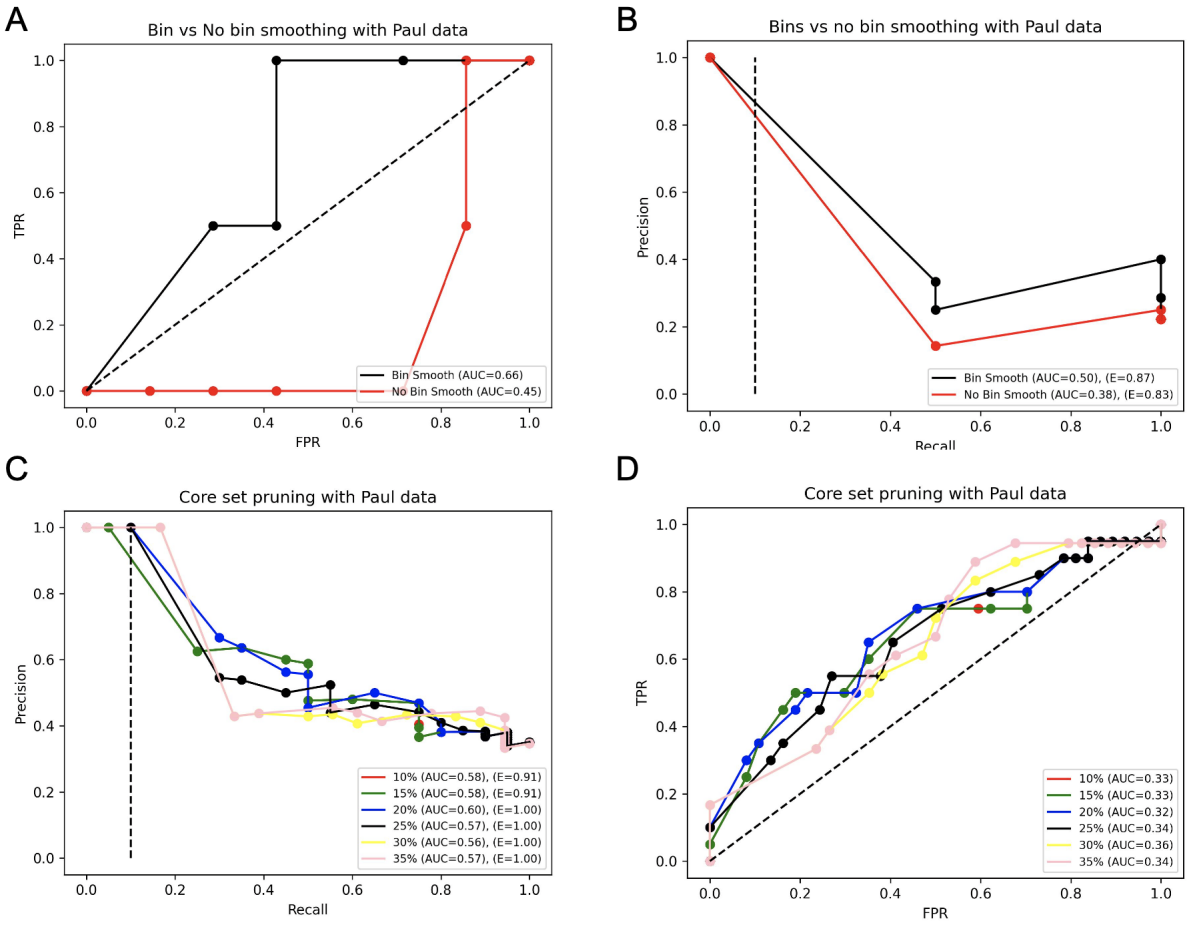
Suppl Figure 12. The 30 spline subsampled smooth data representations for FRY are plotted with expression plotted on the y-axis and pseudotime on the x-axis. This was using the best set of params (lowest AIC and CV) determined through the AIC and CV balancing depicted in Figure 5 and described in the methods.



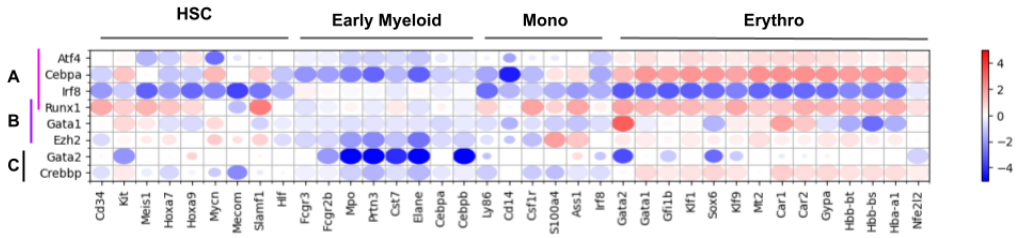
Suppl Figure 13. The 30 spline subsampled smooth data representations for Ankrd43 are plotted with expression plotted on the y-axis and pseudotime on the x-axis. This was using the worst set of params (highest AIC and CV) determined through the AIC and CV balancing depicted in Figure 5 and described in the methods.



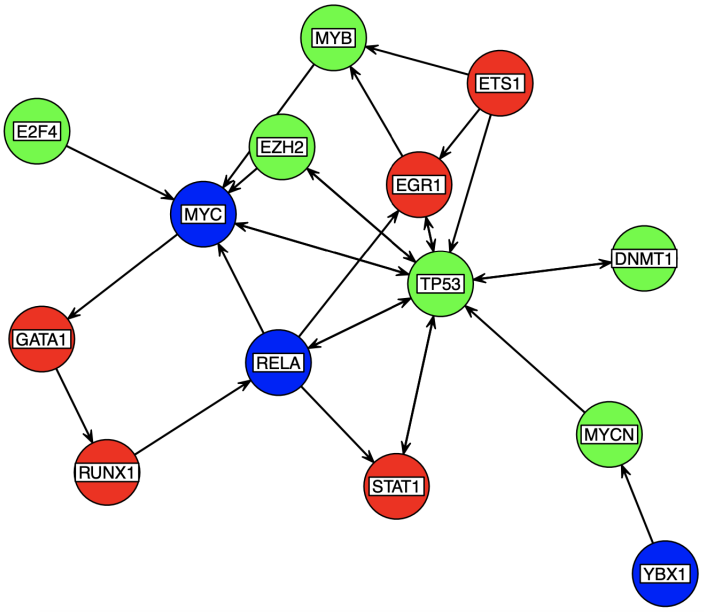
Suppl Figure 14. The 30 spline subsampled smooth data representations for FRY are plotted with expression plotted on the y-axis and pseudotime on the x-axis. This was using the worst set of params (highest AIC and CV) determined through the AIC and CV balancing depicted in Figure 5 and described in the methods.



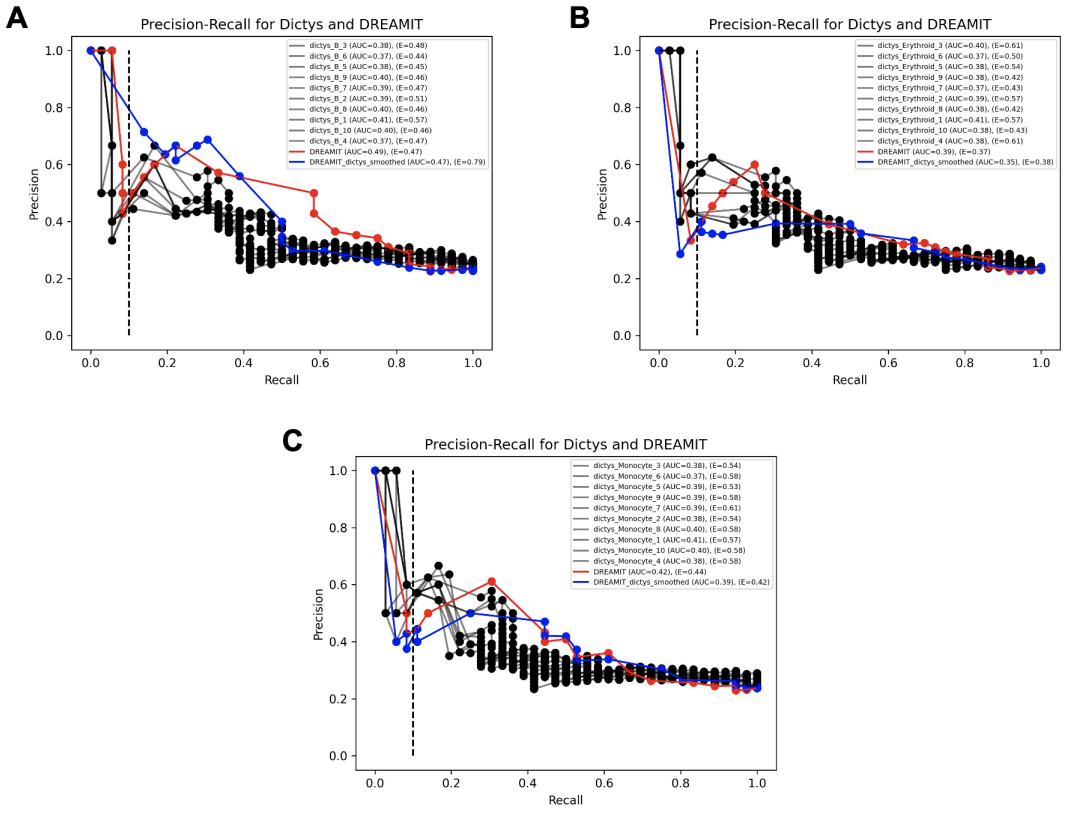
Suppl Figure 15. Using the same paul data plotted in Figure 2A the binning and pruning processes implemented by DREAMIT were compared to alternatives. Using binning for our spline smoothing (black) and spline smoothing without binning the raw data (red) is shown on a ROC plot (A) and precision-recall plot (B). Different thresholds for the pruning process were compared using precision-recall (C) and ROC (D). Early precision is plotted at a recall level of 0.1.



