Unsupervised methods for pattern discovery in high-throughput genomic data

by

Kristina Lynn Buschur

B.A., Kenyon College, 2011

Submitted to the Graduate Faculty of

the School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Kristina Lynn Buschur

It was defended on

April 26, 2019

and approved by

Veronica Hinman, Professor, Department of Biological Sciences, Carnegie Mellon University

Dennis Kostka, Assistant Professor, Department of Developmental Biology

Steffi Oesterreich, Professor, Department of Pharmacology and Chemical Biology

Dissertation Director: Panayiotis V. Benos, Professor, Department of Computational and Systems Biology

Unsupervised methods for pattern discovery in high-throughput genomic data

Kristina Lynn Buschur, PhD

University of Pittsburgh, 2019

Large –omics experiment datasets are being generated at an increasingly fast pace. They present bountiful opportunities for insight into complex diseases and systems but also new challenges in analysis. Novel approaches are needed to make sense of these high-throuput data and especially to consider them jointly for a more complete picture of the system's biology. In this dissertation, we have focused on improving clustering in high-throughput biological datasets by developing a variety of new features that are specifically tailored to reflect the biological properties of the systems we are trying to understand. We started by proposing new features for representing transcription factor binding sites that capture both the DNA sequence composition of the binding region and the TF-DNA binding strength. We observed that these new features aided clustering for improved DNA binding motif discovery. Next, we presented a new method, single sample network perturbation assessment (ssNPA), and demonstrated how causal network learning algorithms could be used to build features that capture the complex interactions of variables within biological systems such as gene regulatory networks and cluster samples based on how these networks are deregulated in different subtypes. We validated this method in a murine liver cell development dataset and with transcriptomic datasets comparing breast cancer and lung adenocarcinoma tumor samples to normal tissue. Then we used ssNPA to describe new subtypes of chronic obstructive pulmonary disease (COPD) that were based on their relative gene network deregulation compared to normal samples. Finally, we applied causal network modeling techniques to two datasets of chronic lung diseases, exploring the systems biology of lung function

decline in COPD at the body systems level and cell type interactions in idiopathic pulmonary fibrosis (IPF) at the scale of the gene expression in single cells.

Table of Contents

List of Tables

List of Figures

Preface

This dissertation is dedicated jointly to my parents, Roy and Julie Buschur, who have given me everything, and my husband, Karl Nagy, who builds me up every day.

I would like to thank many family and friends for their overwhelming love and support during this work and beyond, especially Jonathon Buschur and Amy Hutchison, Michael and Ethel Nadasi, Louis and Evelyn Buschur, Gayle Hoover and Joseph Nagy, Ben Buschur, Kathy Rowlands, Lisa Nadasi, Michele and Vince Brackman, Jeanne Nagy and Matt Vail, Ericka Mochan, Timothy Song, Rory Donovan-Maiye, AJ Sedgewick, and Laura Tipton. I would also like to express my sincere gratitude and appreciation to my advisor, Takis Benos, for all his guidance, support, and humor.

This work would not have been possible without the contributions of numerous colleagues and collaborators. I would like to thank the members of my thesis committee, Veronica Hinman, Dennis Kostka, and Steffi Oesterreich, for all their time and generous feedback and guidance. Julia Zeitlinger, Qiye He, and Jeff Johnston contributed to the ChIP-nexus analysis. Maria Chikina was a generous collaborator in the development of the ssNPA algorithm. Frank Sciurba, Craig Riley, Craig Hersh, Abby Saferali, Peter Castaldi, and Michael Cho collaborated on the COPDGene subtyping project, and AJ Sedgewick, Ivy Shi, Vineet Raghu, Dimitris Manatakis, and Frank Sciurba collaborated on the SCCOR lung function decline study. Finally, the IPF project was done in collaboration with Bob Lafyatis, Nina Morse, Tracy Tabib, John Sembrat, Ana Mora, and Mauricio Rojas.

1.0 Introduction

Large genomic datasets are becoming increasingly abundant. High-throughput sequencing cost is significantly decreasing and the answers to questions related to the entire human genome, transcription, and regulome and their connections to development and disease are more accessible than ever before. The speed and ease with which we can collect these datasets have now far exceeded our ability to analyze and interpret them. New methods are needed for us to make the most out this treasure trove of data, and there are a huge variety of machine learning methods being developed to fill the gap. This dissertation aims to contribute to that work through taking a careful consideration of the biology of particular system and developing approaches that are specifically targeted to important applications and questions regarding systems biology and disease.

The majority of this work focuses on unsupervised learning with the goal of clustering biological objects, ranging from single transcription factor genomic binding sites all the way up to subjects diagnosed with chronic lung disease. One of the main goals of clustering is subtype identification. Subtyping is useful in a number of ways. For experiments in which we are trying find a signal that is too weak to detect in a heterogeneous collection of samples, separating the samples into several smaller groups with more consistent characteristics can strengthen the signal of interest and make downstream analyses easier and more productive. Given a set of samples that have been grouped together in a cluster, patterns can often emerge that would not have otherwise been obvious.

A significant challenge in designing a successful clustering experiment is choosing the right feature space in which to represent the objects for clustering. Without careful attention to feature selection, there is no guarantee that objects will be grouped together into clusters in a way

that is relevant to the questions one is trying to answer or even meaningful in any way. We developed a number of new features that are specifically tailored to reflect the biological properties of a given object. For example, in Chapter 2 we propose new features for representing and partitioning whole-genome transcription factor occupancy datasets (e.g. ChIP-seq, ChIP-exo, and ChIP-nexus). We have found that features that capture both the DNA sequence composition of the binding region and the TF-DNA binding strength improve clustering performance with the goals of DNA-binding motif discovery and understanding genomic organization.

In Chapter 3, we used sophisticated causal network learning algorithms to devise features that reflect how complicated systems of variables (e.g. gene expression measurements) interact and how those interactions change among control samples and case subtypes. With objects represented in this feature space we could group them together into clusters in ways that directly corresponded to the deregulation of their gene regulatory networks. We call our new method Single Sample Network Perturbation Assessment (ssNPA). While the examples we have provided here are limited to gene expression variables, this method is flexible and can consider the interactions among a huge variety of variables generated by high-throughput experiments such as genetic variant, methylation, or protein expression data. Even clinical measurements could be incorporated into the networks used for feature calculation. We validated this method in a variety of transcriptomic datasets, including single cell RNA-seq measurements of liver cell development in mice, RNA-seq studies of human breast cancer and lung adenocarcinoma.

Then, in Chapter 4, we explored an application of ssNPA for discovering subtypes of chronic obstructive pulmonary disease (COPD) based on gene expression, which, in the literature, has proven to be a particularly difficult problem. Finally, in Chapter 5, we further explore causal

network modeling and how these methods can be applied to the study of chronic lung diseases and adapted to make use of large, multi-modal biological data.

2.0 Quantitative *k*-mer transcription factor peak clustering

Transcription factor (TF)-DNA *in vivo* binding is a complex procedure, which depends on the cell type and pathways the target genes are involved. A TF can have altered binding specificities, depending on the context, which may include chromatin status and interacting TFs. This is expected to affect the peak properties in the chromatin immunoprecipitation experiments. In this article, we investigate whether these contextual differences are truly reflected in and can be recovered from genome-wide occupancy data (ChIP-seq/nexus). We systematically evaluate various peak features, including peak shape and sequencing depth, and pre-clustering strategies, using simulated and experimental data (publicly available and new datasets). Comparison of peak shape and quantitative *k*-mer frequencies shows that the latter can capture more properties of TF binding and are thus more useful in identifying TF binding sites and motifs by reducing false binding site prediction. We conclude that quantitative *k*-mer peak clustering can aid a variety of downstream analyses for motif discovery and genomic annotation. We show that it outperforms previously described aggregation plot shape features on synthetic and real ChIP-seq datasets, and we provide examples of the use of our method to explore real ChIP-seq and ChIP-nexus data from human cell lines and *Drosophila* embryos, where biologically meaningful patterns are discovered. The new approach we propose leads to the identification of motif subclasses (co-binding TFs) and discovery of new biological knowledge.

2.1 Background

Transcription factors (TF) recognize short DNA signals at the open chromatin regions (e.g. promoters, enhancers) and regulate gene expression. A powerful technique to identify the genomic regions bound by a certain TF is chromatin immunoprecipitation (ChIP) with TF-specific antibody, followed by sequencing of the precipitated regions. ChIP-seq has provided useful insights in the regulatory circuits of many organisms (1-3). For a more general assessment of open chromatin regions, DNase-seq (4, 5), FAIRE-seq (6), and ATAC-seq (7) are used.

Classical ChIP-seq and open chromatin assessment methods do not identify the exact location of transcription factor binding sites (TFBS), but rather an extended region where binding may have occurred. This problem of resolution has been dramatically improved with the introduction of ChIP-exo (8) and our more recent variant ChIP-nexus (9), which pinpoint more accurately the exact DNA position where the immunoprecipitated protein is cross-linked to DNA. However, since protein-protein crosslinks are also frequent during formaldehyde fixation, indirect binding through other TFs could still be detected with this method, so motif detection remains a more complex problem.

Once the genomic regions that are assessed by a given experiment have been identified, we are still faced with the task of interpreting and finding meaningful patterns in these large biological datasets. Previous work on ChIP-seq data has sought to partition peaks into clusters to identify different patterns within them (10-12). However, these efforts have focused almost exclusively on histone modification datasets with peaks aligned to the nearest transcription start site (TSS). These histone datasets typically have broad peaks that can have various shapes, so the read densities in bins over the length of the peak interval (aggregation plots) were used as features to cluster the peaks. Sequence specific TF binding sites, however, present a unique problem that,

for a number of reasons, cannot be adequately addressed by making use of these aggregation plots. First, a TFBS can occur nearly anywhere in the genome, in genic or intergenic regions. Therefore, aligning them based on their nearest TSS binding site is not very practical. Typically, we analyze a TF ChIP-seq dataset by aligning them according to the TF DNA motif or the peak summit. Assuming the TF motif can be found near the summit of the peak, this results in all the peaks being co-localized after alignment. Furthermore, sequence specific TFs tend to produce much narrower and Gaussian-shaped peaks than those found in histone modification datasets. Thus, the shape of the peaks may not vary tremendously within a dataset (or datasets), so peak shape is not a promising feature for partitioning these samples. Instead, we have decided to explore sequence-based features for partitioning TF ChIP-seq/nexus datasets for pattern discovery. MuMoD is another method that relies on sequence information to partition TF ChIP-seq datasets, but it is specifically focused on motif discovery for partitioning and is limited to a single motif (mode of binding) per cluster (13).

In this paper, we investigate whether the characteristics of the ChIP peaks can be used to address the problem of multiple TF binding and the related problem of TF submotif identification in ChIP-seq/exo/nexus data. Specifically, we compare the peak shape features, the sequence features and combinations of these as well as different peak clustering methods to assess which can better partition the different classes of TF binding. Using the best performing methods, we analyze ChIP-seq and ChIP-nexus data from human cell lines and Drosophila embryos, where meaningful associations are discovered. Used for partitioning high-throughput TF occupancy datasets, our results show promise for identifying new DNA-binding motifs and co-binding factors in future analytic strategies.

2.2 Materials and Methods

We sought to identify a method for optimal partitioning of TF ChIP-seq peaks in a way that captures the differences in DNA binding, either in sequence composition or in peak shape or both. To this end, we compared a number of feature space representations and clustering methods of the peaks in both synthetic and real (ChIP-seq, ChIP-nexus) datasets.

2.2.1 Feature extraction

We represent each ChIP-seq peak by a feature vector. This feature vector includes both peak shape and peak sequence features. We compared various feature vectors to determine the best performing ones for TF ChIP-seq datasets.

2.2.1.1 Shape

Shape features capture the shape and height of each peak from the ChIP-seq data. The number of extracted features is equal to the number of bins that span the length of the genomic interval the peak covers. These features are known as aggregation plots (APs), and previous work using these types of features has applied bins of various sizes (11). For our purposes, we used a bin size of 1 because of the relatively short peak lengths, especially in the ChIP-nexus data. Bin size of 1 corresponds each feature to a single base pair. For example, for a set of peaks of length 100, each peak is represented in $\mathbb{Z}_{\geq 0}^{100}$ feature space. The value of each feature is the number of mapped reads covering that position. If we had long peak lengths and were to use bins that span more than one base as features, then the value of each feature would be the average number of reads that cover each position in that bin.

2.2.1.2 Sequence

Sequence features were calculated to reflect the sequence composition of the peaks. For this purpose, we calculate the *k*-mer frequencies over the peak. For each *k*-mer, we counted the number of times it appears in the peak sequence using a sliding window of length *k*. We considered only canonical *k*-mers, i.e. each *k*-mer and its reverse complement were treated as one, to account for TF binding on either strand of the DNA. So, for each *k* there are $\frac{1}{2}\left(4^k + 4^{\frac{k}{2}}\right)$ *k*-mer features when *k* is even and $\frac{4^k}{2}$ features when *k* is odd. Then, we divided each *k*-mer count by the expected count for that *k*-mer. We calculated the expected count by using Jellyfish (14) to calculate the *k*-mer frequencies for a set of background sequences. Then, we multiplied the sum of *k*-mers present in the sequence by the background *k*-mer frequencies to get the expected value for each *k*-mer. By dividing by the expected number of each *k*-mer, we aim to amplify the functionally relevant differences among sequences. For our synthetic and human datasets, the background sequences were the 1 kb regions upstream of the hg19 transcription start sites. For the *Drosophila* ChIP-nexus datasets, we took the 1 kb regions upstream of the dm3 transcription start sites as background.

These sequence features do not contain any information about the shape or height of the peak. During our performance comparisons we used a range of values for *k* from 3 to 6, since TF binding sites are typically 6-12 bases long.