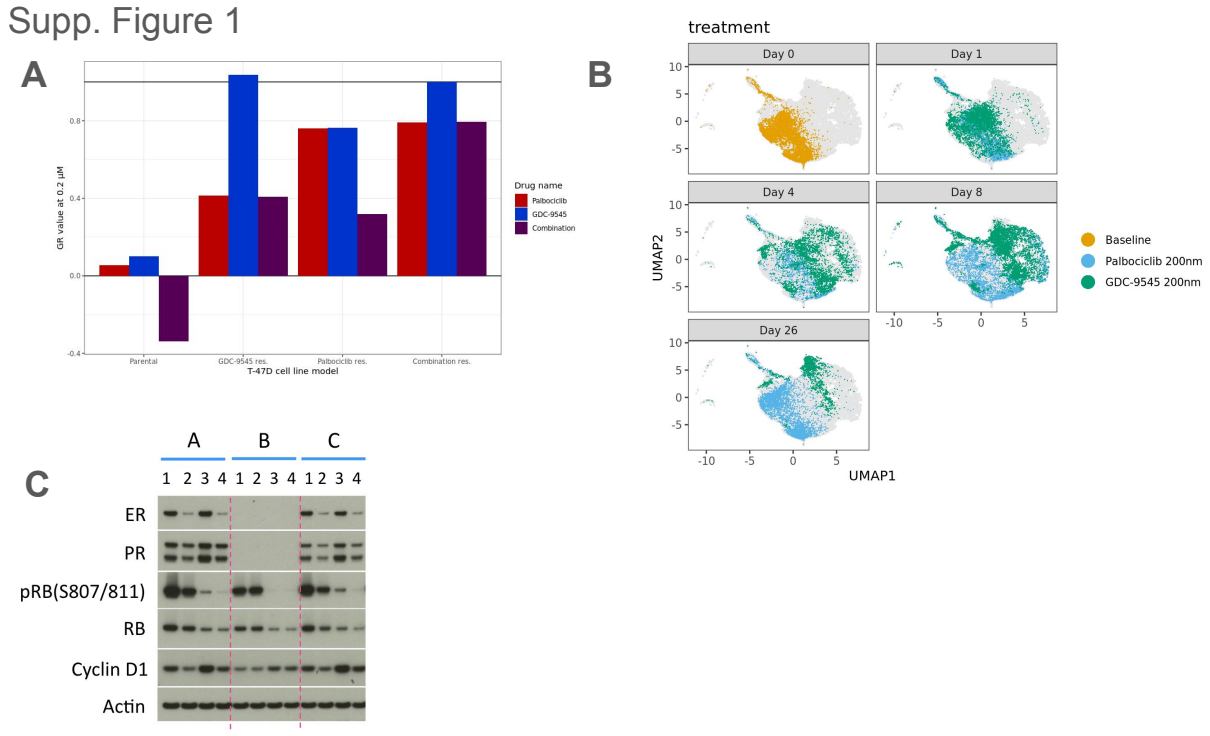
Suppl Figure 16. TFs scored by DREAMIT and the Perturb-seq experiment from Lara-Astiaso et. al. [85] were plotted against a set of markers for HSC, early myeloid, monocyte, and erythrocyte cell types. The color of each dot represents the logarithm base-2 of the fold change (a ceiling of -5 and +5 was implemented) and the size of a dot represents (1 - the adjusted P-value) reflecting the differential expression of a marker gene after a TF is knocked out. The fold change and P-values were obtained from the Lara-Astiaso et. al. publication. DREAMIT found TF set A (pink - Atf4, Cebpa, Irf8, Runx1) to be significant in the monocyte lineage, the TF set B (purple - Runx1, Gata1, Ezh2) to be significant in the erythrocyte lineage, and the TF set C (black - Gata2, Crebbp) to not be significant in either lineage.



Suppl Figure 17. The TF-TF network displays directionality and regulation from TRRUST. The color of each TF node represents the cell type; red for blood marker, blue for stem marker, and green for others. These annotations were obtained from the TF-Marker database.

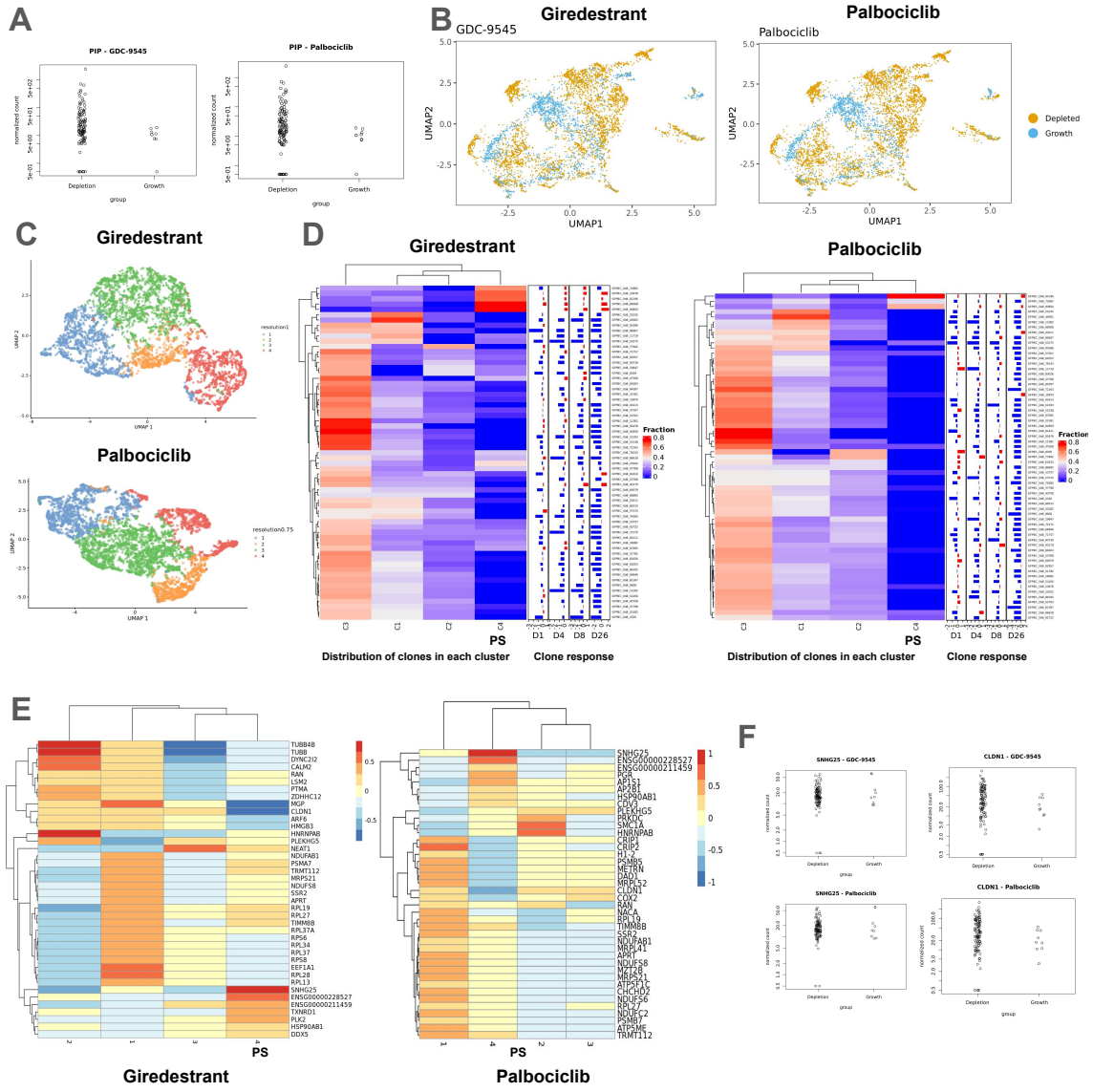


Suppl Figure 18. DREAMIT and Dictys precision-recall for tissue specificity from the TF-Marker database is plotted for three different branches for the (A) B-cell lineage, (B) Erythroid lineage, and (C) Monocyte lineage. Dictys results are plotted in black, DREAMIT results using raw data and the DREAMIT spline smoothing approach are plotted in red, and DREAMIT results using pre-processed and smoothed data from Dictys are plotted in blue. The early precision AUC is taken at 0.1 recall where the dashed line is plotted.



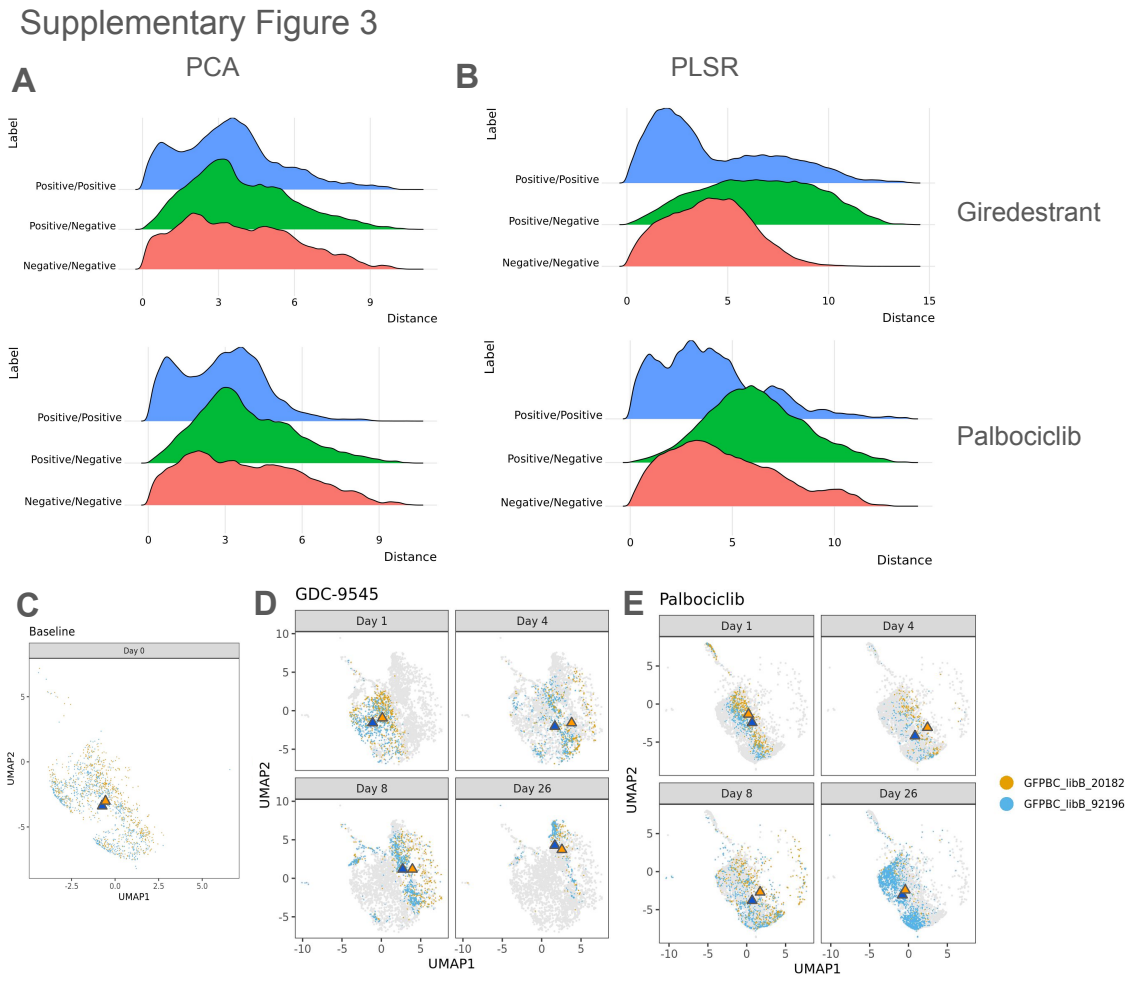
Suppl Figure 19. (A) GR value for parental T-47D, giredestrant-treated PS population, or palbociclib-treated PS population, or combination-treated PS population treated for 7 days with either 0.2µM of giredestrant, 0.2µM of palbociclib, or their combination. (B) Transcriptomic changes are observed in palbociclib treated cells (blue), but return to a baseline (orange) state, while giredestrant-treated cells (green) have remained changed. (C) Western Blot of parental T-47D [A], giredestrant-treated PS population [B], or palbociclib-treated PS population [C], treated for 24 hours with either DMSO [1], giredestrant [2], palbociclib [3], or their combination [4].