2.2.1.3 Scaled sequence

Scaled sequence features were also *k*-mer features that account for both the binding strength and the sequence content of the peaks. In this case, the number of times the *k*-mer appears in the peak sequence was multiplied by the peak height (number of peak reads containing that *k*-

mer) at that position. Assuming that the differences between various classes of peaks are more likely to occur at or near the summit of a peak, these features will amplify that signal and the clustering step should be able to separate the classes more easily. Like the sequence features, we only used canonical *k*-mers, resulting in the same number of total features for a given value of *k*. Additionally, as with the sequence features, we divided each feature entry by the expected number of each *k*-mer for that peak. In this case, the total number of *k*-mers for a given peak is the sum of all the *k*-mers in the peak, where each *k*-mer present in the sequence has been multiplied by the height of the peak at that position. We used the same sets of sequences to calculate background *k*-mer frequencies as described for the sequence features. We tested all values of *k* from 3 to 6.

2.2.2 Feature clustering

We tested a variety of standard and more recently developed clustering methods to determine which resulted in the best partitioning of the ChIP-seq peaks.

2.2.2.1 k-means

k-means was performed using the MATLAB function kmeans with 10 replicates and *c* clusters. When the number of true classes was known, we used that value for *c*. Estimating the number of true classes when it is not already known is an open and well-documented problem that is beyond the scope of this work. Thus, we chose to use either $c=2$ or $c=3$ to begin to investigate datasets in which the underlying classes were not already characterized. However, this is a parameter the user should vary based on the dataset he is using or the questions she wants to explore. We tested k-means clustering using both the squared Euclidean distance (subsequently referred to as Euclidean) and 1 minus the sample correlation (correlation) as distance metrics.

2.2.2.2 *k*-medoids

k-medoids was performed with the MATLAB function kmediods, using the same parameters as k-means. We tested this clustering method using both Euclidean and correlation distance metrics.

2.2.2.3 Probabilistic partitioning

We implemented the probabilistic partitioning method described by Nair et al. (11) in MATLAB. Their approach uses expectation maximization (EM) to optimize a mixture model. With this method, each peak has a probability for being a member of each cluster, and the highest probability at the end of the optimization is used for cluster assignment. We tested both the basic implementation of this method, which we refer to as "Partition," and the modified version that only takes into account the "shape" of the features, which we call "Partition (Shape)." Because we were representing our peaks with features that included sequence information in addition to shape, this method does not exactly correspond to peak shape, but we still refer to it in this way for consistency with previous work. Clusters were seeded randomly as previously described (11), and a fixed number of 30 EM iterations were carried out. When we test this method on the synthetic datasets, where the true number of clusters is known, we provide the true number of clusters as a parameter to the algorithm.

2.2.3 Clustering performance measures

In the cases where the true underlying classes were known (i.e., the synthetic data and the combined ENCODE datasets), we assessed clustering performance by calculating the percent classification error. With a small number of classes (e.g. <10), it was straightforward to test all

combinations of class to cluster mappings and choose the labeling that minimized percent classification error. Classification error was calculated by the number of peaks that were assigned to the wrong cluster divided by the total number of peaks.

2.2.4 Synthetic data generation

We first used a collection of synthetic datasets to evaluate the various peak representations and clustering methods. We simulated the ChIP-seq peaks in terms of both shape and TF binding strength, the latter as proportional to the sequence reads in each peak. Thus, the synthetic data were generated in two steps.

1. First we simulated the shape of the peaks. Similarly to (11), we generated Gaussian peak shapes with mean *m*, standard deviation *s*, and coverage *f*. This was achieved by sampling from Poisson distributions to generate 100 bin counts, covering the entire length of the peak. The parameters of the Poisson distribution could vary among classes and along the length of the peak. We tested different combination of parameters with *m* varying from 35 to 50, *s* from 10 to 25, and *f* from 5 to 1,000.

2. Next, we simulated the peak sequences. Each sequence was 100 bp long, and the background was generated from a 3rd order Markov model of human promoter sequences (1 kb upstream of the transcription start site). A transcription factor target site derived from a synthetic motif was inserted into the sequence, centered at the peak mean. Synthetic motifs were generated by randomly selecting columns of high information content (IC > 1.8 bits), medium information content (1.6 bits ≤ IC ≤ 1.8 bits), or low information content (0.8 bits ≤ IC ≤ 1.2 bits) from motifs in the JASPAR database (15). The length of the synthetic DNA motifs was between 8 and 12 bases long, depending on the experiment.

For each simulated dataset, we generated 1,000 peaks (samples). To simulate a ChIP-seq experiment for a TF that can bind with altered specificity depending on cellular or genomic context, we used the primary motif for class 1. Then for the rest of the classes, we shuffled the nucleotide labels of a subset of the columns to create derivative motifs. We guaranteed that the nucleotide with the highest frequency in each column was changed in the derivative motif. This preserves the information content of the column, but changes the base preference at that position. Which motif column labels are shuffled greatly impacts the *k*-mer composition of the resulting sequences, so we exhaustively tested all combinations of columns. For example, with a motif of length 10, we performed experiments with 210 different combinations of columns changed in the derivative motif. This ranged from changing 0 columns (the motifs for each class were the original) all the way to shuffling labels for all 10 columns. Thus, we were able to observe the full distribution of the effect of altered TF binding specificity on clustering performance. In further experiments, we set k=5 and only shuffled columns 1 and 6 of the motif of length 10, which results in all *k*-mers spanning the motif to differ between classes.

We did not simulate sequencing reads directly, so we needed to slightly modify how the scaled sequence features were calculated in the synthetic datasets. For each *k*-mer in the peak sequence, we added the height of its first base (proportional to a theoretical number of reads at that base) to the frequency of this *k*-mer feature.

We generated a collection of synthetic datasets to explore the effects of a variety of parameters including the number of classes, peak means and standard deviations of each class, coverage, TF motif length, TF motif information content, and the number of columns that differed between the TF motifs of different classes. All running times presented are wall-clock times calculated on a computer with a 2.7 GHz Intel Core i5 processor and 16 GB of memory.

2.2.5 ENCODE ChIP-seq data

We used three human ChIP-seq datasets from ENCODE to test whether the different feature calculation methods, including our scaled sequence method, can distinguish among different TF ChIP-seq peaks from the same cell line. Alignments were downloaded for GABPA, GATA2, and YY1, as well as the corresponding controls. All ChIP-seq experiments were done in erythroleukemia cells (K562). Full accession information is provided in Table S1.

Table 1 GEO accession information for ENCODE datasets

Transcription Factor	Cells	Accession Number	File Type	Bio Replicate	Control Accession
GABPA	human K562	ENCFF000QAH	bam	1	ENCFF772PJM
GATA2	human K562	ENCFF000QAT	bam	1	ENCFF772PJM
YY1	human K562	ENCFF000QKE	bam	1	ENCFF772PJM

Peaks were called for the alignment files with MACS (16) using default parameters, and the top 5,000 peaks with lowest FDR for each TF were used for feature calculation and clustering. For use in calculating the scaled sequence features, aligned reads were extended in the 3' direction by length *d*, where *d* is the distance between the summits of the forward and reverse strand reads, as calculated by MACS. This helps to ensure that any TF binding signals are covered by the short sequencing reads, and will, we expect, improve clustering performance.

To determine if the clustering step could distinguish among different transcription factors' binding sites, peak features from the different datasets were pooled together, and the concatenated feature sets were clustered. To account for different levels of sequencing coverage among peaks, read counts were scaled such that each experiment had a total of 20 million reads.

2.2.6 Drosophila ChIP-nexus data

ChIP-nexus experiments were performed in duplicates as previously described (9) with 10 µg of rabbit polyclonal antibodies against the C-term region (aa 168-345) of Drosophila Scute, which were raised and affinity-purified by Genescript. The chromatin extracts were prepared from 2-4h AEL *Drosophila* embryos from mothers with the genotype Tlrm9/Tlrm10.

ChIP-nexus produces the same kind of data as ChIP-exo (17), but the library preparation is more efficient and includes a random barcode to correct amplification bias (9). For both methods, the output data are the first bases of each read, which corresponds to the beginning of the protected DNA region due to TF binding.

ChIP-nexus data are in the form of chromosome:position:orientation, which denotes the first protected base in each unique read. In order to call the peaks of these reads, reads were extended in the 3' direction to a length of 50 bp. This corresponds to ~5 turns of the DNA and would cover the length of the TF binding site as well as any co-bound binding motifs. Increasing this threshold did not improve the results (data not shown). Peaks were called with GEM using the suggested settings for ChIP-exo data (17). We used the following parameters: k_min=6 and k_max=13 (range for *k*-mer length), smooth=3 (the width in bp over which to smooth the read distribution), and no duplicate read filtering because of the barcoding used in the ChIP-nexus technique. Additionally, we used the ChIP-exo starting read distribution file provided with the GEM software. The algorithm identified single base pair loci for binding events, which we extended 50 bp in both directions to capture the entire peak region. For each replicate, we analyze the top 5,000 peaks, according to largest –log(q-value), as calculated by GEM. Through this preprocessing, we take advantage of the increased accuracy of the ChIP-nexus technique for

discovering binding sites, while providing a suitable input for our feature calculation and partitioning methods.

2.2.7 Drosophila CTCF ChIP-seq data

Drosophila CTCF-GFP ChIP-seq data from whole organism embryo (0-16 hrs) were downloaded from ENCODE: https://www.encodeproject.org/experiments/ENCSR661BEZ/. The data are derived from the modERN Project (Kevin White, University of Chicago). To make them comparable to ChIP-nexus data, the optimal IDR thresholded peaks were centered and trimmed to 100 bp.

2.2.8 Motif discovery and peak annotation

Motif detection was performed on the peak clusters using HOMER (18). Default parameters were used except for the following: size=entire peak length; len=6, 8, 10,12 to increase the search space to include motifs of length 6, 8, 10, or 12; and finally S=10 to optimize 10 motifs. HOMER was also used for peak annotation.

2.3 Results and Discussion

2.3.1 *k*-mer length selection considerations

In order to determine the optimal *k*-mer length to use for calculating the sequence-related features, we tested lengths ranging from 3 to 6. We did not consider *k*=1 (GC content) or *k*=2 (too

few possible features). There are several points to consider when deciding which *k*-mer length to use:

1. The number of possible *k*-mer features, *z*, increases exponentially with increasing *k*. This results in a significant increase in the time and memory required for feature calculation and clustering.

2. The maximum possible number of non-zero features for a sequence of length s is given by *s*-*k*+1. So, for example, using *k*=6, a sequence of length *s*=100 could have a maximum of 95 different *k*-mers represented. This is only a very small fraction of the 2,080 *k*-mer features possible.

3. DNA sequences are often not represented well by a simple multinomial model, so we need to find a value for *k* that is large enough that any motif signals are distinguishable from sequence background. Indeed, we used a third-order Markov model to generate the background sequences in our synthetic datasets, so we would not expect *k*=3 to be sufficient to distinguish between classes.

We compared the classification performance for *k* in the range of 3 to 6 in distinguishing between two classes of sequences that differed in the PWMs used to generate the motif sequence signal (Figure 1A). We tested this with PWMs of 8, 10, and 12 columns and exhaustively shuffled every possible combination of columns to differ between the PWMs for each class (for example, for a PWM with 10 columns, we created 2^{10}=1,024 datasets). Scaled *k*-mer features were calculated for each peak, and k-means with Euclidean distance was used to partition the peaks into two clusters. In general, we observed that increasing the length of the PWM led to lower classification error for all values of *k*. In other words, a longer motif would produce a stronger signal, which is expected and in turn, it improves clustering performance. For *k*=3, there is a very large spread of

classification error across all motif lengths, with the bulk of the datasets falling in the range of 0-15% classification error. We observe a steep drop in errors at k=4, with much tighter distributions. At k=5, classification error is near 0% for almost all datasets, particularly when the PWM length is 8 or greater. Finally, k=6 shows a very minimal improvement that is really not enough to justify the significant increases in memory and runtime we experienced for these datasets. We excluded the case where the k-mer length and motif length were both equal to 6 because this would correspond to the full difference between classes being captured in a single feature.

Increasing from k=5 to k=6 corresponded to a very large increase in the average time required to both calculate the features and perform k-means clustering (Figure 1B). The feature calculation time was measured for our simplified data simulation process, but the increased demands at k=6 are exacerbated in real data sets for which we need to process large numbers of sequencing reads. Thus, we decided to proceed using k-mers of length k=5, although this remains a parameter the user can set according to his or her particular needs.

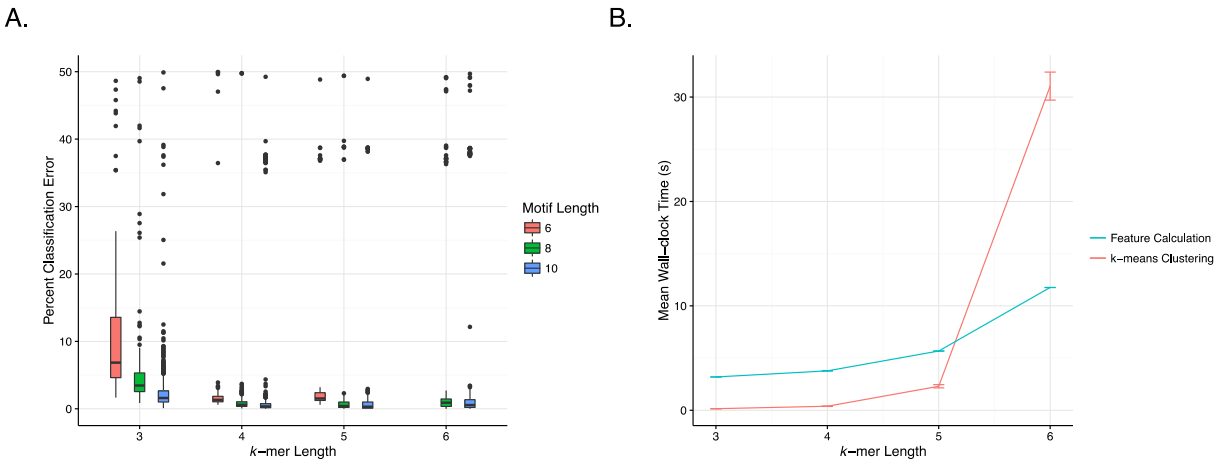

Figure 1 (A) Percent classification error decreases with both increasing *k*-mer length and TF motif length. For each dataset, two classes were constructed, each with 1000 samples. Each sample is a synthetic peak of length 100 bp, simulated with mean 50, standard deviation 10, and coverage 1000. These samples were represented with our scaled sequence features and clustered with k-means with Euclidean distance. Each distribution shown represented 2^l datasets, where l is equal to the motif length. The distribution covers all possible combinations of the columns that can differ between the motifs of the two classes. The case where the *k*-mer length and motif length were both equal to 6 is excluded. (B) The mean time required for both feature calculation and k-means clustering increases exponentially with the length of the *k*-mer used to calculated the scaled sequence features. Error bars show the standard error of the mean.