Supp. Figure 2

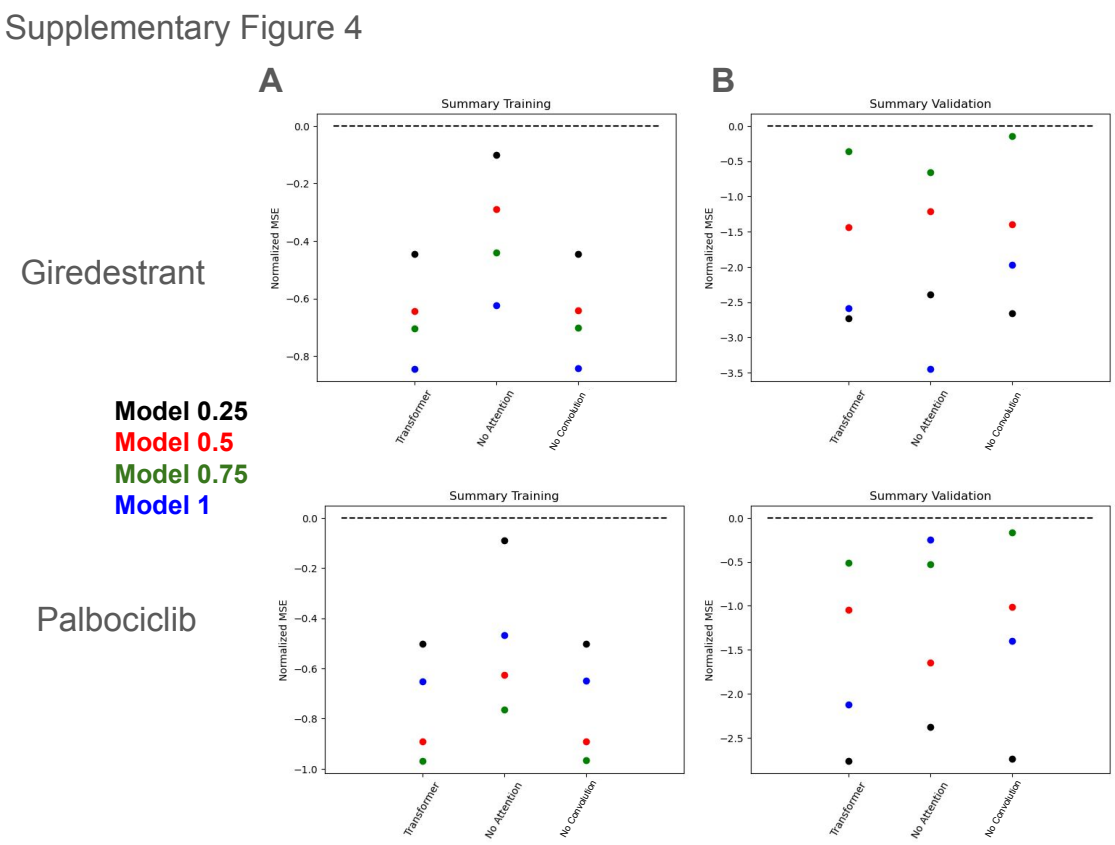


Suppl Figure 20. The PIP gene is differentially expressed in PS clones compared to NS clones in giredestrant and palbociclib (A). PS clones (blue) are not separated from NS clones (orange) when PCA dimension reduction is used to create a UMAP projection (B). Clustering results produced from the PLSR approach are plotted for giredestrant and palbociclib-treated cells (C).

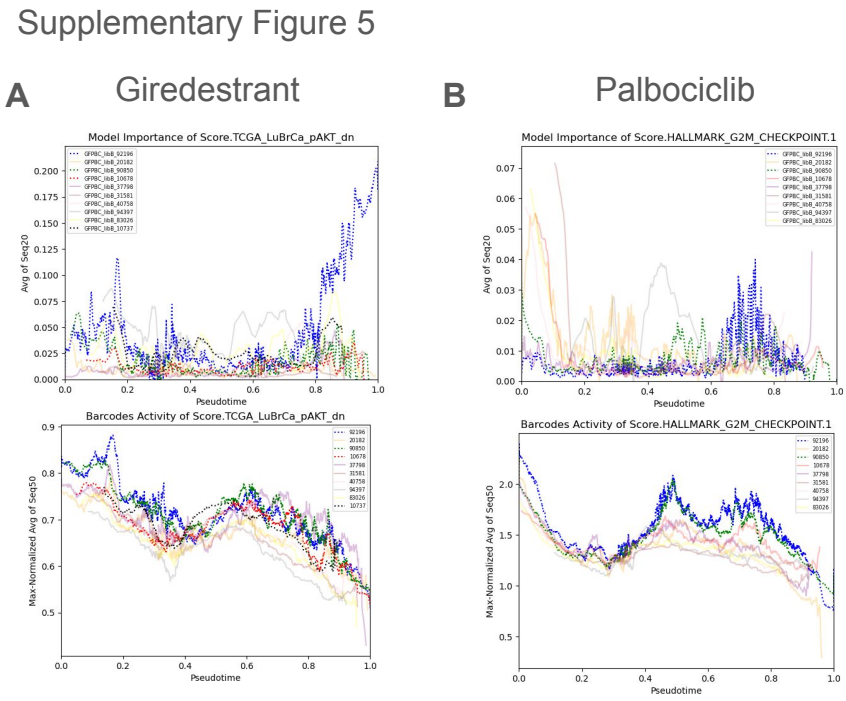
Heatmaps depict the fraction of clones (y-axis) within a given PLSR cluster (x-axis), where cluster 4 is the PS cluster in both giredestrant and palbociclib treatment. The barplots on the right side show the increase (in red) or decrease (in blue) in the clonal population at days 1, 4, 8, and 26 after treatment (D). Genes found to be differentially expressed via findMarkers are plotted relative to the expression in the PS cluster 4 (E). Genes found through the PLSR approach, SNHG25 and CLDN1, were not shown to be differentially expressed when comparing PS clones to NS clones (F).



Suppl Figure 21. Ridge plots depict the pairwise distance between two PS cells (blue), two NS cells (red), and one PS cell and one NS cell (green) when PCA dimension reduction is used (A). The distance between PS to PS cells and NS to NS cells is reduced, while PS and NS cells are more separated with PLSR (B). The transcriptomes of a NS clone (orange) and PS clone (blue) and their centroids (triangles) are plotted together pre-treatment, showing they are not separated (C). When treated with giredestrant (D) or palbociclib (E) the transcriptomes of these clones change throughout treatment, but the clone centroids remain together and are not separated by treatment.

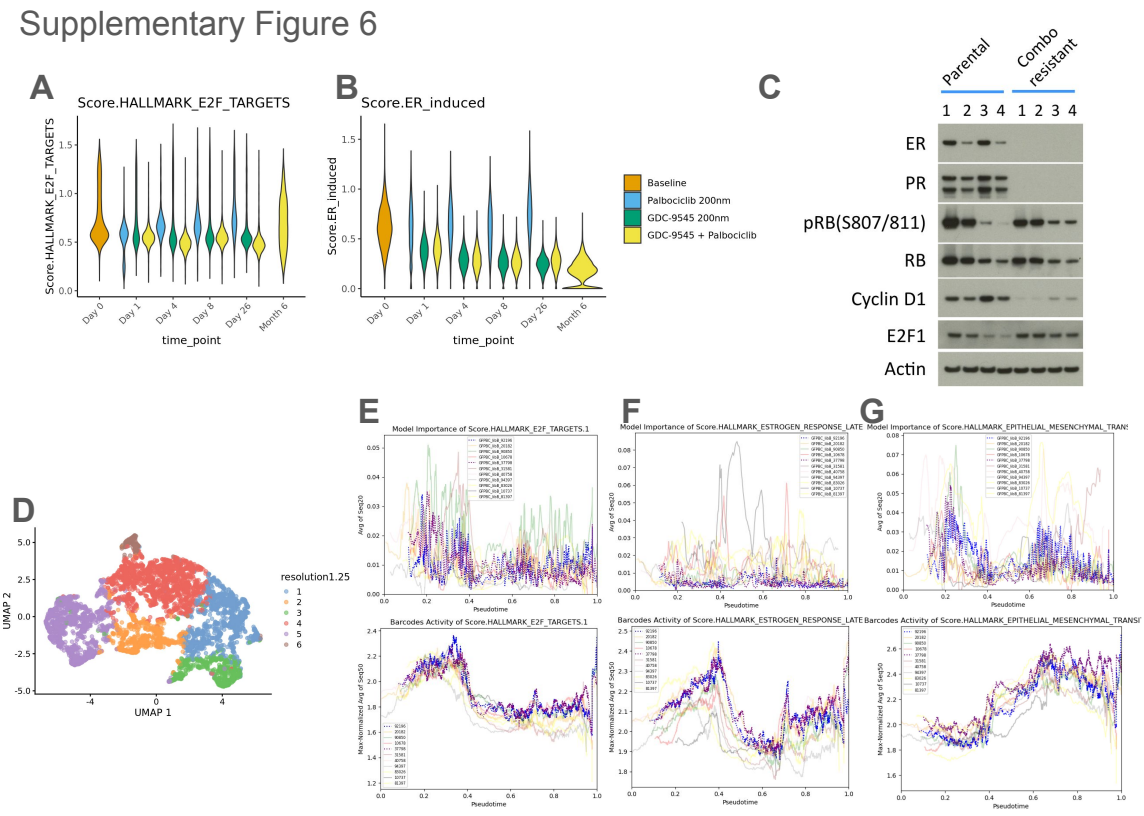


Suppl Figure 22. The model error in training is plotted relative to multivariate regression (dashed line) for modeling up till each of the four testing increments at 0.25 (black), 0.5 (red), 0.75 (green), and the full fraction of the clonal data (blue). Error is represented as the mean error for the modeled barcodes for the full TraCSED model (left), TraCSED without the attention layers (middle), and TraCSED without the convolutional layers (right) for both giredestrant and palbociclib (A). The model error is then also plotted in the same way for the validation data (B).