2.3.2 Scaled *k*-mer features improve separation of co-localized peaks over peak shape or sequence alone

We compared several types of features in their ability to separate simulated peaks according to their motif class. First, we considered shape features for which there is a feature for every genomic position along the length of the peak and its value is the number of mapped reads

covering that position (also known as an aggregation plot). These features were used previously for partitioning histone peaks. In their case, they binned the read counts into 20-50 bp (11), compared to our single base bins, because they did not consider sequence features. Next, we calculated sequence features. For each *k*-mer of length 5 (treating a *k*-mer and its reverse complement as one), we counted the number of times it appears in the peak sequence using a sliding window. Finally, we considered scaled sequence features. These are similar to sequence features, but, for every *k*-mer, the number of times the *k*-mer appears in the peak sequence was multiplied by the peak height (number of reads containing that *k*-mer) at that position. This way we are capturing information about both the shape (binding strength) and the sequence of the peak.

Table 2 Average percent classification error (standard error of the mean) with different combinations of features and clustering methods.

Features	*k*-means Euclidean	*k*-means Correlation	*k*-medoids Euclidean	*k*-medoids Correlation	Partition	Partition Shape Only
Shape	49.44 (0.01)	-	48.47 (0.00)	-	49.07 (0.02)	49.05 (0.02)
Sequence	7.65 (0.49)	2.28 (0.15)	13.01 (0.33)	10.28 (0.27)	50.00 (0.00)	50.00 (0.00)
Scaled Sequence	0.98 (0.13)	-	0.92 (0.06)	-	50.00 (0.00)	50.00 (0.00)

We calculated each type of features for the simulated datasets. Each dataset contained two motif classes and 1,000 samples (peaks) of each class. The simulated binding sites were of length 10, and we created one dataset for each possible subset of the 10 columns for which the nucleotide labels were shuffled in the TF motifs that generated the sites that differed between the two classes. Thus, we tested 1,024 datasets. All generated peaks were 100 bp long, with a shape of mean m=50, standard deviation s=10, and coverage f=1000. We partitioned the peaks using six different methods: k-means with either Euclidean or correlation distance, k-medoids with either Euclidean

or correlation distance, and the probabilistic partitioning method proposed by Nair et al. (11) with either the basic or shape-only implementation (Table 2). We could not use the correlation-based methods on the shape or scaled sequence features because the standard deviation across the features for the peaks was too small. For the other clustering methods, the scaled sequence features resulted in the lowest average percent classification error: 0.98% for k-means (Euclidean) and 0.92% for k-medoids (Euclidean). Both probabilistic partitioning methods performed very poorly for all features, but in different ways. In the case of the shape features, the two clusters were typically of approximately equal size, but the classes were randomly distributed between them. For the sequence and scaled sequence features, all peaks in both classes were assigned to a single cluster. Overall, these results make sense considering that all the peaks were co-localized and had the same shape, similar to what we might expect to see in real ChIP-seq datasets in which the peaks are centered around their summits (presumably containing the TF binding site). This differs from the aggregation plots (our shape features) often used for analyzing histone ChIP-seq datasets, in which peaks are commonly aligned by their distance to the nearest transcription start site or some other external locus and, thus, you would not necessarily expect all the peaks to be co-localized. These results agree with those found by Nair et al. (11), where co-localized peaks proved to be a much more difficult task for their probabilistic partitioning method. Furthermore, it is not surprising that the scaled sequence features outperformed the sequence features because the differences between the peak classes occur at the peak summits, and this signal is amplified in the scaled sequence features.

Table 3 Percent classification error decreases with increasing difference between the peak means for the two classes. m_1=50 and m_2 varies between 35 and 50. The peak standard deviation for both classes is set to 10.

Method	Features	$m_2 = 35$	$m_2 = 40$	$m_2 = 45$	$m_2 = 50$
Shape	k-means (Euclidean)	0	0	0	49.8
	k-medoids (Euclidean)	0	0	0	48.7
	Partition	0	0	0	50
	Partition (Shape)	0	0	0	49.8
Sequence	k-means (Euclidean)	1.95	2.05	1.95	1.9
	k-means (Correlation)	1.95	2	2.15	1.95
	k-medoids (Euclidean)	3.15	2.95	2.85	4.6
	k-medoids (Correlation)	5.7	5.95	4.65	7.5
	Partition	50	50	50	50
	Partition (Shape)	50	50	50	50
Scaled Sequence	k-means (Euclidean)	1.4	1.45	1.35	1.4
	k-medoids (Euclidean)	1.3	1.4	1.35	1.5
	Partition	50	50	50	50
	Partition (Shape)	50	50	50	50

We also tested the various features and clustering methods in cases where the peaks in the two classes had different shapes- either different means or different standard deviations. We compared the percent classification errors for these cases (Table 3 and Table 4). As the differences between the shapes of the peaks in the two classes were increased, the shape features began to perform very well across all methods. However, the sequence and scaled sequence features also performed well and more reliably across all shape differences when paired with k-means or k-medoids.

Table 4 Percent classification error decreases with increasing difference between the peak standard deviations for the two classes. s_1=10 and s_2 varies between 10 and 25. The peak mean for both classes is set to 50.

Method	Features	$s_2 = 10$	$s_2 = 15$	$s_2 = 20$	$s_2 = 25$
Shape	k-means (Euclidean)	49.8	0	0	0
	k-medoids (Euclidean)	48.7	0	0	0
	Partition	50	0	0	0
	Partition (Shape)	49.8	0	0	0
Sequence	k-means (Euclidean)	1.9	2.05	2.15	2.25
	k-means (Correlation)	1.95	2.5	2.25	2.65
	k-medoids (Euclidean)	4.6	4.7	4.75	3.05
	k-medoids (Correlation)	7.5	5.7	4.75	6.75
	Partition	50	50	50	50
	Partition (Shape)	50	50	50	50
Scaled Sequence	k-means (Euclidean)	1.4	0.65	0.4	0.25
	k-medoids (Euclidean)	1.5	1.1	0.65	0.65
	Partition	50	50	50	50
	Partition (Shape)	50	50	50	50

2.3.3 Clustering performance improves with increasing motif information content

To investigate how the information content (IC) of the TF binding motif impacts clustering performance, we generated three synthetic datasets. The first had low IC, that is, each column of the motif had IC that was between 0.8 bits and 1.2 bits. Medium IC motifs had columns whose IC was between 1.6 bits and 1.8 bits. Finally, high IC was a column with IC greater than 1.8 bits. As we would expect, higher IC in the motif led to lower percent classification error when we used both sequence and scaled sequence features (Table 5). Higher IC in the motif used to generate the sequence resulted in the two classes being more reliably different from each other and were easier to partition correctly. Classification performance was invariably poor over all levels of motif IC when we used the shape features because the motif used has no effect on these.

Table 5 Percent classification error decreases with motif information content (low, medium, or high) in simulated datasets with two classes.

Method	Features	Low	Medium	High
Shape	*k*-means (Euclidean)	49.5	48.8	49.8
	k-medoids (Euclidean)	48.45	48.45	48.7
	Partition	49.65	49.95	50
	Partition (Shape)	49.8	47.65	49.8
Sequence	*k*-means (Euclidean)	49.95	15.25	1.9
	k-means (Correlation)	48.05	10.7	1.95
	k-medoids (Euclidean)	47.55	15.2	4.6
	k-mediods (Correlation)	48.65	17.7	7.5
	Partition	50	50	50
	Partition (Shape)	50	50	50
Scaled Sequence	*k*-means (Euclidean)	49.95	14.65	1.4
	k-medoids (Euclidean)	49.4	14.55	1.5
	Partition	50	50	50
	Partition (Shape)	50	50	50

We also investigated how clustering performance changed with sequencing coverage in simulated datasets of co-localized peaks (Table 6). Generally, there was higher classification error in the low coverage datasets, particularly for the scaled sequence features, underscoring the need to for sufficient coverage for these approaches to be useful in analyzing a dataset. Overall, k-means with Euclidean distance is fast and performed well, so we used it with scaled sequence features (with a *k*-mer length of 5) for all subsequent experiments.

Table 6 Percent classification error decreases with increasing peak coverage.

Method	Features	$f = 5$	$f = 10$	$f = 50$	$f = 100$	$f = 500$	$f = 1000$
Shape	k-means (Euclidean)	48.5	49.9	49.15	49.75	49.4	49.8
	k-medoids (Euclidean)	49.7	49.05	49.5	49.3	49.9	48.7
	Partition	48.75	48.8	49.1	49.75	48.5	50
	Partition (Shape)	48.75	48.3	48.2	48.75	49.6	49.8
Sequence	k-means (Euclidean)	2.1	2.25	2.8	2.3	2.2	1.9
	k-means (Correlation)	1.9	2.35	3.1	2.05	1.7	1.95
	k-medoids (Euclidean)	3.15	3.55	2.85	4.3	3.1	4.6
	k-medoids (Correlation)	5.7	6.4	5.6	6.05	3.85	7.5
	Partition	50	50	50	50	50	50
	Partition (Shape)	50	50	50	50	50	50
Scaled Sequence	k-means (Euclidean)	44.15	35.05	5.1	1.95	1.5	1.4
	k-medoids (Euclidean)	43.75	34.85	4.9	1.7	1.45	1.5
	Partition	50	50	50	50	50	50
	Partition (Shape)	50	50	50	50	50	50

2.3.4 Scaled sequence features are sufficient to distinguish among different transcription factor binding sites

Next, we wanted to test how our scaled sequence features performed in partitioning real human ChIP-seq peaks for different transcription factors that had been artificially pooled together. This is a much more difficult problem because an in vivo ChIP-seq experiment will typically produce a mixed sample of peaks, including those where the precipitated DNA was directly bound to the targeted transcription factor, indirectly bound through a co-binding transcription factor or complex of proteins, or even connected to the primary binding site through a DNA loop. We downloaded human ENCODE ChIP-seq datasets for GABPA, GATA2, and YY1 in K562 cells. These represent a diverse sample of TF families, each with different canonical DNA-binding

motifs (Figure 2). These TFs were selected so that they have motifs with different degrees of similarity. All share the short pattern AAG, but the similarity of YY1 and GABPA motifs extends further to the TGGC motif with one base insertion ("A") in the YY1 motif. For the purposes of combining datasets from different experiments, we scaled all read counts such that each dataset would have a total of 20 million reads.

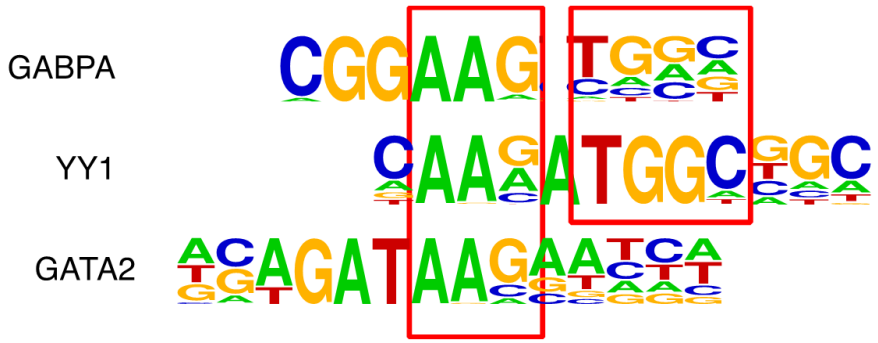

Figure 2 JASPAR motifs for GABPA, GATA2, and YY1 transcription factors.

First, we tested how well scaled sequence features with a *k*-mer length of 5 and *k*-means clustering were able to partition pairs of these peak datasets into their true classes (Figure 3A, Columns 1-3). The number of clusters was set to the true number of classes. For the GABPA and YY1 datatsets, cluster 1 contained a majority of the YY1 peaks and cluster 2 was mostly GABPA peaks. Still, each cluster contained a fair number of the other TF's peaks. This is probably due to the more extended similarity of the canonical binding motifs for these two proteins, which makes it more difficult to distinguish between these TFs. The other two pair separated much more cleanly. Between GABPA and GATA2, Cluster 1 contained a majority of the GABPA peaks, plus very few GATA2 peaks. Cluster 2 was almost exclusively GATA2 peaks, with some GABPA peaks included as well. Finally, GATA2 and YY1 peaks separated almost exactly according to TF. For these data we would not expect to see a perfect partitioning because each TF class likely

starts off as a mix of different types of peaks. Therefore, these results show that the scaled k-mer features are able to capture differences between the sequences and shapes of the ChIP-seq peaks of two different transcription factors, and are likely useful in exploratory analysis of ChIP-seq datasets that contain two classes of peaks.