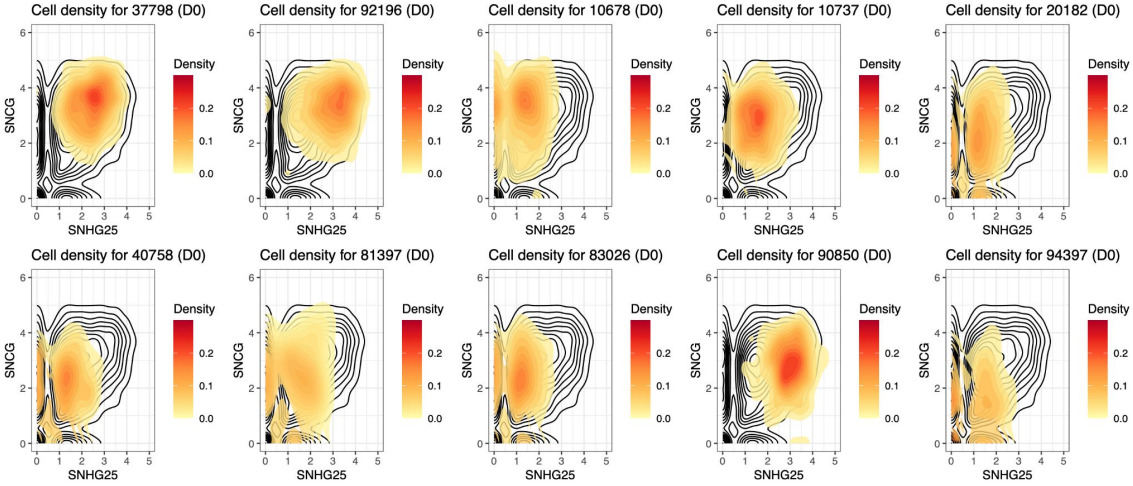
Suppl Figure 23. Smoothed model importance (averaging 20 adjacent cells) was plotted for PS clones (bold, dashed lines) and NS clones (faint, solid lines) for the AKT pathway in giredestrant treatment (upper panel). Smoothed pathway activity (averaging 50 adjacent cells) was also plotted (lower panel, A). Similarly, the smoothed model importance (upper panel) and pathway activity (lower panel) were plotted for the G2M checkpoint pathway in palbociclib

treatment (B).



Suppl Figure 24. Pathway scores for E2F targets (A) and ER induced activity (B) is plotted across pre-treatment (orange) and for the duration of treatment with palbociclib (blue), giredestrant (green), and the giredestrant-palbociclib combination therapy (yellow). The cluster results from the PLSR approach are plotted for the giredestrant and palbociclib combination treatment (C). Smoothed model importance (averaging 20 adjacent cells) was plotted for PS clones (bold, dashed lines) and NS clones (faint, solid lines)(upper panel). Smoothed pathway activity (averaging 50 adjacent cells) was also plotted (lower panel). This was done for the E2F targets pathway (C), Estrogen Response Late (D), and EMT pathway (E).

Supplementary Figure 7



Suppl Figure 25. Distribution of cells at baseline based on the expression of SNHG25 and SNCG. Black outline shows all cells; color density shows cells from each clone individually.

Bibliography

[1] Hervé Abdi. Partial least squares regression and projection on latent structure regression (pls regression). *WIREs Computational Statistics*, 2(1):97–106, January 2010.

[2] Kimberle A Agle, Rebecca A Vongsa, and Michael B Dwinell. Calcium mobilization triggered by the chemokine cxcl12 regulates migration in wounded intestinal epithelial monolayers. *Journal of Biological Chemistry*, 285(21):16066–16075, 2010.

[3] Moharram Ahmadnejad, Naser Amirizadeh, Roya Mehrasa, Ahmad Karkhah, Mahin Nikougoftar, and Arezoo Oodi. Elevated expression of DNMT1 is associated with increased expansion and proliferation of hematopoietic stem cells co-cultured with human MSCs. *Blood Res.*, 52(1):25–30, March 2017.

[4] Salman Ali, Aaron F Hirschfeld, Matthew L Mayer, Edgardo S Fortuno, Nathan Corbett, Maia Kaplan, Shirley Wang, Julia Schneiderman, Christopher D Fjell, Jin Yan, et al. Functional genetic variation in nfkbia and susceptibility to childhood asthma, bronchiolitis, and bronchopulmonary dysplasia. *The Journal of Immunology*, 190(8):3949–3958, 2013.

[5] Fernando Almazán, Carmen Galán, and Luis Enjuanes. The nucleoprotein is required for efficient coronavirus genome replication. *Journal of virology*, 78(22):12683–12688, 2004.

[6] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, April 2010.

[7] Pau Badia-I-Mompel, Lorna Wessels, Sophia Müller-Dott, Rémi Trimbour, Ricardo O Ramirez Flores, Ricard Argelaguet, and Julio Saez-Rodriguez. Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.*, 24(11):739–754, November 2023.

[8] Fatemeh Behjati Ardakani, Kathrin Kattler, Tobias Heinen, Florian Schmidt, David Feuerborn, Gilles Gasparoni, Konstantin Lepikhov, Patrick Nell, Jan Hengstler, Jörn Walter, et al. Prediction of single-cell gene expression for transcription factor analysis. *GigaScience*, 9(11):giaa113, 2020.

[9] Fatemeh Behjati Ardakani, Kathrin Kattler, Tobias Heinen, Florian Schmidt, David Feuerborn, Gilles Gasparoni, Konstantin Lepikhov, Patrick Nell, Jan Hengstler, Jörn Walter, and Marcel H Schulz. Prediction of single-cell gene expression for transcription factor analysis. *Gigascience*, 9(11), October 2020.

[10] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[11] Vance W Berger and YanYan Zhou. Kolmogorov–smirnov test: Overview. *Wiley statsref: Statistics reference online*, 2014.

[12] Vance W Berger and Yanyan Zhou. *Kolmogorov-Smirnov Test: Overview*. John Wiley & Sons, Ltd, Chichester, UK, September 2014.

[13] Hyo-eun C Bhang, David A Ruddy, Viveksagar Krishnamurthy Radhakrishna, Justina X Caushi, Rui Zhao, Matthew M Hims, Angad P Singh, Iris Kao, Daniel Rakiec, Pamela Shaw, Marissa Balak, Alina Raza, Elizabeth Ackley, Nicholas Keen, Michael R Schlabach, Michael Palmer, Rebecca J Leary, Derek Y Chiang, William R Sellers, Franziska Michor, Vesselina G Cooke, Joshua M Korn, and Frank Stegmeier. Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nature Medicine*, 21(5):440–448, April 2015.

[14] Alessio Biagioni, Shima Tavakol, Nooshin Ahmadirad, Masoumeh Zahmatkeshan, Lucia Magnelli, Ali Mandegary, Hojjat Samareh Fekri, Malek Hossein Asadi, Reza Mohammadinejad, and Kwang Seok Ahn. Small nucleolar ¡scp¿rna¡/scp¿ host genes promoting epithelial–mesenchymal transition lead cancer progression and metastasis. *IUBMB Life*, 73(6):825–842, May 2021.

[15] Brent A. Biddy, Wenjun Kong, Kenji Kamimoto, Chuner Guo, Sarah E. Waye, Tao Sun, and Samantha A. Morris. Single-cell mapping of lineage and identity in direct reprogramming. *Nature*, 564(7735):219–224, December 2018.

[16] Daniel Blanco-Melo, Benjamin E Nilsson-Payant, Wen-Chun Liu, Skyler Uhl, Daisy

Hoagland, Rasmus Møller, Tristan X Jordan, Kohei Oishi, Maryline Panis, David Sachs, et al. Imbalanced host response to sars-cov-2 drives development of covid-19. *Cell*, 181(5):1036–1045, 2020.

[17] Paolo Boldi, Massimo Santini, and Sebastiano Vigna. Pagerank as a function of the damping factor. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, page 557–566, New York, NY, USA, 2005. Association for Computing Machinery.

[18] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[19] Michael Buchert, Charbel Darido, Ebba Lagerqvist, Anna Sedello, Chantal Cazevieille, Frank Buchholz, Jean–Francois Bourgaux, Julie Pannequin, Dominique Joubert, and Frédéric Hollande. The symplekin/zonab complex inhibits intestinal cell differentiation by the repression of aml1/runx1. *Gastroenterology*, 137(1):156–164.e3, July 2009.

[20] Carter T Butts. network: a package for managing relational data in r. *Journal of statistical software*, 24:1–36, 2008.

[21] Robrecht Cannoodt, Wouter Saelens, Louise Deconinck, and Yvan Saeys. dyngen: a multi-modal simulator for spearheading new single-cell omics analyses. *BioRxiv*, 2020.

[22] Margherita T Cantorna and Juhi Arora. Two lineages of immune cells that differentially express the vitamin D receptor. *J. Steroid Biochem. Mol. Biol.*, 228(106253):106253, April 2023.

[23] Margherita T Cantorna and Juhi Arora. Two lineages of immune cells that differentially express the vitamin D receptor. *J. Steroid Biochem. Mol. Biol.*, 228(106253):106253, April 2023.

[24] Zhuoxiao Cao, Xinghui Sun, Basak Icli, Akm Khyrul Wara, and Mark W Feinberg. Role of krüppel-like factors in leukocyte development, function, and disease. *Blood*, 116(22):4404–4414, November 2010.

[25] Rachel Cavill, Jos Kleinjans, and Jacob-Jan Briede. Dtw4omics: comparing patterns in biological time series. *PLoS one*, 8(8):e71823, 2013.

[26] Rachel Cavill, Jos Kleinjans, and Jacob-Jan Briedé. DTW4Omics: comparing patterns in biological time series. *PLoS One*, 8(8):e71823, August 2013.

[27] Michaël Cerezo, Xiaoxiao Sun, and Shensi Shen. Editorial: Non-genetic adaptive drug resistance in cancer. *Frontiers in Cell and Developmental Biology*, 10, November 2022.

[28] Matthew T. Chang, Frances Shanahan, Thi Thu Thao Nguyen, Steven T. Staben, Lewis Gazzard, Sayumi Yamazoe, Ingrid E. Wertz, Robert Piskol, Yeqing Angela Yang, Zora Modrusan, Benjamin Haley, Marie Evangelista, Shiva Malek, Scott A. Foster, and Xin Ye. Identifying transcriptional programs underlying cancer drug response with trace-seq. *Nature Biotechnology*, 40(1):86–93, September 2021.

[29] Geng Chen, Baitang Ning, and Tieliu Shi. Single-cell rna-seq technologies and related computational data analysis. *Frontiers in genetics*, page 317, 2019.

[30] Geng Chen, Baitang Ning, and Tieliu Shi. Single-cell RNA-seq technologies and related computational data analysis. *Front. Genet.*, 10:317, April 2019.

[31] Hui Chen, Hudan Liu, and Guoliang Qing. Targeting oncogenic myc as a strategy for cancer treatment. *Signal transduction and targeted therapy*, 3(1):1–7, 2018.

[32] Huidong Chen, Luca Albergante, Jonathan Y Hsu, Caleb A Lareau, Giosuè Lo Bosco, Jihong Guan, Shuigeng Zhou, Alexander N Gorban, Daniel E Bauer, Martin J Aryee, David M Langenau, Andrei Zinovyev, Jason D Buenrostro, Guo-Cheng Yuan, and Luca Pinello. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat. Commun.*, 10(1):1903, April 2019.