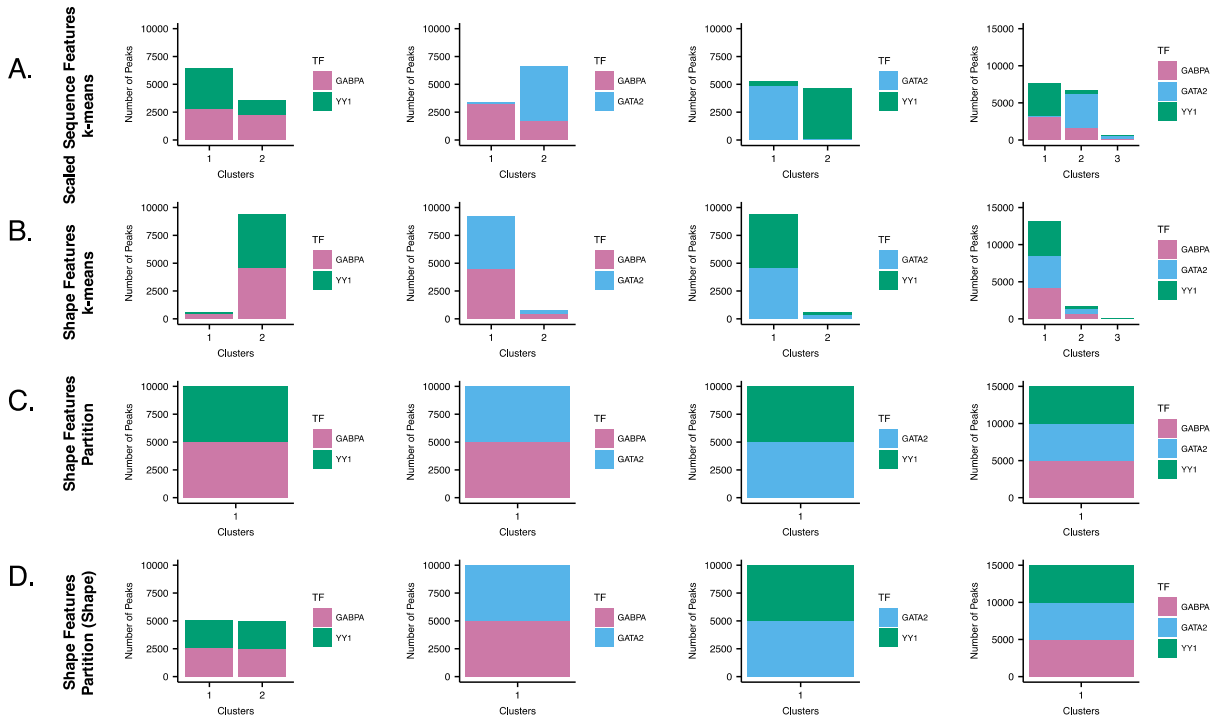

Figure 3 (A) Scaled sequence features and k-means clustering can separate artificially pooled ChIP-seq peaks according to transcription factor. The method performed well in separating pairs of datasets, including GABPA and YY1 (column 1), GABPA and GATA2 (column 2), and GATA2 and YY1 (column 3). It was also able to distinguish among TFs when all three of GABPA, GATA2, and YY1 were pooled together (column 4), although less cleanly. This combination of features and clustering method far outperformed (B) shape features with k-means clustering, (C) the partition method, and (D) partition with shape only method, which could not distinguish between or among TFs.

To further probe the usefulness of this approach, we wanted to see if our method could separate three classes of TF ChIP-seq peaks when they were all pooled together. We combined

all the GABPA, GATA2, and YY1 peaks resulting in a dataset of 15,000 peaks that we partitioned into three clusters (Figure 3A, Column 4). The results showed some interesting relationships between the transcription factor peaks. As we saw when only looking at the pairs, GABPA and YY1 were the most difficult to separate. They were mostly grouped together into cluster 1. The GATA2 peaks were grouped almost exclusively into cluster 2, and shared that cluster with only a few peaks from the other two transcription factors. Cluster 3 contained only a few peaks from each TF. Overall, these results confirm that the scaled sequence features with k-means clustering are not just separating the peaks randomly; rather, they are being grouped together in biologically meaningful ways.

We additionally wanted to compare our scaled sequence features to shape features in these real ChIP-seq datasets. We calculated shape features for the ENCODE datasets and artificially pooled the peaks for the different TFs as before. Then we clustered the shape features with k-means (Figure 3B), the partition method (Figure 3C), and the partition method with shape only (Figure 3D). In all cases the shape features were not adequate to distinguish between the different TFs. In almost all cases, the peaks for both or all three TFs were assigned to a single cluster. k-means with shape features created two or three clusters, but one cluster contained nearly all the peaks and the other(s) contained very few peaks and represented all the TFs. Shape features with the shape only partition method did find two evenly sized clusters in the case of GABPA and YY1, but peaks for the two TFs were distributed nearly randomly between the two clusters. These results further confirm that shape information alone is not adequate to separate ChIP-seq peaks with different TF-binding motifs.

2.3.5 Peak partitioning in a Drosophila Scute ChIP-nexus dataset enriches for E-box motifs

We were able to show that our approach is able to separate ChIP-seq datasets that have been artificially pooled together. However, this is obviously not very useful for real-world situations in which we do not already understand the subclasses present in a sample. Thus, we wanted apply our partitioning approach to a high-resolution ChIP sequencing data, so we performed experiments on *Drosophila* embryos with the ChIP-nexus protocol (9). Since basic-helix-loop-helix (bHLH) transcription factors bind as heterodimers with different partners and have remarkable binding specificity (9, 19), we analyzed the occupancy of the bHLH transcription factor Scute. Together with achaete, Scute is a highly conserved transcription factor that confers neural identity from *Drosophila* to mammals (20). To obtain a large homogenous population of suitable cells, we collected *Drosophila* embryos from maternal mutants that consist entirely of neuronal precursors (Tlrm9/Tlrm10) (21).

When we applied our scaled sequence feature partitioning method on the Scute ChIP-nexus dataset, 5,000 ChIP-nexus peaks were partitioned into 2 clusters. The analysis of both replicates showed similar results. The peaks were split nearly evenly into clusters for both replicates (52.3% of peaks in cluster 1 for replicate 1 and 53.2% in cluster 1 for replicate 2). When we used HOMER to search for enriched motifs in each cluster for both replicates, we found a very close agreement in the motifs identified in the corresponding clusters across replicates (Figure 4). There were four motifs (red, green, blue, and pink) that were found in all clusters in both replicates. The DREF motif is frequently associated with promoters, where it is bound by DNA-replication factor (DREF) (22). But the same motif is also bound by BEAF-32, an insulator or architectural protein (blue). A motif that closely matches CTCF (maroon) was found in cluster 1 but not cluster 2, which

suggests that these peaks might have some functional connection to insulators. Interestingly, an E-box motif (red), consistent with the known binding specificities of Scute or related bHLH factors in mammals CAGCTG (23), was identified in all clusters, but the motifs differed slightly in the middle two positions (the dinucleotide sequence between CA and TG). Cluster 2 had a GC in these positions with high information content (for both replicates). Cluster 1, however, had lower information content at these positions with almost equal probability of either G or C at each, indicating that they may be low-affinity binding sites.

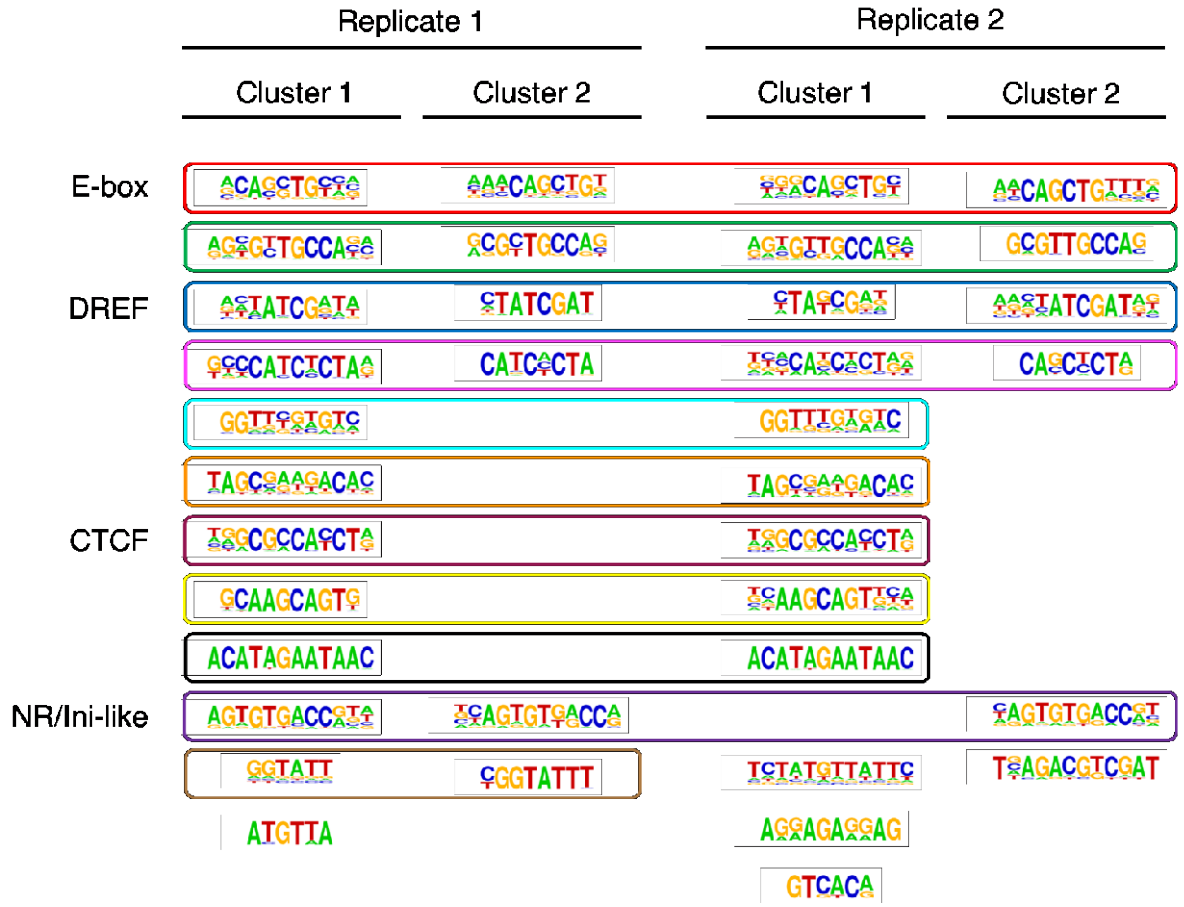

Figure 4 Similar motifs were found in the same clusters for each Scute ChIP-nexus replicate. Matching motifs are grouped by color. If there was a close match for a known TF, the motif is labeled.

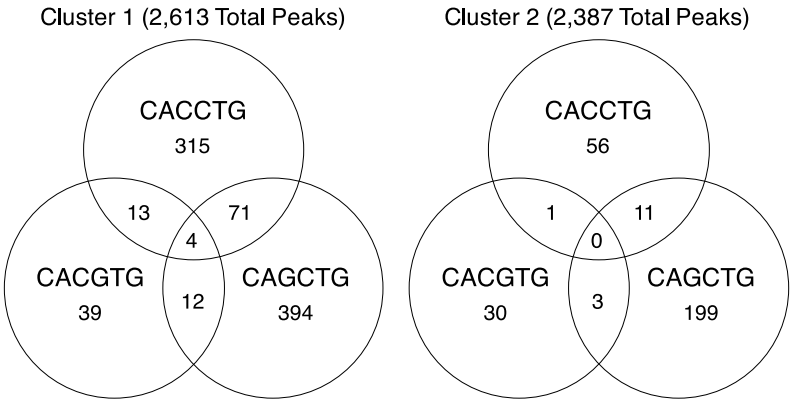

Figure 5 Number of *Drosophila* Scute ChIP-nexus peaks (replicate 1) that contain E-box-related motifs for each cluster.

When we checked the actual sequences at the peaks in each cluster, we found that there were 403 peaks that contained the string CACCTG (or its reverse complement CAGGTG) in cluster 1 compared to only 68 peaks in cluster 2 (Figure 5). There were very similar results for replicate 2 (Figure 6). Additionally, cluster 1 contained more than twice as many peaks with the string CAGCTG compared to cluster 2 (481 peaks versus 213). Thus, cluster 2 appears to contain more high-affinity binding sites, while cluster 1 contains additional binding motifs such as CTCF. Interestingly, the canonical motif CACGTG appears in the least number of peaks in both replicas (68 and 34, respectively) compared to the non-canonical ones.

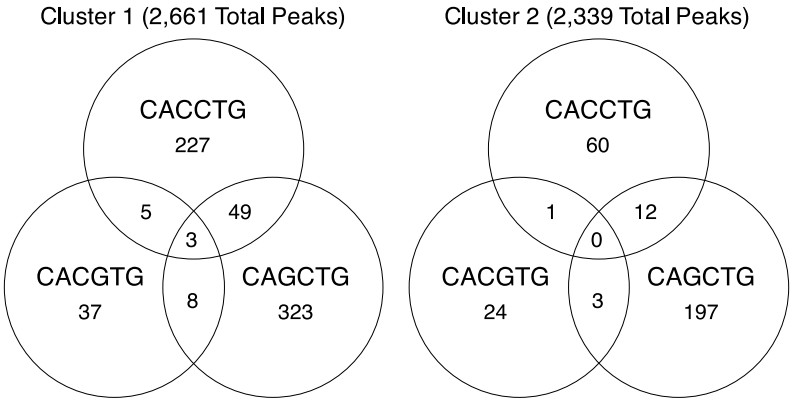

Figure 6 Number of *Drosophila* Scute ChIP-nexus peaks (replicate 2) that contain E-box-related motifs for each cluster.