[33] Nanshan Chen, Min Zhou, Xuan Dong, Jieming Qu, Fengyun Gong, Yang Han, Yang Qiu, Jingli Wang, Ying Liu, Yuan Wei, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. *The lancet*, 395(10223):507–513, 2020.

[34] Ji Yeh Choi and Heungsun Hwang. Bayesian generalized structured component analysis. *Br. J. Math. Stat. Psychol.*, 73(2):347–373, May 2020.

[35] Kwangbom Choi, Yang Chen, Daniel A Skelly, and Gary A Churchill. Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. *Genome biology*, 21(1):1–16, 2020.

[36] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L

Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):1–19, 2016.

[37] Ibiayi Dagogo-Jack and Alice T. Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15(2):81–94, November 2017.

[38] Priya Nimish Deo and Revati Deshmukh. Pathophysiology of keratinization. *Journal of oral and maxillofacial pathology: JOMFP*, 22(1):86, 2018.

[39] Atul Deshpande, Li-Fang Chu, Ron Stewart, and Anthony Gitter. Network inference with granger causality ensembles on single-cell transcriptomics. *Cell reports*, 38(6):110333, 2022.

[40] Atul Deshpande, Li-Fang Chu, Ron Stewart, and Anthony Gitter. Network inference with granger causality ensembles on single-cell transcriptomics. *Cell Rep.*, 38(6):110333, February 2022.

[41] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[42] Benjamin L. Emert, Christopher J. Cote, Eduardo A. Torre, Ian P. Dardani, Connie L. Jiang, Naveen Jain, Sydney M. Shaffer, and Arjun Raj. Variability within rare cell states enables multiple paths toward drug resistance. *Nature Biotechnology*, 39(7):865–876, February 2021.

[43] Janan T Eppig. Mouse genome informatics (MGI) resource: Genetic, genomic, and biological knowledgebase for the laboratory mouse. *ILAR J.*, 58(1):17–41, July 2017.

[44] Janan T Eppig, Cynthia L Smith, Judith A Blake, Martin Ringwald, James A Kadin, Joel E Richardson, and Carol J Bult. Mouse genome informatics (mgi): resources for mining mouse genetic, genomic, and biological data in support of primary and translational research. In *Systems Genetics*, pages 47–73. Springer, 2017.

[45] Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, and Zhang YZ. A new coronavirus associated with human respiratory disease in China. *Nature. 2020 Mar;579(7798):265-269. doi:*, 10., 2020.

[46] Eva M Fast, Audrey Sporrij, Margot Manning, Edroaldo Lummertz Rocha, Song Yang, Yi Zhou, Jimin Guo, Ninib Baryawno, Nikolaos Barkas, David Scadden, Fernando Camargo, and Leonard I Zon. External signals regulate continuous transcriptional states in hematopoietic stem cells. *Elife*, 10, December 2021.

[47] Jason D Fernandes, Angie S Hinrichs, Hiram Clawson, Jairo Navarro Gonzalez, Brian T Lee, Luis R Nassar, Brian J Raney, Kate R Rosenbloom, Santrupti Nerli, Arjun A Rao, et al. The ucsc sars-cov-2 genome browser. *Nature Genetics*, 52(10):991–998, 2020.

[48] Laura Ferreira and David B Hitchcock. A comparison of hierarchical methods for clustering functional data. *Communications in Statistics-Simulation and Computation*, 38(9):1925–1949, 2009.

[49] Christian Fougner, Helga Bergholtz, Jens Henrik Norum, and Therese Sørlie. Re-definition of claudin-low as a breast cancer phenotype. *Nature Communications*, 11(1), April 2020.

[50] Kirsten L. Frieda, James M. Linton, Sahand Hormoz, Joonhyuk Choi, Ke-Huan K. Chow, Zakary S. Singer, Mark W. Budde, Michael B. Elowitz, and Long Cai. Synthetic recording and in situ readout of lineage information in single cells. *Nature*, 541(7635):107–111, November 2016.

[51] Clayton E Friedman, Quan Nguyen, Samuel W Lukowski, Abbigail Helfer, Han Sheng Chiu, Jason Miklas, Shiri Levy, Shengbao Suo, Jing-Dong Jackie Han, Pierre Osteil, et al. Single-cell transcriptomic analysis of cardiac differentiation from human pscs reveals hopx-dependent cardiomyocyte maturation. *Cell stem cell*, 23(4):586–598, 2018.

[52] Clayton E Friedman, Quan Nguyen, Samuel W Lukowski, Abbigail Helfer, Han Sheng Chiu, Jason Miklas, Shiri Levy, Shengbao Suo, Jing-Dong Jackie Han, Pierre Osteil, Guangdun Peng, Naihe Jing, Greg J Baillie, Anne Senabouth, Angelika N Christ, Timothy J Bruxner, Charles E Murry, Emily S Wong, Jun Ding, Yuliang Wang, James Hudson, Hannele Ruohola-Baker, Ziv Bar-Joseph, Patrick P L Tam, Joseph E Powell, and Nathan J Palpant. Single-cell transcriptomic analysis of cardiac differentiation from human PSCs reveals HOPX-dependent cardiomyocyte maturation. *Cell Stem Cell*, 23(4):586–598.e8, October 2018.

[53] Julien J. Ghislain and Eleanor N. Fish. Application of genomic dna affinity chromatog-

raphy identifies multiple interferon--regulated stat2 complexes. *Journal of Biological Chemistry*, 271(21):12408–12413, May 1996.

[54] Yogesh Goyal, Gianna T. Busch, Maalavika Pillai, Jingxin Li, Ryan H. Boe, Emanuelle I. Grody, Manoj Chelvanambi, Ian P. Dardani, Benjamin Emert, Nicholas Bodkin, Jonas Braun, Dylan Fingerman, Amanpreet Kaur, Naveen Jain, Pavithran T. Ravindran, Ian A. Mellis, Karun Kiani, Gretchen M. Alicea, Mitchell E. Fane, Syeda Subia Ahmed, Haiyin Li, Yeqing Chen, Cedric Chai, Jessica Kaster, Russell G. Witt, Rossana Lazcano, Davis R. Ingram, Sarah B. Johnson, Khalida Wani, Margaret C. Dunagin, Alexander J. Lazar, Ashani T. Weeraratna, Jennifer A. Wargo, Meenhard Herlyn, and Arjun Raj. Diverse clonal fates emerge upon drug treatment of homogeneous cancer cells. *Nature*, 620(7974):651–659, July 2023.

[55] Rachel L Graham and Ralph S Baric. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. *Journal of virology*, 84(7):3134–3146, 2010.

[56] Lisa E Gralinski, Timothy P Sheahan, Thomas E Morrison, Vineet D Menachery, Kara Jensen, Sarah R Leist, Alan Whitmore, Mark T Heise, and Ralph S Baric. Complement activation contributes to severe acute respiratory syndrome coronavirus pathogenesis. *MBio*, 9(5):e01753–18, 2018.

[57] Matthew E Grunewald, Yating Chen, Chad Kuny, Takashi Maejima, Robert Lease, Dana Ferraris, Masanori Aikawa, Christopher S Sullivan, Stanley Perlman, and Anthony R Fehr. The coronavirus macrodomain is required to prevent parp-mediated inhibition of

virus replication and enhancement of ifn expression. *PLoS pathogens*, 15(5):e1007756, 2019.

[58] Wei-jie Guan, Zheng-yi Ni, Yu Hu, Wen-hua Liang, Chun-quan Ou, Jian-xing He, Lei Liu, Hong Shan, Chun-liang Lei, David SC Hui, et al. Clinical characteristics of coronavirus disease 2019 in china. *New England journal of medicine*, 382(18):1708–1720, 2020.

[59] Jingtao Guo, Xichen Nie, Maria Giebler, Hana Mlcochova, Yueqi Wang, Edward J Grow, Robin Kim, Melissa Tharmalingam, Gabriele Matilionyte, Cecilia Lindskog, et al. The dynamic transcriptional cell atlas of testis development during human puberty. *Cell stem cell*, 26(2):262–276, 2020.

[60] Piyush B. Gupta, Christine M. Fillmore, Guozhi Jiang, Sagi D. Shapira, Kai Tao, Charlotte Kuperwasser, and Eric S. Lander. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell*, 146(4):633–644, August 2011.

[61] Catherine Gutierrez, Aziz M. Al'Khafaji, Eric Brenner, Kaitlyn E. Johnson, Satyen H. Gohil, Ziao Lin, Binyamin A. Knisbacher, Russell E. Durrett, Shuqiang Li, Salma Parvin, Anat Biran, Wandi Zhang, Laura Rassenti, Thomas J. Kipps, Kenneth J. Livak, Donna Neuberg, Anthony Letai, Gad Getz, Catherine J. Wu, and Amy Brock. Multifunctional barcoding with clonmapper enables high-resolution study of clonal dynamics during tumor evolution and treatment. *Nature Cancer*, 2(7):758–772, July 2021.

[62] Marc Hafner, Mario Niepel, Mirra Chung, and Peter K Sorger. Growth rate inhibition

metrics correct for confounders in measuring sensitivity to cancer drugs. *Nature Methods*, 13(6):521–527, May 2016.

[63] Heonjong Han, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, Chan Yeong Kim, Muyoung Lee, Eunbeen Kim, et al. Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic acids research*, 46(D1):D380–D386, 2018.

[64] Heonjong Han, Jae-Won Cho, Sangyoung Lee, Ayoung Yun, Hyojin Kim, Dasom Bae, Sunmo Yang, Chan Yeong Kim, Muyoung Lee, Eunbeen Kim, Sungho Lee, Byunghee Kang, Dabin Jeong, Yaeji Kim, Hyeon-Nae Jeon, Haein Jung, Sunhwee Nam, Michael Chung, Jong-Hoon Kim, and Insuk Lee. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.*, 46(D1):D380–D386, January 2018.

[65] Yuhua He, Shuifang Xu, Yi Qi, Jinfang Tian, and Fengying Xu. Long noncoding rna snhg25 promotes the malignancy of endometrial cancer by sponging microrna-497-5p and increasing fasn expression. *Journal of Ovarian Research*, 14(1), November 2021.

[66] Z. He, A. Maynard, A. Jain, T. Gerber, R. Petri, H. C. Lin, M. Santel, K. Ly, J. S. Dupré, L. Sidow, and Sanchis Calleja. Lineage recording in human cerebral organoids. *Nature methods*, 19(1):90–99, 2022.

[67] Ryan M Hekman, Adam J Hume, Raghuveera Kumar Goel, Kristine M Abo, Jessie Huang, Benjamin C Blum, Rhiannon B Werder, Ellen L Suder, Indranil Paul, Sadhna

Phanse, et al. Actionable cytopathogenic host responses of human alveolar type 2 cells to sars-cov-2. *Molecular cell*, 80(6):1104–1122, 2020.