Finally, we looked at the genomic annotations of the peaks in each cluster. The majority of the Scute peaks covered promoter and TSS regions (Figure 7). However, cluster 1 contained the majority of peaks that mapped to introns (377 peaks in cluster 1 compared to 56 in cluster 2 for replicate 1 and, for replicate 2, 306 peaks in cluster 1 and 89 peaks in cluster 2.) Taken together, these results suggest that there could be biologically meaningful differences in function between the Scute binding sites in each of the clusters we discovered.

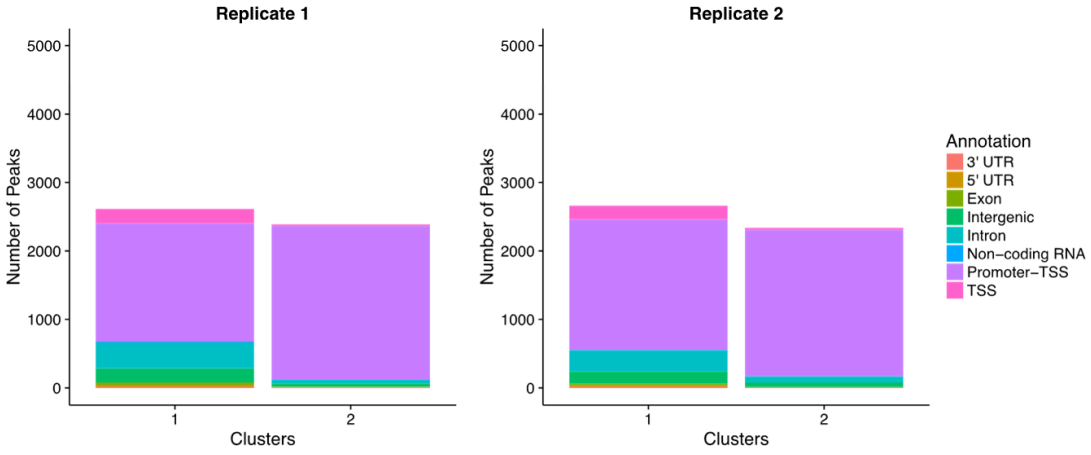

Figure 7 Scute peak annotation by cluster

2.3.6 Non-canonical Scute sites coincide with insulator protein CTCF binding sites in cluster 1 but not in cluster 2

Given that CTCF motifs were only found in cluster 1 of the Scute ChIP-nexus data, we further investigated the potential interactions between CTCF and Scute. We tested whether the predicted CTCF motifs were real and whether they co-occur in the proximity of the Scute E-boxes. For this reason, we analyzed a CTCF-GFP ChIP-seq dataset from the modERN project (see Materials and Methods). We found that CTCF ChIP-seq peaks are overrepresented in the ChIP-

nexus peaks of cluster 1 (Fisher's exact test p-value=10-34), but not in cluster 2. Furthermore, the CTCF peaks are co-occurring with Scute bound peaks in cluster-1 that contain the non-canonical E-box motifs CAGCTG (p-value=10-26) and CACCTG (p-value=10-8), but not with the other canonical motif CACGTG (p-value=0.4). Co-occurrence of CTCF binding with E-box motifs in cluster 2 was non-significant (p-values 0.1 to 0.9).

This shows that using the quantitative *k*-mer method we can partition the ChIP-nexus peaks into subsets with distinct biological functions. Furthermore, given that the peaks were ~100 bp long, our findings indicate a possible physical interaction between Scute binding to non-canonical E-boxes and the CTCF insulator protein.

2.4 Conclusions

Identifying the TF binding subclasses on different regulatory regions and the corresponding binding preferences is a very important step in understanding gene regulation in a cell. In this paper, we evaluated various methods for partitioning peaks of TF chromatin bound regions using peak shape and sequencing depth features. We found that quantitative *k*-mer sequence features improve separation of co-localized peaks on the basis of the TFs that bind to them. As expected, clustering performance increases with the information content of the motifs of the corresponding TFs.

We applied this method to two genome-wide occupancy *Drosophila* datasets: a ChIP-nexus dataset for Scute. In the ChIP-nexus dataset, we identified two clusters, both containing the expected E-boxes (but with slightly different submotifs). In both clusters, we found the non-canonical motif CAGCTG to be predominant. Non-canonical motif CACCTG was also found in

large number of peaks in cluster 1 but not cluster 2. Finally, the canonical E-box motif CACGTG was also present but only in 8-11% of all peaks. This shows that non-canonical E-boxes may play a more significant role than previously thought in Drosophila development. Another interesting finding is the low number of co-occurrence of these three motifs in the same peak.

Besides E-boxes, we also noted motifs that match other TF proteins. In both clusters, for example, DREF motif was present. DREF regulates cell cycle through p53. Cluster 1 also contained a number of other motifs including CTCF, which were not present in cluster 2. Further analysis of published CTCF ChIP-seq data showed that there is a significant number of co-occurrences of non-canonical E-box motifs and CTCF binding in cluster 1, but not cluster 2. Given the short size of ChIP-nexus peaks (we extended them by 100 bp), it is highly likely that the motif co-occurrence imply physical interaction between Scute and CTCF. There was not significance co-occurrence of the canonical E-box motif and CTCF binding in either cluster.

3.0 Single sample network perturbation assessment

Complex diseases involve perturbation in multiple pathways and a major challenge in clinical genomics is characterizing the heterogeneity of disease samples with respect to the molecular networks involved. The aim is to identify in each patient sample the underlying mechanism of disease thereby improving diagnosis and personalizing treatment. Thus far, the methods that have been developed to address this challenge have a common theme of relying on external databases of pathways to quantify pathway activity scores. The drawback of this approach is that it ignores the dependencies present in the data and that pathways are incomplete and may not accurately represent the specific tissue and disease under study.

We present a new approach, Single Sample Network Perturbation Assessment (ssNPA), for subtyping samples based on deregulation of their gene expression networks. Instead of relying on pathway prior knowledge, our method proceeds by learning a causal graphical model directly from control data. Network neighborhood deregulation of an individual sample can then be quantified via the error incurred in predicting the expression of each gene from its Markov blanket.

We evaluate the performance of ssNPA by assessing class assignment of single samples in several datasets for which the true classes are known. Using a single-cell RNAseq dataset of liver cell development we demonstrate that ssNPA can partition the cells according to embryonic stage and differentiated cell type. We further validate ssNPA on two cancer datasets where we demonstrate that ssNPA-derived clusters show significantly different survival outcomes and correlate with known molecular subtypes. In all analyses ssNPA consistently outperforms alternative approaches, including competing methods that estimate pathway activity from prior information, highlighting the advantage of our network-based approach.

3.1 Background

Gene expression profiling by RNA-sequencing has become routine tool in biomedical research. Similarly, on the clinical side, RNA-seq has now been introduced as a cost-effective diagnostic tool (24, 25). Moreover, recent technological advances have made the assessment of gene expression at single cell level (scRNA-seq) feasible, opening new avenues to developmental biology and the study of dynamic networks (26-28). Consequently, the number of large RNAseq datasets keeps growing with hundreds or thousands of samples representing a single clinical or cellular condition. As a result, the scientific questions have shifted away from simple differential expression to characterizing the molecular heterogeneity of disease phenotypes. One simple way to characterize sample heterogeneity is via clustering and/or dimensionality reduction. This approach will often reveal distinct sample groups within the population but it treats all genes equivalently, ignoring the fact that genes are organized in regulatory networks. On the other end of the spectrum there has been considerable development in methods that quantify pathway activation on a single sample level (ssGSEA (29), PLAGE (30), GSVA (31), Pathifier (32)). However, these methods rely heavily on existing pathway information (e.g., from KEGG, BioCarta, The Nature Pathway Interaction Database), which may be incomplete, not well annotated or irrelevant to the studied phenotype or condition. Other methods (e.g. (33)) quantify a sample-to-sample similarity with the aim of identifying similarities and differences between cell functions.

In this paper, we present a different approach for assessing, qualitatively and quantitatively, how the gene network from a set of control samples is perturbed in a newly presented single sample. Our approach, Single Sample Network Perturbation Assessment (ssNPA), uses causal modelling to first learn the gene expression interaction network from a set of reference samples.

For each new sample the method then assesses which parts of the "reference sample network" are deregulated.

Causal graphs have been used in the past to learn gene networks from expression data (34-36) or gene features that are highly predictive of certain phenotypes (37-41). Our ssNPA approach learns a causal graph from expression data and for every gene it builds a predictive model based on its Markov blanket. Applying the model to a new sample produces a vector of residuals which quantifies the network level gene deregulation (NLGD). The NLGD vectors can then be used to cluster samples into groups and assess their group characteristics (e.g. developmental time, survival, molecular mechanisms of phenotype, etc.) or to assign an individual patient to a disease subcluster. We use this property to evaluate ssNPA on existing datasets, for which the ground truth is known. Specifically, we show that ssNPA separates well the developmental trajectory and the differentiated cell type in a mouse liver cell development scRNA-seq dataset (42). We also use RNA-seq data from The Cancer Genome Atlas (TCGA) (lung and breast cancer datasets) (43, 44) to demonstrate ssNPA can separate the samples in the corresponding datasets according to patient survival and molecular subtypes with better accuracy than alternative approaches.

3.2 Materials and Methods

3.2.1 Liver Cell Development Data

A murine liver cell development scRNA-seq dataset was obtained from (42) (GEO:GSE90047). The experiment measured the gene expression of 447 cells over the course of embryonic days E10.5-E17.5. Cells were first sorted with fluorescence-activated cell sorting

(FACS) according to the cell surface markers Delta-like (DLK) to identify hepatocytes and epithelial cell adhesion molecule (EpCAM) to distinguish cholangiocytes.

3.2.2 TCGA Data

Breast invasive carcinoma (BRCA) and lung adenocarcinoma (LUAD) RNA-seq data from The Cancer Genome Atlas (TCGA) project were downloaded from the Broad Firehose (Broad Institute TCGA Genome Data Analysis Center, 2016). The BRCA dataset consists of 1,100 cancer samples and 112 normal samples (43). The LUAD dataset consists of 517 cancer samples and 59 normal samples (44).

3.2.3 Sample clustering

In order to better evaluate the efficiency of the various methods for single sample subtyping, we performed sample clustering using Seurat (45) and we examined various external characteristics of the clusters. Samples were clustered in their feature space. First, the samples are projected into principal component space. The number of principal components to retain in the projection is determined heuristically by identifying the elbow of the scree plot. Then clustering is performed with a graph-based clustering that constructs the shared nearest neighbour graph and then optimizes the modularity function (46). Finally, the clusters are visualized with a nonlinear dimensionality reduction (t-SNE) (47).

3.2.4 Comparison to other methods

ssNPA methods were compared to the commonly used gene expression-based clustering and to two other comparable algorithms: Pathifier (32) and single-sample gene set enrichment analysis (ssGSEA) (29). All methods were tested on the same input data and reference sample selections. For Pathifier, we provided gene lists for all KEGG pathways and used the R implementation with the quantify_pathways_deregulation() function and default parameters. For ssGSEA we used the gene sets from the C2 collection of the Molecular Signatures Database version 3.0 (48) provided in the GSVAdata R package and the implementation of ssGSEA provided within the GSVA() function of the GSVA R package with default parameters. For fairness, we use an equal number of principal components for clustering with each method. The number of principal components is set to the maximum number of principal components identified by the elbow of the scree plot for any single method.

3.2.5 Software availability

An R software package has been developed and will be freely available upon the publication of the paper.

3.3 Results

3.3.1 ssNPA algorithm description

ssNPA learns the global gene expression network as a directed (causal) graph from a set of reference samples using FGES (49). FGES calculates a directed acyclic graph (DAG) over all data by maximizing the Bayesian Information Criterion (BIC) score of the data given the model (network). The BIC score is given by the formula:

$$BIC = -2 \cdot \mathcal{L}(\mathcal{D}) + PD \cdot df \cdot \ln n$$

where $\mathcal{L}(\mathcal{D}) = \ln P(\mathcal{D}|\theta, \mathcal{M})$ is the maximum log-likelihood of the data given the model and its parameters; PD is a penalty value ("penalty discount") that controls sparsity (PD=1 in the standard BIC definition); df is the degrees of freedom; and n is the sample size. This score is decomposable, and the total BIC of the graph is the sum of the BIC of its nodes and their parents.

FGES starts with an empty graph then adds single edges while the BIC score increases. Next, the algorithm removes single edges while the BIC score increases.

The Markov blanket of a gene G_i, MB(G_i), consists of the parents, children and spouses of G_i in the graph. Once the graph has been learned from the reference (control) samples, then ssNPA uses the Markov blanket around each gene, G_i, to build a predictor of its expression. This is because in a directed graph:

$$Ind(G_i, X|MB(G_i)), \text{ for every } X \notin MB(G_i).$$

Therefore, a highly predictive regression model can be learned for each gene:

$$G_i = \beta_{0,i} + \sum_{G_k \in MB(G_i)} \beta_{k,i} \cdot G_k + \varepsilon$$

Then for each new sample this model can be used to calculate the deviation of the expression of G_i in this sample compared to the reference samples. So, the new sample can be represented as a vector of deviations of expression of every gene from the reference samples. Given that genes are connected through the network of interactions, in this way, we assess both the topology and the magnitude of network perturbations. This procedure is summarized in Figure 8.

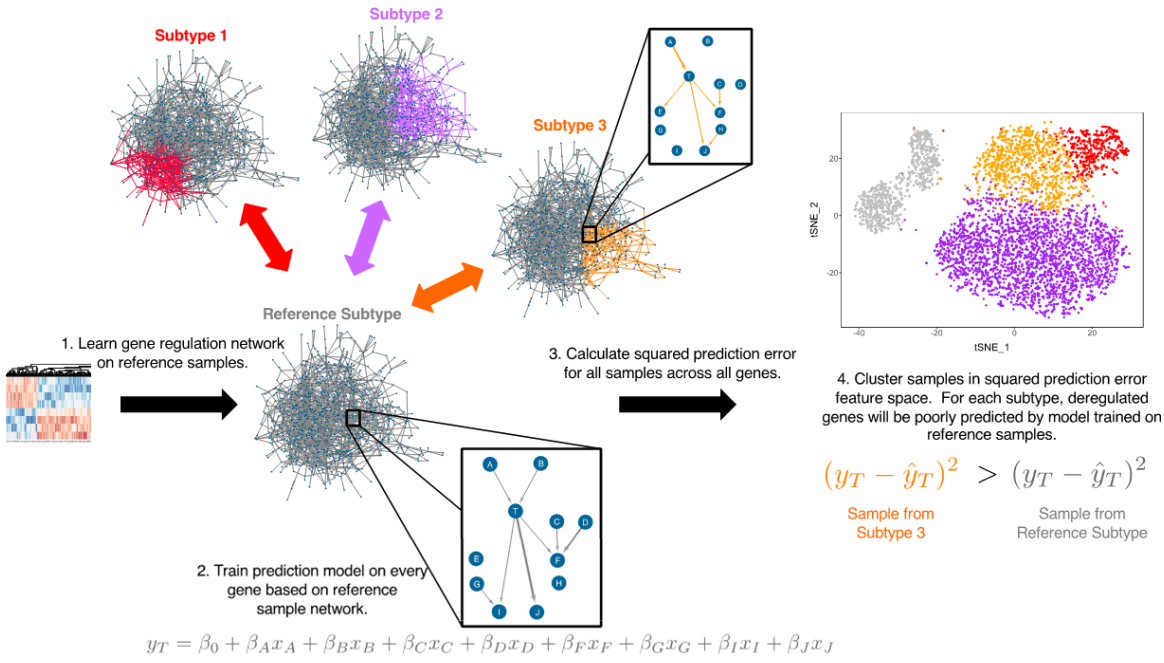

Figure 8 Overview of the single-sample network perturbation assessment through causal network (ssNPA) algorithm.

For comparison purposes, we also implemented ssNPA-LR, in which causal learning is substituted by lasso regression, resulting in an undirected graph. The ssNPA analysis procedure has the following steps:

1. *Data preparation.* For speed and accuracy, in this paper we selected the top 3,000 most variant genes for scRNA-seq or RNA-seq data. The RNA-seq counts were transformed to log2 counts per million through mean-variance modeling by the voom function (Limma v. 3.32.10) (50).

2. *Reference samples.* For disease data, we used the controls as reference sample set. For the liver scRNA-seq, we tested each stage (as determined by external cell markers) as potential reference group.

3. *Gene network learning (ssNPA).* A directed graph was learned on the expression data for the reference group of samples (FGES algorithm) (49). For this work, we scan over a number of PD values in the range [4, 12] and we choose a PD for each dataset that balances grouping the reference samples together while not overfitting. The Markov blanket around every gene in this network is used for predicting its expression on any given sample; and deviation from the observed value is a measure of network perturbation.

4. *Feature selection (ssNPA-LR).* In this case, we used the glmnet package in R (v. 2.0.16) to learn a lasso regression prediction model for every gene across the reference samples (51).

We chose each sparsity parameter (λ) with 10-fold cross validation, selecting the value of λ corresponding to minimum mean cross-validated error.

3.3.2 ssNPA correctly identifies embryonic stage and cell type in murine liver cells from single cell RNA-seq data

We used a recently published liver development scRNA-seq dataset to test ssNPA and compare it to other methods. This dataset is composed of multiple types of liver cells samples at a series of developmental timepoints. The early hepatoblast cell differentiates into two lineages (hepatoblasts and cholangiocytes). In this dataset the time point and cell-identity is experimentally controlled and thus can serve as the ground truth. We hypothesize that information regarding the cell-type and developmental stage is reflected in the gene expression data.

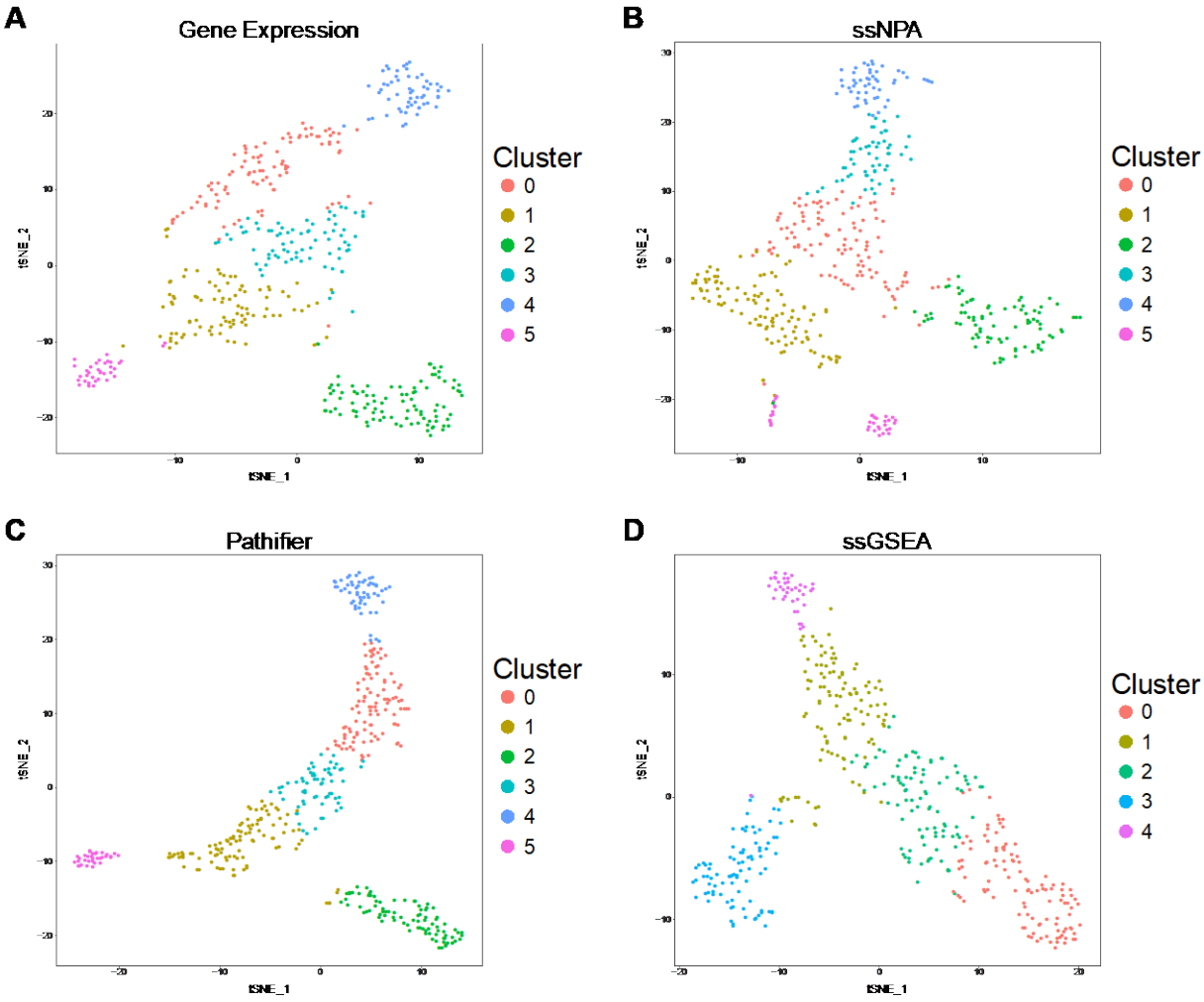

Figure 9 Cluster assignments with (A) gene expression, (B) ssNPA, (C) Pathifier, and (D) ssGSEA of murine liver cell scRNA-seq samples. ssNPA was used with PD=5 and the E14.5 cells were provided as the reference set to both ssNPA and Pathifier. Clustering for all methods was performed with the first 10 principal components.

When gene expression data used directly for clustering (46) we identified six clusters (Figure 9A), which separated well the extreme developmental time points: cells measured at day E10.5 and hepatocytes from day E17.5 (Figure 10A). However, all of the differentiated cholangiocytes were grouped together in a single cluster and although they were somewhat stratified within the cluster, their embryonic stage was not distinguishable. The remaining clusters

contained a mix of intermediate timepoints (days), and hence they did not accurately represent the developmental trajectory.

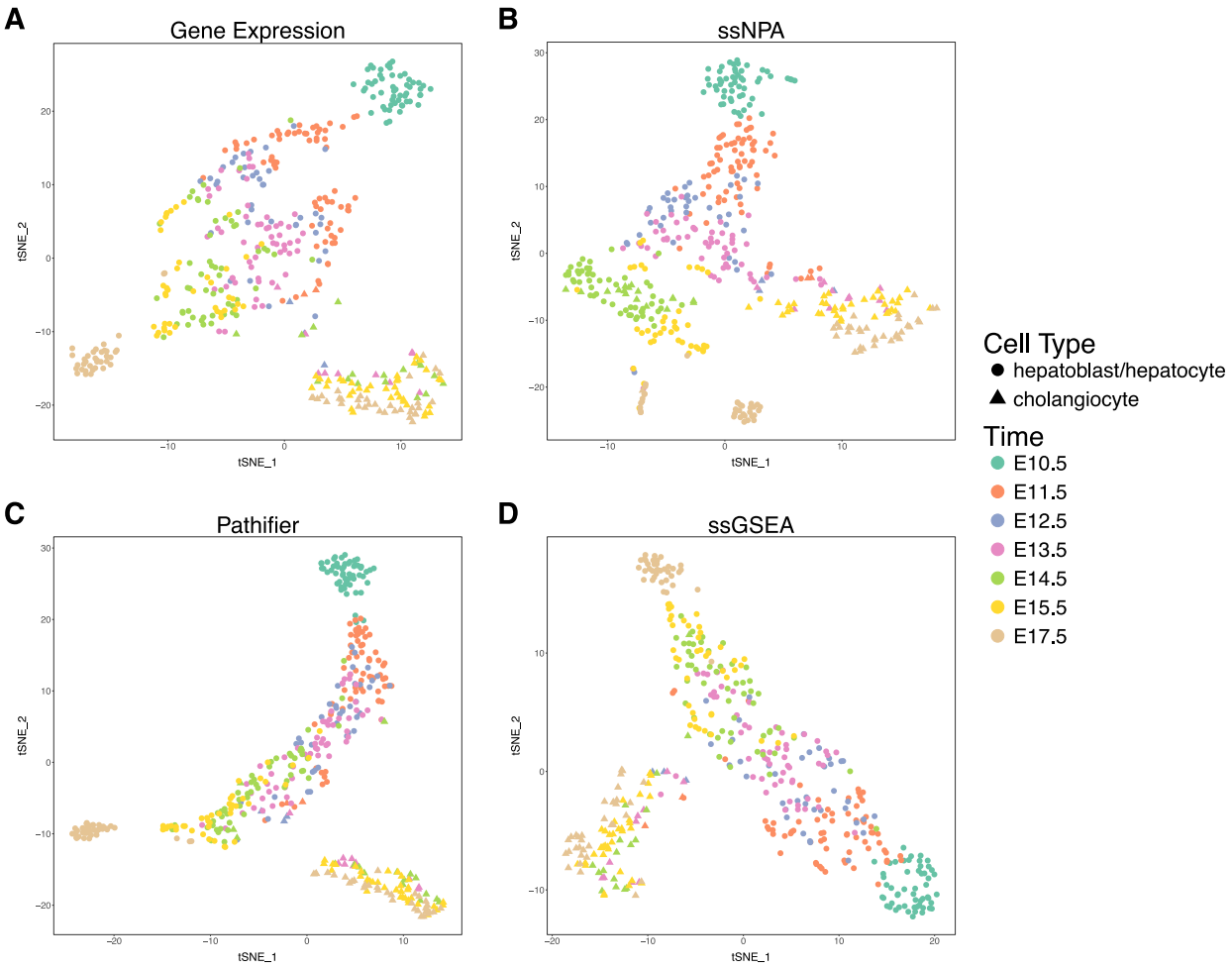

Figure 10 Comparison of how well (A) gene expression, (B) ssNPA, (C) Pathifier, and (D) ssGSEA separate murine liver cell scRNA-seq samples by developmental stage and cell type. ssNPA was used with the E14.5 cells as the reference set and PD=5. Pathifier was also applied with the E14.5 cells as the reference set. Clustering for every method was performed with the first ten principal components.

By contrast, the six identified clusters based on the network perturbation features of ssNPA (Figure 9B) separated well all stages (Figure 10B). In particular, hepatoblasts from days E10.5 and E11.5 as well as mature cholangiocytes and E17.5 hepatocytes were separated into four distinct

clusters. Since there was not an obvious reference set of samples in this dataset, we examined the utility of each group as a potential reference. We additionally evaluated a range of PD (penalty discount) parameter values for FGES [4,12]. We found the late intermediate stages (E14.5 or E15.5) to show better performance than the extremes, when they were used as reference set, while the choice of PD had less impact on clustering performance (Figure 11).