[68] Jason I Herschkowitz, Karl Simin, Victor J Weigman, Igor Mikaelian, Jerry Usary, Zhiyuan Hu, Karen E Rasmussen, Laundette P Jones, Shahin Assefnia, Subhashini Chandrasekharan, Michael G Backlund, Yuzhi Yin, Andrey I Khramtsov, Roy Bastein, John Quackenbush, Robert I Glazer, Powel H Brown, Jeffrey E Green, Levy Kopelovich, Priscilla A Furth, Juan P Palazzo, Olufunmilayo I Olopade, Philip S Bernard, Gary A Churchill, Terry Van Dyke, and Charles M Perou. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biology*, 8(5):R76, 2007.

[69] Laurie Herviou, Giacomo Cavalli, Guillaume Cartron, Bernard Klein, and Jérôme Moreaux. EZH2 in normal hematopoiesis and hematological malignancies. *Oncotarget*, 7(3):2284–2296, January 2016.

[70] Christian H Holland, Jovan Tanevski, Javier Perales-Patón, Jan Gleixner, Manu P Kumar, Elisabetta Mereu, Brian A Joughin, Oliver Stegle, Douglas A Lauffenburger, Holger Heyn, et al. Robustness and applicability of transcription factor and pathway analysis tools on single-cell rna-seq data. *Genome biology*, 21(1):1–19, 2020.

[71] Christian H Holland, Jovan Tanevski, Javier Perales-Patón, Jan Gleixner, Manu P Kumar, Elisabetta Mereu, Brian A Joughin, Oliver Stegle, Douglas A Lauffenburger, Holger Heyn, Bence Szalai, and Julio Saez-Rodriguez. Robustness and applicability of tran-

scription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol.*, 21(1):36, February 2020.

[72] Jenny Hsu, Julia Arand, Andrea Chaikovsky, Nancie A Mooney, Janos Demeter, Caileen M Brison, Romane Oliverio, Hannes Vogel, Seth M Rubin, Peter K Jackson, and Julien Sage. E2F4 regulates transcriptional activation in mouse embryonic stem cells independently of the RB family. *Nat. Commun.*, 10(1):2939, July 2019.

[73] Jenny Hsu and Julien Sage. Novel functions for the transcription factor E2F4 in development and disease. *Cell Cycle*, 15(23):3183–3190, December 2016.

[74] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. *The lancet*, 395(10223):497–506, 2020.

[75] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, 5(9):e12776, September 2010.

[76] Leonie M Kamminga, Leonid V Bystrykh, Aletta de Boer, Sita Houwer, José Douma, Ellen Weersing, Bert Dontje, and Gerald de Haan. The polycomb group gene ezh2 prevents hematopoietic stem cell exhaustion. *Blood*, 107(5):2170–2179, March 2006.

[77] Shigeki Katoh, Nobuhiro Matsumoto, Kiyoyasu Fukushima, Hiroshi Mukae, Jun-ichi Kadota, Shigeru Kohno, and Shigeru Matsukura. Elevated chemokine levels in bron-

choalveolar lavage fluid of patients with eosinophilic pneumonia. *Journal of allergy and clinical immunology*, 106(4):730–736, 2000.

[78] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

[79] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, May 2015.

[80] Ashley C Kramer, Jenna Weber, Ying Zhang, Jakub Tolar, Ying Y Gibbens, Margaret Shevik, and Troy C Lund. TP53 modulates oxidative stress in gata1+ erythroid cells. *Stem Cell Reports*, 8(2):360–372, February 2017.

[81] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[82] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Erratum: Estimating mutual information [phys. rev. e69, 066138 (2004)]. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 83(1), January 2011.

[83] Chen L, Liu W, Zhang Q, Xu K, Ye G, Wu W, Sun Z, Liu F, Wu K, Zhong B, Mei Y, Zhang W, Chen Y, Li Y, Shi M, Lan K, and Liu Y. Rna based mNGS approach identifies a

novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. *Emerg Microbes Infect. 2020 Feb 5;9(1):313-319. doi:*, 10., 2020.

[84] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, December 2008.

[85] David Lara-Astiaso, Ainhoa Goñi-Salaverri, Julen Mendieta-Esteban, Nisha Narayan, Cynthia Del Valle, Torsten Gross, George Giotopoulos, Tumas Beinortas, Mar Navarro-Alonso, Laura Pilar Aguado-Alvaro, Jon Zazpe, Francesco Marchese, Natalia Torrea, Isabel A. Calvo, Cecile K. Lopez, Diego Alignani, Aitziber Lopez, Borja Saez, Jake P. Taylor-King, Felipe Prosper, Nikolaus Fortelny, and Brian J. P. Huntly. In vivo screening characterizes chromatin factor functions during normal and malignant hematopoiesis. *Nature Genetics*, 55(9):1542–1554, August 2023.

[86] Celine Lefebvre, Presha Rajbhandari, Mariano J Alvarez, Pradeep Bandaru, Wei Keat Lim, Mai Sato, Kai Wang, Pavel Sumazin, Manjunath Kustagi, Brygida C Bisikirska, Katia Basso, Pedro Beltrao, Nevan Krogan, Jean Gautier, Riccardo Dalla-Favera, and Andrea Califano. A human b-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol. Syst. Biol.*, 6(1):377, June 2010.

[87] Xiangjie Li, Kui Wang, Yafei Lyu, Huize Pan, Jingxiao Zhang, Dwight Stambolian, Katalin Susztak, Muredach P Reilly, Gang Hu, and Mingyao Li. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nat. Commun.*, 11(1):2338, May 2020.

[88] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417–425, December 2015.

[89] Kevin Litchfield, Chey Loveday, Max Levy, Darshna Dudakia, Elizabeth Rapley, Jeremie Nsengimana, D Tim Bishop, Alison Reid, Robert Huddart, Peter Broderick, et al. Large-scale sequencing of testicular germ cell tumour (tgct) cases excludes major tgct predisposition gene. *European urology*, 73(6):828–831, 2018.

[90] Yinglei Liu, Boqun Xu, Manhua Liu, Haifeng Qiao, Siming Zhang, Junjun Qiu, and Xiaoyan Ying. Long non-coding rna snhg25 promotes epithelial ovarian cancer progression by up-regulating comp. *Journal of Cancer*, 12(6):1660–1668, 2021.

[91] Kenneth J Livak and Thomas D Schmittgen. Analysis of relative gene expression data using real-time quantitative pcr and the 2- ct method. *methods*, 25(4):402–408, 2001.

[92] Alexandra Louey, Damián Hernández, Alice Pébay, and Maciej Daniszewski. Automation of organoid cultures: Current protocols and applications. *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, 26(9):1138–1147, 2021.

[93] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

[94] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12), December 2014.

[95] Shaolei Lu, Kamaljeet Singh, Shamlal Mangray, Rose Tavares, Lelia Noble, Murray B Resnick, and Evgeny Yakirevich. Claudin expression in high-grade invasive ductal carcinoma of the breast: correlation with the molecular subtype. *Modern Pathology*, 26(4):485–495, April 2013.

[96] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

[97] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.*, 15(6):e8746, June 2019.

[98] Nicholas W Lukacs. Role of chemokines in the pathogenesis of asthma. *Nature Reviews Immunology*, 1(2):108–116, 2001.

[99] Aaron T.L. Lun, Davis J. McCarthy, and John C. Marioni. A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. *F1000Research*, 5:2122, October 2016.

[100] Fábio Madeira, Matt Pearce, Adrian Tivey, Prasad Basutkar, Joon Lee, Ossama Edbali, Nandana Madhusoodanan, Anton Kolesnikov, and Rodrigo Lopez. Search and sequence analysis tools services from embl-ebi in 2022. *Nucleic Acids Research*, 2022.

[101] Fábio Madeira, Matt Pearce, Adrian R N Tivey, Prasad Basutkar, Joon Lee, Ossama Edbali, Nandana Madhusoodanan, Anton Kolesnikov, and Rodrigo Lopez. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.*, 50(W1):W276–W279, July 2022.

[102] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1(S1):S7, March 2006.

[103] Frank J. Massey. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68, March 1951.

[104] N. D. Maulding, S. Seiler, A. Pearson, et al. Dual RNA-Seq analysis of SARS-CoV-2 correlates specific human transcriptional response pathways directly to viral expression. *Sci Rep 12*, 1329, 2022.

[105] Sunnie Grace McCalla, Alireza Fotuhi Siahpirani, Jiaxin Li, Saptarshi Pyne, Matthew Stone, Viswesh Periyasamy, Junha Shin, and Sushmita Roy. Identifying strengths and weaknesses of methods for computational network inference from single-cell RNA-seq data. *G3 (Bethesda)*, 13(3), March 2023.

[106] Davis J McCarthy, Kieran R Campbell, Aaron T L Lun, and Quin F Wills. Scater: preprocessing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics*, 33(8):1179–1186, January 2017.

[107] Satoko Morohashi, Tomomi Kusumi, Fuyuki Sato, Hiroki Odagiri, Hiroki Chiba, Shuichi Yoshihara, Kenichi Hakamada, Mutsuo Sasaki, and Hiroshi Kijima. Decreased expression of claudin-1 correlates with recurrence status in breast cancer. *International Journal of Molecular Medicine*, August 2007.

[108] M Müller, editor. *Dynamic Time Warping. Information Retrieval for Music and Motion*. Springer, Berlin, Heidelberg; Berlin Heidelberg, 2007.

[109] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[110] Fionn Murtagh and Pierre Legendre. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of classification*, 31(3):274–295, 2014.

[111] Leann Myers and Maria J Sirois. Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences*, 12, 2004.

[112] Leann Myers and Maria J Sirois. *Spearman correlation coefficients, differences between*. John Wiley & Sons, Inc., Hoboken, NJ, USA, August 2006.

[113] Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W; China Novel Coronavirus Investigating, and Research Team. A novel coronavirus from patients with pneumonia in china. *2020 Feb 20;382(8):727-733. doi:*, 10., 2020.

[114] Jairo Navarro Gonzalez, Ann S Zweig, Matthew L Speir, Daniel Schmelter, Kate R Rosenbloom, Brian J Raney, Conner C Powell, Luis R Nassar, Nathan D Maulding, Christopher M Lee, et al. The ucsc genome browser database: 2021 update. *Nucleic Acids Research*, 49(D1):D1046–D1057, 2021.

[115] Lee E Neilson, Joseph F Quinn, and Nora E Gray. Peripheral blood NRF2 expression as a biomarker in human health and disease. *Antioxidants (Basel)*, 10(1):28, December 2020.

[116] Yulia Newton, Adam M Novak, Teresa Swatloski, Duncan C McColl, Sahil Chopra, Kiley Graim, Alana S Weinstein, Robert Baertsch, Sofie R Salama, Kyle Ellrott, et al. Tumormap: exploring the molecular similarities of cancer samples in an interactive portal. *Cancer research*, 77(21):e111–e114, 2017.

[117] Robert G Nichols and Emily R Davenport. The relationship between the gut microbiome and host gene expression: a review. *Human genetics*, 140(5):747–760, 2021.

[118] Bernice Nounamo, Yibo Li, Peter O'Byrne, Aoife M Kearney, Amir Khan, and Jia Liu. An interaction domain in human samd9 is essential for myxoma virus host-range determinant m062 antagonism of host anti-viral function. *Virology*, 503:94–102, 2017.

[119] Seog Bae Oh, Takayuki Endoh, Arthur A Simen, Dongjun Ren, and Richard J Miller. Regulation of calcium currents by chemokines and their receptors. *Journal of neuroimmunology*, 123(1-2):66–75, 2002.

[120] SHP Oliveira and NW Lukacs. The role of chemokines and chemokine receptors in eosinophil activation during inflammatory allergic reactions. *Brazilian Journal of Medical and Biological Research*, 36:1455–1463, 2003.

[121] Yaara Oren, Michael Tsabar, Michael S. Cuoco, Liat Amir-Zilberstein, Heidie F. Cabanos, Jan-Christian Hütter, Bomiao Hu, Pratiksha I. Thakore, Marcin Tabaka, Charles P.

Fulco, William Colgan, Brandon M. Cuevas, Sara A. Hurvitz, Dennis J. Slamon, Amy Deik, Kerry A. Pierce, Clary Clish, Aaron N. Hata, Elma Zaganjor, Galit Lahav, Katerina Politi, Joan S. Brugge, and Aviv Regev. Cycling cancer persister cells arise from lineages with distinct programs. *Nature*, 596(7873):576–582, August 2021.

[122] Sreekumar Othumpangat, John D Noti, Cynthia M McMillen, and Donald H Beezhold. Icam-1 regulates the survival of influenza virus in lung epithelial cells during the early stages of infection. *Virology*, 487:85–94, 2016.

[123] Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, and Shi ZL. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature. 2020 Mar;579(7798):270-273. doi:*, 10., 2020.

[124] Rodrigo Alexandre Panepucci, Lucila Habib B. Oliveira, Dalila Luciola Zanette, Rita de Cassia Viu Carrara, Amélia Goes Araujo, Maristela Delgado Orellana, Patrícia Vianna Bonini de Palma, Camila C.B.O. Menezes, Dimas Tadeu Covas, and Marco Antonio Zago. Increased levels of notch1, nf-b, and other interconnected transcription factors characterize primitive sets of hematopoietic stem cells. *Stem Cells and Development*, 19(3):321–332, March 2010.

[125] V Pant, A Quintás-Cardama, and G Lozano. *The p53 pathway in hematopoiesis: lessons from mouse models, implications for humans. Blood, The Journal of the*, volume 120. American Society of Hematology, Washington, DC, 2012.

[126] Giovanni Pasquini, Jesus Eduardo Rojo Arias, Patrick Schäfer, and Volker Busskamp. Automated methods for cell type annotation on scrna-seq data. *Computational and Structural Biotechnology Journal*, 19:961–969, 2021.

[127] Giovanni Pasquini, Jesus Eduardo Rojo Arias, Patrick Schäfer, and Volker Busskamp. Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.*, 19:961–969, January 2021.

[128] F Paul and Y Arkin. a., giladi a., jaitin da, kenigsberg e., keren-shaul h., et al.(2015). *Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. Cell*, 163(7):1663–1677.

[129] Franziska Paul, Ya'ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, David Lara-Astiaso, Meital Gury, Assaf Weiner, Eyal David, Nadav Cohen, Felicia Kathrine Bratt Lauridsen, Simon Haas, Andreas Schlitzer, Alexander Mildner, Florent Ginhoux, Steffen Jung, Andreas Trumpp, Bo Torben Porse, Amos Tanay, and Ido Amit. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 164(1-2):325, January 2016.

[130] Maalavika Pillai, Emilia Hojel, Mohit Kumar Jolly, and Yogesh Goyal. Unraveling non-genetic heterogeneity in cancer with dynamical models and computational tools. *Nature Computational Science*, 3(4):301–313, April 2023.

[131] A. A. Pollen, A. Bhaduri, M. G. Andrews, T. J. Nowakowski, O. S. Meyerson, M. A. Mostajo-Radji, Di Lullo, Alvarado E., Bedolli B., Dougherty M., M. L., and I. T. Fid-

des. Establishing cerebral organoids as models of human-specific brain evolution. *Cell*, 176(4):743–756, 2019.

[132] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and T M Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat. Methods*, 17(2):147–154, February 2020.

[133] Aditya Pratapa, Amogh P Jalihal, Jeffrey N Law, Aditya Bharadwaj, and TM Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature methods*, 17(2):147–154, 2020.

[134] Xiaojie Qiu, Qi Mao, Ying Tang, Li Wang, Raghav Chawla, Hannah Pliner, and Cole Trapnell. Reversed graph embedding resolves complex single-cell developmental trajectories. February 2017.

[135] Florian Rambow, Aljosja Rogiers, Oskar Marin-Bejar, Sara Aibar, Julia Femel, Michael Dewaele, Panagiotis Karras, Daniel Brown, Young Hwan Chang, Maria Debiec-Rychter, Carmen Adriaens, Enrico Radaelli, Pascal Wolter, Oliver Bechter, Reinhard Dummer, Mitchell Levesque, Adriano Piris, Dennie T. Frederick, Genevieve Boland, Keith T. Flaherty, Joost van den Oord, Thierry Voet, Stein Aerts, Amanda W. Lund, and Jean-Christophe Marine. Toward minimal residual disease-directed therapy in melanoma. *Cell*, 174(4):843–855.e19, August 2018.

[136] Ranmali Ranasinghe and Rajaraman Eri. Modulation of the ccr6-ccl20 axis: A potential therapeutic target in inflammation and cancer. *Medicina*, 54(5):88, 2018.

[137] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, 47(W1):W191–W198, 2019.

[138] Lee P. Richman, Yogesh Goyal, Connie L. Jiang, and Arjun Raj. Clonocluster: A method for using clonal origin to inform transcriptome clustering. *Cell Genomics*, 3(2):100247, February 2023.

[139] Alejo E. Rodriguez-Fraticelli, Caleb Weinreb, Shou-Wen Wang, Rosa P. Migueles, Maja Jankovic, Marc Usart, Allon M. Klein, Sally Lowell, and Fernando D. Camargo. Single-cell lineage tracing unveils a role for tcf15 in haematopoiesis. *Nature*, 583(7817):585–589, July 2020.

[140] Alexander Roesch, Mizuho Fukunaga-Kalabis, Elizabeth C. Schmidt, Susan E. Zabierowski, Patricia A. Brafford, Adina Vultur, Devraj Basu, Phyllis Gimotty, Thomas Vogt, and Meenhard Herlyn. A temporarily distinct subpopulation of slow-cycling melanoma cells is required for continuous tumor growth. *Cell*, 141(4):583–594, May 2010.

[141] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

[142] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.*, 37(5):547–554, May 2019.

[143] Katrin Schmidt, Lorena Martinez-Gamboa, Susan Meier, Christian Witt, Christian

Meisel, Leif G Hanitsch, Mike O Becker, Doerte Huscher, Gerd R Burmester, and Gabriela Riemekasten. Bronchoalveoloar lavage fluid cytokines and chemokines as markers and predictors for the outcome of interstitial lung disease in systemic sclerosis patients. *Arthritis research & therapy*, 11(4):1–11, 2009.

[144] Bolette Hellung Schønning, Maja Bévort, Sanne Mikkelsen, Mia Andresen, Peter Thomsen, Henrik Leffers, and Bodil Norrild. Human papillomavirus type 16 e7-regulated genes: regulation of s100p and adp/atp carrier protein genes identified by differential-display technology. *Journal of General Virology*, 81(4):1009–1015, 2000.

[145] Lea Schuh, Michael Saint-Antoine, Eric M. Sanford, Benjamin L. Emert, Abhyudai Singh, Carsten Marr, Arjun Raj, and Yogesh Goyal. Gene networks with transcriptional bursting recapitulate rare transient coordinated high expression states in cancer. *Cell Systems*, 10(4):363–378.e12, April 2020.

[146] Sydney M. Shaffer, Margaret C. Dunagin, Stefan R. Torborg, Eduardo A. Torre, Benjamin Emert, Clemens Krepler, Marilda Beqiri, Katrin Sproesser, Patricia A. Brafford, Min Xiao, Elliott Eggan, Ioannis N. Anastopoulos, Cesar A. Vargas-Garcia, Abhyudai Singh, Katherine L. Nathanson, Meenhard Herlyn, and Arjun Raj. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*, 546(7658):431–435, June 2017.

[147] Sydney M. Shaffer, Benjamin L. Emert, Raúl A. Reyes Hueros, Christopher Cote, Guillaume Harmange, Dylan L. Schaff, Ann E. Sizemore, Rohit Gupte, Eduardo Torre, Abhyudai Singh, Danielle S. Bassett, and Arjun Raj. Memory sequencing reveals heritable

single-cell gene expression programs associated with distinct cellular behaviors. *Cell*, 182(4):947–959.e17, August 2020.

[148] Sreenath V. Sharma, Diana Y. Lee, Bihua Li, Margaret P. Quinlan, Fumiyuki Takahashi, Shyamala Maheswaran, Ultan McDermott, Nancy Azizian, Lee Zou, Michael A. Fischbach, Kwok-Kin Wong, Kathleyn Brandstetter, Ben Wittner, Sridhar Ramaswamy, Marie Classon, and Jeff Settleman. A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell*, 141(1):69–80, April 2010.

[149] Dongyuan Song and Jingyi Jessica Li. Pseudotimede: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell rna sequencing data. *Genome biology*, 22(1):1–25, 2021.

[150] Dongyuan Song and Jingyi Jessica Li. PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data. *Genome Biol.*, 22(1):124, April 2021.

[151] M. L. Speir, A. Bhaduri, N. S. Markov, P. Moreno, T. J. Nowakowski, I. Papatheodorou, A. A. Pollen, B. J. Raney, L. Seninge, W. J. Kent, and M. Haeussler. Ucsc Cell Browser: visualize your single-cell data. *Bioinformatics*, 37(23):4578–4580, 2021.

[152] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19(1):1–16, 2018.

[153] Kelly Street, Davide Risso, Russell B. Fletcher, Diya Das, John Ngai, Nir Yosef, Eliza-

beth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19(1), June 2018.

[154] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19(1):477, June 2018.

[155] G Studzinski, E Garay, R Patel, J Zhang, and X Wang. Vitamin D receptor signaling of monocytic differentiation in human leukemia cells: Role of MAPK pathways in transcription factor activation. *Curr. Top. Med. Chem.*, 6(12):1267–1271, June 2006.