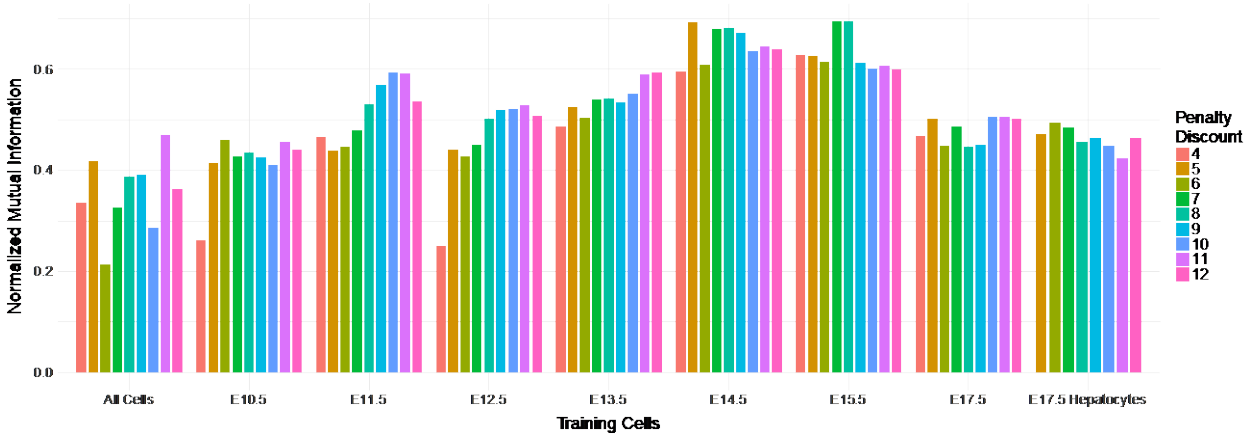

Figure 11 ssNPA clustering performance assessed by normalized mutual information measures how well murine liver cells are separated by their developmental stage and cell type and is a function of both the penalty discount (PD) parameter chosen for learning the reference causal network with FGES and the cells chosen as the reference set on which to learn the reference network. Clustering for all methods was performed with the first ten principal components. Time point E14.5 with PD=5 was chosen for learning the reference network in all subsequent ssNPA analyses of this dataset.

Next, we compared ssNPA to Pathifier and ssGSEA. Both methods quantify gene interactions, but require pathway information from an external database. For Pathifier we used the KEGG pathway database (52). It also identified six distinct clusters (Figure 9C), but with the exception of E10.5, it did not separate the developmental stages very well (Figure 10C). All of the intermediate stage hepatocytes were mixed together and distributed in three clusters. Furthermore, the runtime of Pathifier was very long compared to ssNPA (on the order of hours compared to

minutes), Finally, we tested ssGSEA which calculates a gene set enrichment score for every sample. We used ssGSEA with default parameters and the default gene sets from the C2 collection of the Molecular Signatures Database version 3.0 (48). ssGSEA does not require the user to provide a reference set. Clustering with the ssGSEA produced five clusters (Figure 9D), but in general, these were not well separated according to developmental time point (Figure 10D). The cholangiocytes were grouped into one cluster but the largest cluster contained the hepatoblasts from E10.5 and E11.5. The hepatocytes from E17.5 were grouped together well in another, but the remaining hepatoblasts/hepatocytes spanning E12.5-E15.5 were mixed together and divided between two clusters.

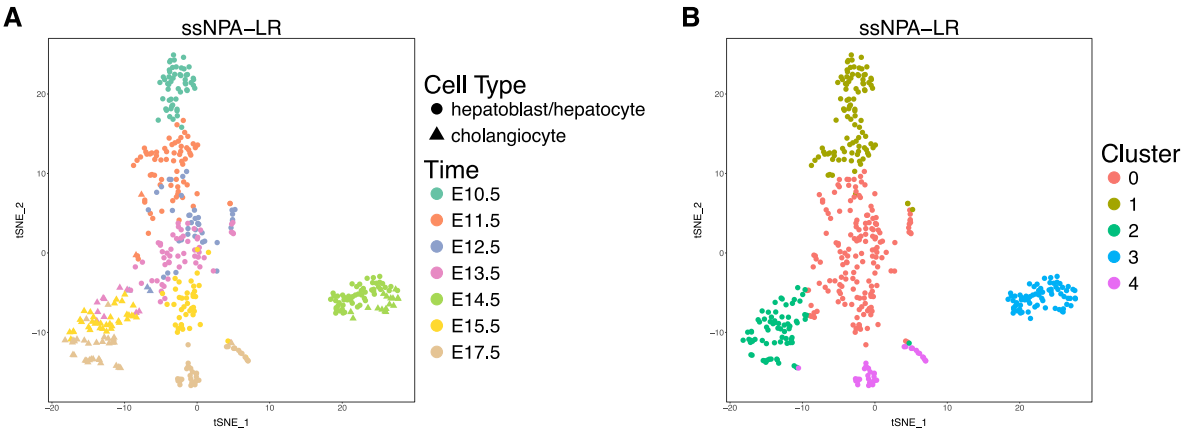

Figure 12 (A) Developmental stage and cell type separation and (B) cluster assignment with ssNPA-LR on a murine liver cell scRNA-seq dataset. We chose E14.5 was the reference group of cells to facilitate comparison with ssNPA. Sparsity parameters (λ) for the lasso regression models were chosen with 10-fold cross validation, selecting the value of λ corresponding to the minimum cross-validated error. Clustering was performed with the first ten principal components.

We additionally developed and tested a variation of ssNPA, the ssNPA-LR algorithm, which uses lasso regression instead of causal learning to choose the features predicting the expression of a gene (Figure 12A). We found five clusters (Figure 12B), which separated well the

early and late developmental stages, but most of the intermediate stage hepatocytes were grouped together in one cluster. The exception to this is the group of E14.5 hepatocyte and cholangiocyte cells, which are grouped tightly together at a large distance from the rest of the developmental trajectory. This suggests that the lasso regression approach might be overfitting to the reference group of cells despite the fact that its optimum parameter was selected through cross-validation to minimize mean error.

Table 7 Comparison of different feature calculation methods. Clustering for every method was performed with the first ten principal components. E14.5 were used as the reference cells for Pathifier, ssNPA, and ssNPA-LR. PD=5 for ssNPA. MI: mutual information; ARI: adjusted Rand index.

Method	Normalized MI	ARI	Avg number of features
ssNPA	**0.693**	**0.500**	**2.5**
Gene expr. (all genes)	0.571	0.367	NA
Gene expr. (top 3,000)	0.565	0.369	NA
Pathifier	0.585	0.399	NA
ssGSEA	0.540	0.343	NA
ssNPA-LR	0.687	0.427	23.1

To quantitatively compare the clustering performances of all methods we used the normalized mutual information (NMI) and the adjusted Rand index (ARI) (Table 7). We found that ssNPA and ssNPA-LR clearly outperform Pathifier, ssGSEA, and gene expression (with either the top 3,000 most highly variant genes or all genes) by maximizing NMI (0.693 and 0.687, respectively). ssNPA also returned the highest ARI of these methods (0.5). However, we note a strong advantage to ssNPA over ssNPA-LR when we consider how many genes they utilized. On average, ssNPA used only 2.5 predictors for every gene, while ssNPA-LR needed 23.1 genes.

Thus, we see that using directed graphs to jointly model the expression of all 3,000 genes offers a clear advantage.

3.3.3 ssNPA separates breast cancer samples according to molecular subtype and shows significant differences in survival

We also applied various methods on breast cancer RNA-seq data from tissues of known subtype. ssNPA-based clustering placed the majority of the basal tumor samples (93.1%) in two clusters along with most of the HER2+ samples (89.2%) (Figure 13A and Figure 14A), while Luminal A and B subtypes were not resolved that well. The other methods produced a similar result with respect to molecular subtype clustering (Figure 13B-Figure 13D and Figure 14B-Figure 14D). The identified clusters of patients differ significantly in terms of survival for all methods (Figure 15), although ssNPA-derived clusters had more significant p-value.

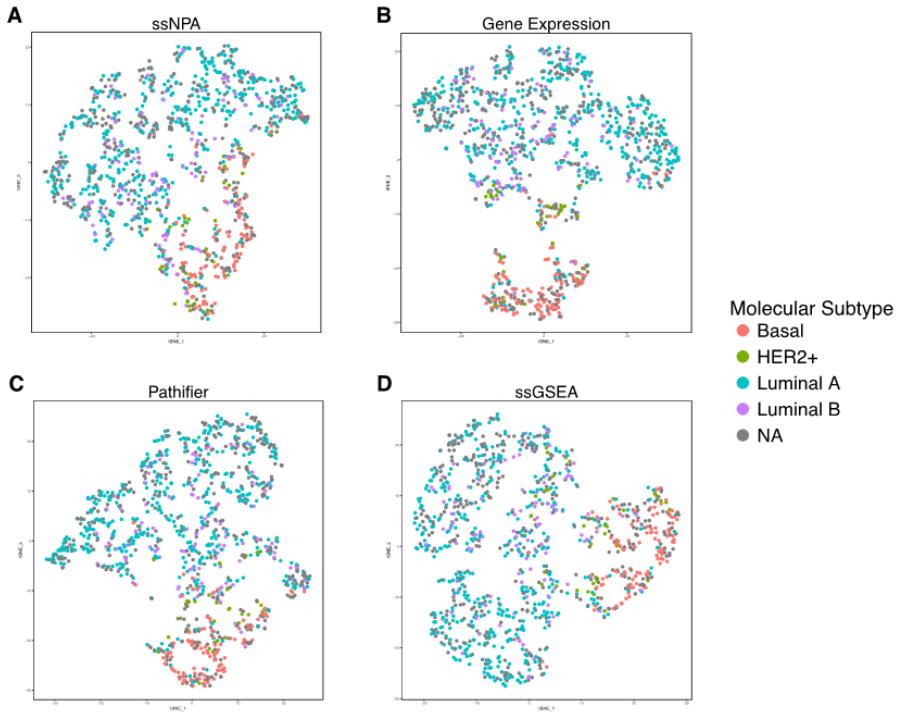

Figure 13 Separation of breast cancer RNA-seq samples according to tumor molecular subtype by (A) ssNPA, (B) gene expression, (C) Pathifier, and (D) ssGSEA. ssNPA was used with PD=8. Clustering for all methods was performed with the first three principal components. Molecular subtype was assigned according to estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) status. We define ER-negative, PR-negative, and HER2-negative as basal (triple-negative); ER-negative, PR-negative, and HER2-positive as HER2+; ER-positive, PR-positive, and HER2-negative as luminal A; and ER-positive, PR-positive, and HER2-positive as luminal B.

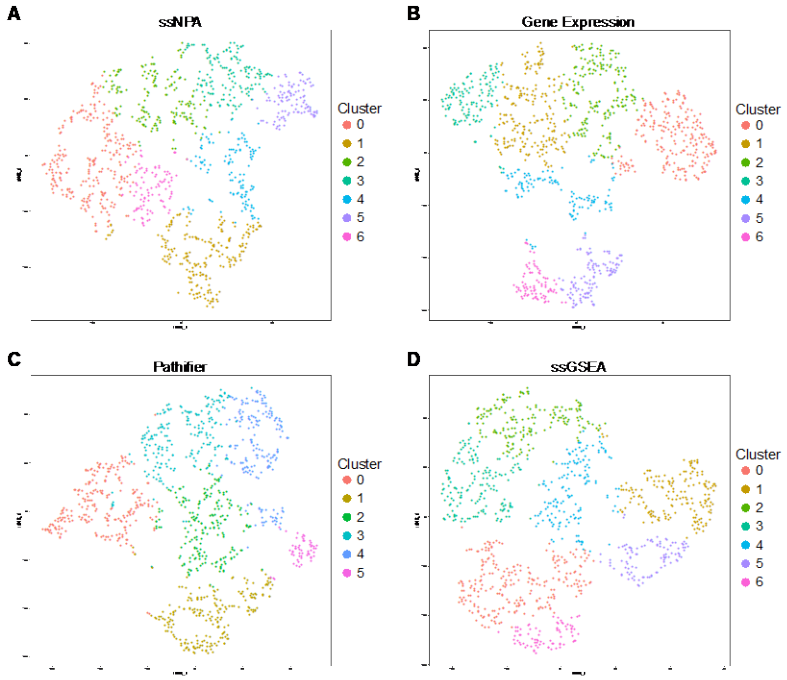

Figure 14 Cluster assignments with (A) ssNPA, (B) gene expression, (C) Pathifier, and (D) ssGSEA of breast cancer RNA-seq samples. ssNPA was used with PD=8. Clustering for all methods was performed with the first three principal components.

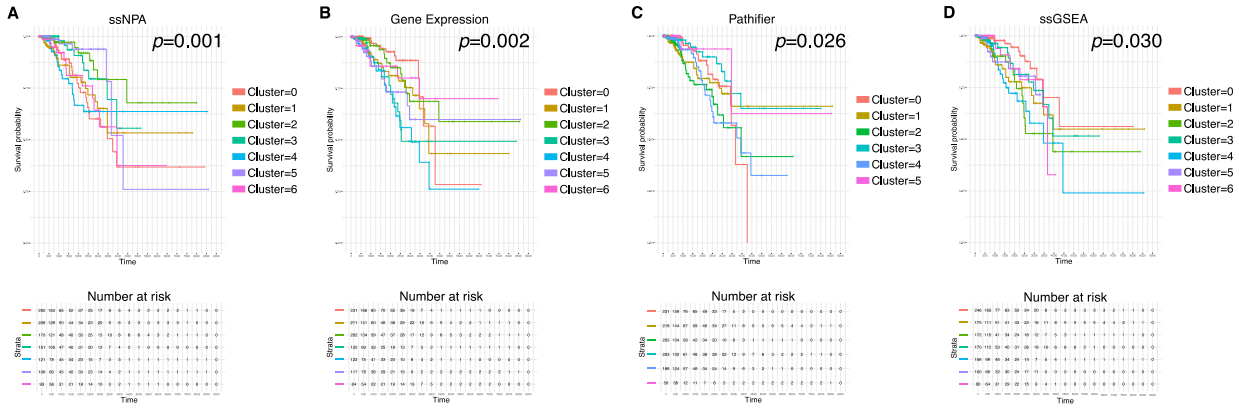

Figure 15 Full breast cancer subject survival analysis by cluster as assigned with (A) ssNPA, (B) gene expression, (C) Pathifier, and (D) ssGSEA. ssNPA was used with PD=8. Clustering for all methods was performed with the first three principal components.