[156] G Studzinski, E Garay, R Patel, J Zhang, and X Wang. Vitamin D receptor signaling of monocytic differentiation in human leukemia cells: Role of MAPK pathways in transcription factor activation. *Curr. Top. Med. Chem.*, 6(12):1267–1271, June 2006.

[157] Yapeng Su, Wei Wei, Lidia Robert, Min Xue, Jennifer Tsoi, Angel Garcia-Diaz, Blanca Homet Moreno, Jungwoo Kim, Rachel H. Ng, Jihoon W. Lee, Richard C. Koya, Begonya Comin-Anduix, Thomas G. Graeber, Antoni Ribas, and James R. Heath. Single-cell analysis resolves the cell state transition and signaling dynamics associated with melanoma drug-induced resistance. *Proceedings of the National Academy of Sciences*, 114(52):13679–13684, December 2017.

[158] Y. Tao, A. Both, R. I. Silveira, K. Buchin, S. Sijben, R. S. Purves, P. Laube, D. Peng, K. Toohey, and M. Duckham. A comparative analysis of trajectory similarity measures. *GIScience Remote Sensing*, 58(5):643–669, 2021.

[159] Adi Laurentiu Tarca, Sorin Draghici, Purvesh Khatri, Sonia S Hassan, Pooja Mittal, Jung-Sun Kim, Chong Jai Kim, Juan Pedro Kusanovic, and Roberto Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25(1):75–82, January 2009.

[160] David Thissen, Lynne Steinberg, and Daniel Kuang. Quick and easy implementation of the benjamini-hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of educational and behavioral statistics*, 27(1):77–83, 2002.

[161] David Thissen, Lynne Steinberg, and Daniel Kuang. Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *J. Educ. Behav. Stat.*, 27(1):77–83, March 2002.

[162] Lu Tian, Yucui Zhao, Marie-José Truong, Chann Lagadec, and Roland P. Bourette. Synuclein gamma expression enhances radiation resistance of breast cancer cells. *Oncotarget*, 9(44):27435–27447, June 2018.

[163] Luyi Tian, Sara Tomei, Jaring Schreuder, Tom S. Weber, Daniela Amann-Zalcenstein, Dawn S. Lin, Jessica Tran, Cindy Audiger, Mathew Chu, Andrew Jarratt, Tracy Willson, Adrienne Hilton, Ee Shan Pang, Timothy Patton, Madison Kelly, Shian Su, Quentin Gouil, Peter Diakumis, Melanie Bahlo, Toby Sargeant, Lev M. Kats, Philip D. Hodgkin, Meredith O'Keeffe, Ashley P. Ng, Matthew E. Ritchie, and Shalin H. Naik. Clonal multi-omics reveals bcor as a negative regulator of emergency dendritic cell development. *Immunity*, 54(6):1338–1351.e9, June 2021.

[164] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the can-

cer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

[165] Sam Tracy, Guo-Cheng Yuan, and Ruben Dries. Rescue: imputing dropout events in single-cell rna-sequencing data. *BMC bioinformatics*, 20(1):1–11, 2019.

[166] Shinsuke Uda. Application of information theory in systems biology. *Biophysical reviews*, 12(2):377–384, 2020.

[167] Shinsuke Uda. Application of information theory in systems biology. *Biophys. Rev.*, 12(2):377–384, April 2020.

[168] Christian Umkehrer, Felix Holstein, Laura Formenti, Julian Jude, Kimon Froussios, Tobias Neumann, Shona M. Cronin, Lisa Haas, Jesse J. Lipp, Thomas R. Burkard, Michaela Fellner, Thomas Wiesner, Johannes Zuber, and Anna C. Obenauf. Isolating live cell clones from barcoded populations using crispra-inducible reporters. *Nature Biotechnology*, 39(2):174–178, July 2020.

[169] Koen Van den Berge, Hector Roux de Bézieux, Kelly Street, Wouter Saelens, Robrecht Cannoodt, Yvan Saeys, Sandrine Dudoit, and Lieven Clement. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat. Commun.*, 11(1):1201, March 2020.

[170] Koen Van den Berge, Hector Roux de Bézieux, Kelly Street, Wouter Saelens, Robrecht Cannoodt, Yvan Saeys, Sandrine Dudoit, and Lieven Clement. Trajectory-based dif-

ferential expression analysis for single-cell sequencing data. *Nature communications*, 11(1):1–13, 2020.

[171] Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12):i237–45, June 2010.

[172] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[173] Juexin Wang, Anjun Ma, Yuzhou Chang, Jianting Gong, Yuexu Jiang, Ren Qi, Cankun Wang, Hongjun Fu, Qin Ma, and Dong Xu. Author correction: scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nat. Commun.*, 13(1):2554, May 2022.

[174] Lingfei Wang, Nikolaos Trasanidis, Ting Wu, Guanlan Dong, Michael Hu, Daniel E Bauer, and Luca Pinello. Dictys: dynamic gene regulatory network dissects developmen-

tal continuum with single-cell multiomics. *Nat. Methods*, 20(9):1368–1378, September 2023.

[175] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[176] Caleb Weinreb, Alejo Rodriguez-Fraticelli, Fernando D. Camargo, and Allon M. Klein. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 367(6479), February 2020.

[177] Alexander J Westermann, Lars Barquist, and Jörg Vogel. Resolving host–pathogen interactions by dual rna-seq. *PLoS pathogens*, 13(2):e1006033, 2017.

[178] Alexander J Westermann, Stanislaw A Gorski, and Jörg Vogel. Dual rna-seq of pathogen and host. *Nature Reviews Microbiology*, 10(9):618–630, 2012.

[179] Ifor R Williams. Chemokine receptors and leukocyte trafficking in the mucosal immune system. *Immunologic research*, 29(1):283–291, 2004.

[180] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.*, 19(1):15, February 2018.

[181] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5, 2018.

[182] F Alexander Wolf, Fiona K Hamey, Mireya Plass, Jordi Solana, Joakim S Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J Theis. PAGA: graph

abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.*, 20(1):59, March 2019.

[183] Roman Wölfel, Victor M Corman, Wolfgang Guggemos, Michael Seilmaier, Sabine Zange, Marcel A Müller, Daniela Niemeyer, Terry C Jones, Patrick Vollmar, Camilla Rothe, et al. Virological assessment of hospitalized patients with covid-2019. *Nature*, 581(7809):465–469, 2020.

[184] Kejin Wu, Ziyi Weng, Qinghua Tao, Gufa Lin, Xiangru Wu, Huiqin Qian, Yichu Zhang, Xiaoyan Ding, Yangfu Jiang, and Yuenian Shi. Stage-specific expression of breast cancer-specific gene -synuclein. *Cancer epidemiology, biomarkers prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 12:920–5, 10 2003.

[185] Yong Xiong, Yuan Liu, Liu Cao, Dehe Wang, Ming Guo, Ao Jiang, Dong Guo, Wenjia Hu, Jiayi Yang, Zhidong Tang, et al. Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in covid-19 patients. *Emerging microbes & infections*, 9(1):761–770, 2020.

[186] Linghui Xu, Siti Khadijah, Shouguo Fang, Li Wang, Felicia PL Tay, and Ding Xiang Liu. The cellular rna helicase ddx1 interacts with coronavirus nonstructural protein 14 and enhances viral replication. *Journal of virology*, 84(17):8571–8583, 2010.

[187] Mingcong Xu, Xuefeng Bai, Bo Ai, Guorui Zhang, Chao Song, Jun Zhao, Yuezhu Wang, Ling Wei, Fengcui Qian, Yanyu Li, et al. Tf-marker: a comprehensive manually curated

database for transcription factors and related markers in specific cell and tissue types in human. *Nucleic acids research*, 50(D1):D402–D412, 2022.

[188] Mingcong Xu, Xuefeng Bai, Bo Ai, Guorui Zhang, Chao Song, Jun Zhao, Yuezhu Wang, Ling Wei, Fengcui Qian, Yanyu Li, Xinyuan Zhou, Liwei Zhou, Yongsan Yang, Jiaxin Chen, Jiaqi Liu, Desi Shang, Xuan Wang, Yu Zhao, Xuemei Huang, Yan Zheng, Jian Zhang, Qiuyu Wang, and Chunquan Li. TF-Marker: a comprehensive manually curated database for transcription factors and related markers in specific cell and tissue types in human. *Nucleic Acids Res.*, 50(D1):D402–D412, January 2022.

[189] Zhe Xu, Lei Shi, Yijin Wang, Jiyuan Zhang, Lei Huang, Chao Zhang, Shuhong Liu, Peng Zhao, Hongxia Liu, Li Zhu, et al. Pathological findings of covid-19 associated with acute respiratory distress syndrome. *The Lancet respiratory medicine*, 8(4):420–422, 2020.

[190] Andy M Yip and Steve Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC bioinformatics*, 8(1):1–14, 2007.

[191] Jiaolong Yu, Jacob Benesty, Gongping Huang, and Jingdong Chen. Optimal single-channel noise reduction filtering matrices from the pearson correlation coefficient perspective. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, April 2015.

[192] Zhuohan Yu, Yanchi Su, Yifu Lu, Yuning Yang, Fuzhou Wang, Shixiong Zhang, Yi Chang, Ka-Chun Wong, and Xiangtao Li. Topological identification and interpreta-

tion for single-cell gene regulation elucidation across multiple platforms using scMGCA. *Nat. Commun.*, 14(1):400, January 2023.

[193] Huijuan Zeng, Jing Pan, Chao Hu, Jiliang Yang, Jiahao Li, Tianbao Tan, Manna Zheng, Yuanchao Shen, Tianyou Yang, Yun Deng, and Yan Zou. Snhg25 facilitates snora50c accumulation to stabilize hdac1 in neuroblastoma cells. *Cell Death amp; Disease*, 13(7), July 2022.

[194] Yong Zhang, Dailing Mao, William T Roswit, Xiaohua Jin, Anand C Patel, Dhara A Patel, Eugene Agapov, Zhepeng Wang, Rose M Tidwell, Jeffrey J Atkinson, et al. Parp9-dtx3l ubiquitin ligase targets host histone h2bj and viral 3c protease to enhance interferon signaling and control viral infection. *Nature immunology*, 16(12):1215–1227, 2015.

[195] Zhang Zhiyu, Zhou Qi, Song Zhen, Zhang Jianglei, and Ouyang Jun. Small nucleolar rna host gene 25 is a long non-coding rna helps diagnose and predict outcomes in prostate cancer. *Cancer Treatment and Research Communications*, 35:100687, 2023.

[196] Bowen Zhou, Amanda Moodie, Anne Blanchard, Etienne Leygue, and Yvonne Myal. Claudin 1 in breast cancer: New insights. *Journal of Clinical Medicine*, 4(12):1960–1976, November 2015.

[197] QING ZHUANG, CAIYUN LIU, LIKE QU, and CHENGCHAO SHOU. Synuclein-promotes migration of mcf7 breast cancer cells by activating extracellular-signal regulated kinase pathway and breaking cell-cell junctions. *Molecular Medicine Reports*, 12(3):3795–3800, May 2015.