3.3.4 ssNPA identifies two triple negative subclusters with different survival rates

We have demonstrated that ssNPA is able to use relative gene network deregulation separate BRCA tumor samples in a clinically meaningful way. We next investigated if ssNPA can provide further biological insight by inspecting the ssNPA-derived subclusters of the most severe, triple-negative tumors. To this end, we used ssNPA to cluster the 116 triple negative tumor samples, using the 37 HER2+ samples as the reference group. The algorithm identified two distinct subclusters of triple negative patients (63 in clusters 0 and 53 in cluster 1), which we found to differ in terms of survival (p=0.056). In particular, cluster 1 has substantially better survival up to 2.5 years than cluster 0 (Figure 16).

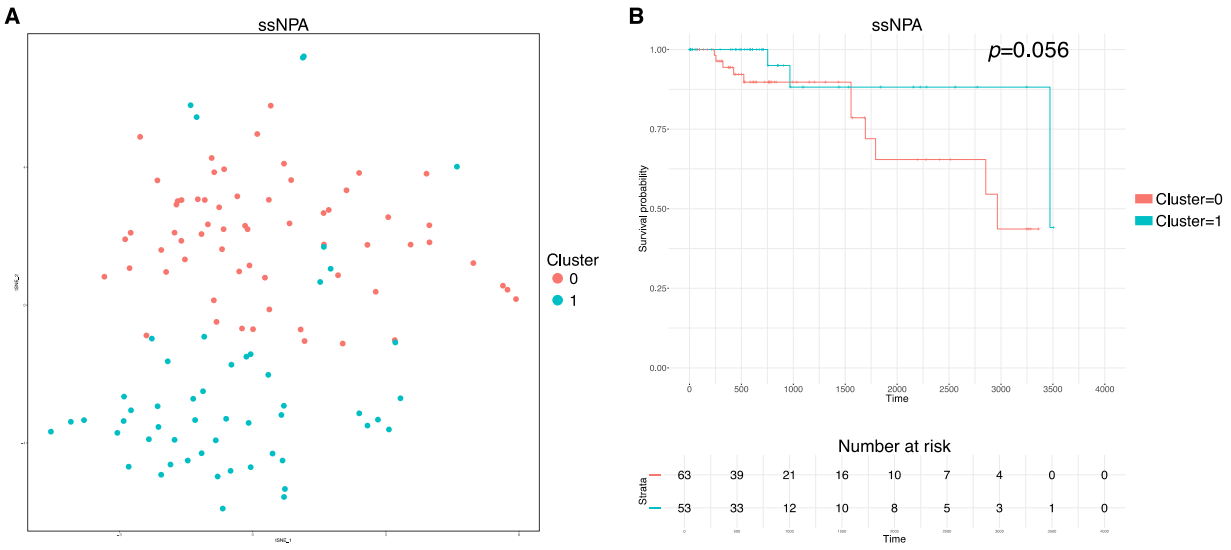

Figure 16 (A) Basal breast cancer patient subclusters. (B) Basal breast cancer patient survival by subcluster. Patients were clustered with ssNPA with the HER2+ patients provided as the reference group and PD=8. The first ten principal components were used for clustering. (B) Survival plot of the patients in the two ssNPA clusters (x-axis: time in days).

In order to investigate which genes contributed to cluster identification we used the magnitude of the PCA loadings for the principal components used in clustering (Table 8). These

genes with the highest loadings are the ones whose network is most deregulated when compared to HER2+ samples (Figure 17A). Many of these genes have well-documented roles in breast cancer malignancy and progression.

Table 8 The genes with the top 5 loadings of the first ten PCs in the BRCA triple-negative dataset. ssNPA was used with PD=8.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
MSLN	0.18	0.20		0.31		0.59	0.25			
MUCL1	0.25				0.21			0.36		0.42
COL9A1				0.29			0.23	0.30		0.18
CSN3		0.20	0.34		0.33					
C4orf40			0.23		0.3	0.22				
PIP	0.22		0.24						0.20	
GAGE12D				0.20		0.2	0.21			
KRT5								0.17	0.25	0.17
KRT13				0.20		0.18	0.19			
LALBA			0.24		0.29					
COL2A1					0.2				0.26	
TFAP2B	0.16					0.22				
DSG1									0.32	
PI3									0.26	
SMR3B			0.24							
FABP4		0.22								
OLFM4				0.22						
ADIPOQ		0.20								
CRISP3	0.18									
EHF		0.18								
C4orf7							0.18			
DMBT1								0.17		
FABP7										0.17
LBP										0.16
CSN1S1								0.15		

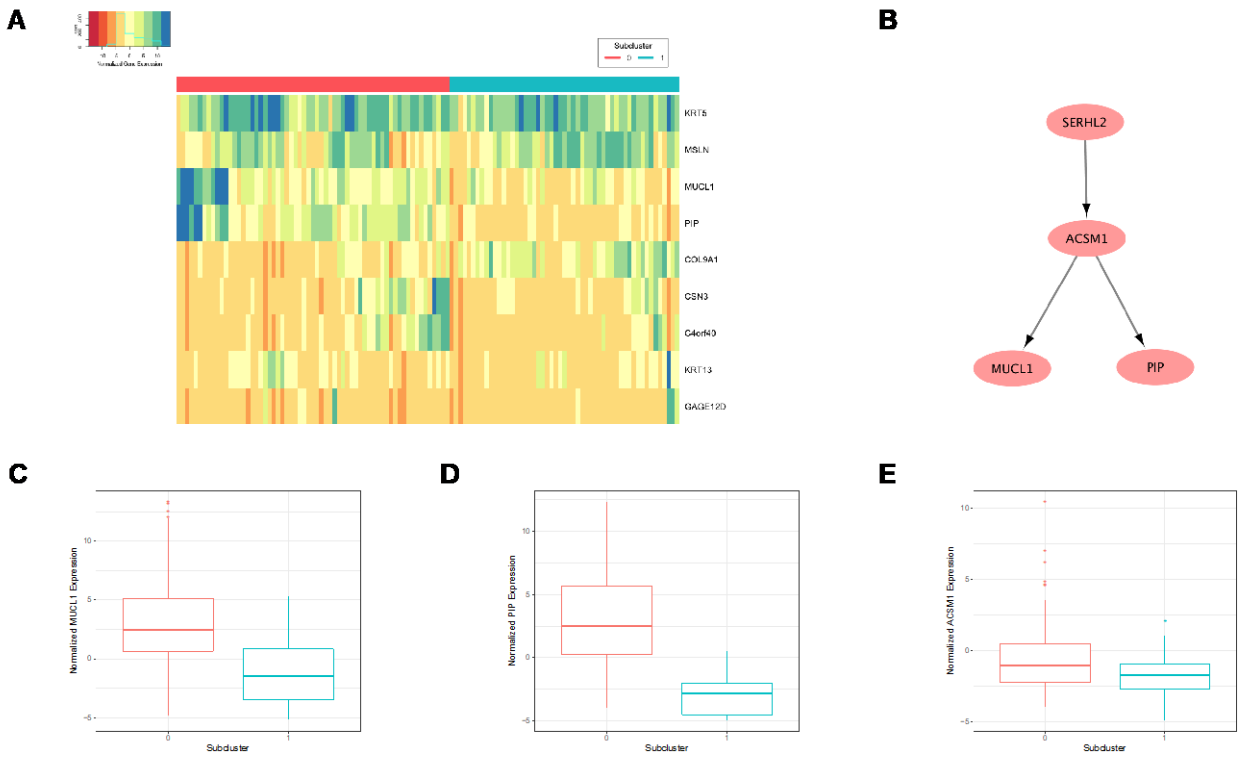

Figure 17 (A) Expression heatmap of the top PCA loading genes that separate the two triple-negative BRCA subclusters. Genes were included if they were among the top five genes for at least three of the first ten principal components that were used for ssNPA clustering. Samples are sorted by subcluster. (B) Relationship among several top PCA loading genes in the subcluster 0 network. The boxplots show the relative expression across subclusters of genes (C) MUCL1, (D) ACSM1, and (E) PIP.

Mesothelin (MSLN) separates the clusters along the first, second, fourth, sixth, and seventh principal components (Table 8). Expression of this gene is positively associated with triple negative breast cancer, and in one study, the triple negative breast cancer patients that express MSLN were older, had more distant metastases, and experienced worse survival outcomes (53). Another study linked overexpression of MSLN to increased ERK1/2 and MMP-9 protein levels and invasive capability in MCF-7 cells, a triple negative breast cancer cell line (54). Mucin-like 1 (MUCL1) had the largest PCA loading for the first principal component, plus strong effects with

the fifth, eighth, and tenth principal components. The expression of MUCL1 is decreased with the inhibition of HER2 activity, and it is proposed to regulate both the phosphorylation of FAK and cell cycle progression through the FAK-JNK pathway (55). In our networks, MUCL1 had no connections in either the HER2+ network or the network with better prognosis, but had a single connection to acyl-CoA synthetase medium-chain family member 1 (ACSM1) in the network of worse prognosis (Figure 17B). ACSM1 expression, in combination with 15-prostaglandin dehydrogenase, has been described as a marker for the molecular apocrine subtype of ER- breast cancer (56). We observed that both MUCL1 and ACSM1 were more highly expressed in the cluster with worse prognosis (Figure 17C and Figure 17D). ACSM1 also did not have any connections in the networks for HER2+ or cluster 1, although in cluster 0, in addition to MUCL1, it was connected to both serine hydrolase-like 2 (SERHL2) and prolactin-induced protein (PIP) (Figure 17B). Interestingly, PIP was another protein with a strong role in separating the clusters along the first, third, and ninth principal components. PIP is known to be the most regulated gene by AR and ERK inhibition in the molecular apocrine subtype and plays an important role in cell invasion and viability for these cells (57). Additionally, PIP has been proposed as a biomarker for early stage breast cancer as it is highly downregulated in these cancer samples compared to normal tissues (58). We observed that PIP was more highly expressed in the triple negative samples from cluster 0 than in those from cluster 1 (Figure 17E). In general, we found that the network for cluster with the worst prognosis (cluster 0) was more densely connected compared to the other cluster (cluster 1) or the HER2+ reference (2,137 edges compared to 1,730 edges and 1,213 edges, respectively).

3.3.5 ssNPA identifies patient subclusters with different survival rates in lung adenocarcinoma

We also tested ssNPA in subtyping patients in the context of lung adenocarcinoma, a disease for which there is no subtyping ground truth. The normal samples were used as reference dataset when needed (ssNPA, Pathifier), but they were omitted during clustering in order to better facilitate the discovery of new disease subtypes. ssNPA features resulted in four clusters with significantly different survival rates (p=3.6e-04, Figure 18A and Figure 19A). Subjects who maintain the greatest survival probability through the first 1,500 days were clustered together in cluster 2. Similarly, subjects with the worst survival are all clustered together in cluster 3. The other two clusters comprised of subjects with an intermediate survival phenotype. While we observe large differences in the survival curves at later time points (after 1,000 days) these have very few subjects and thus contribute little to reported p-value. The only other method that produced significantly different clusters in terms of survival was gene expression alone (five clusters; p=7.1e-04, Figure 18B and Figure 19B). Survival differences in Pathifier and ssGSEA clusters (Figure 18C-D and Figure 19C-D) were not significant (p=0.07 and p=0.08, respectively).

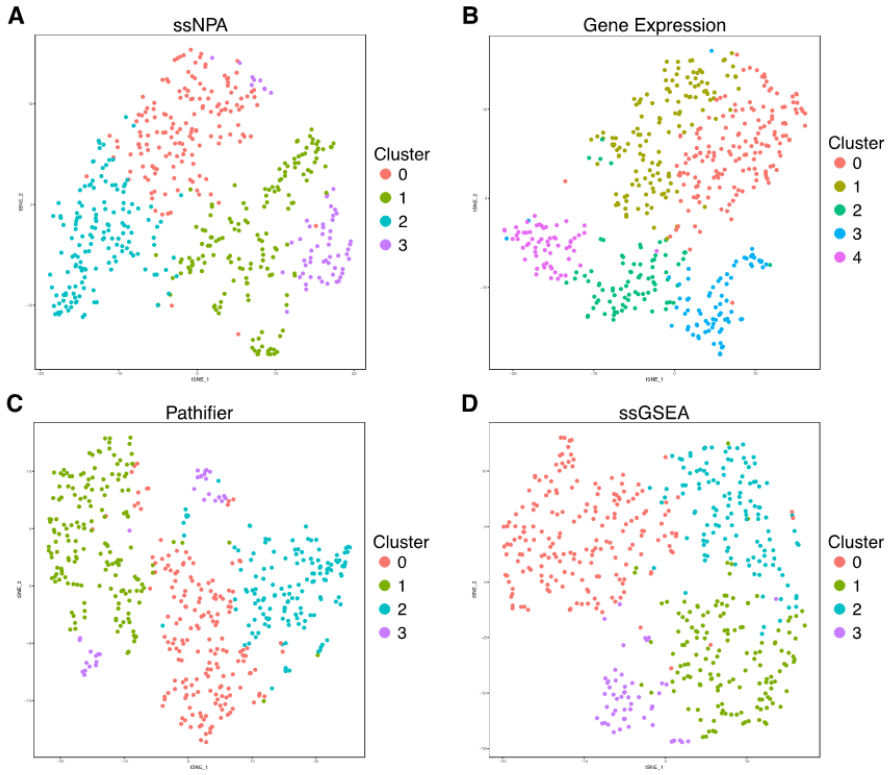

Figure 18 Lung adenocarcinoma RNA-seq sample clusters as discovered with (A) ssNPA, (B) gene expression, (C) Pathifier, and (D) ssGSEA. ssNPA was used with PD=6. Clustering for all methods was performed with the first six principal components.

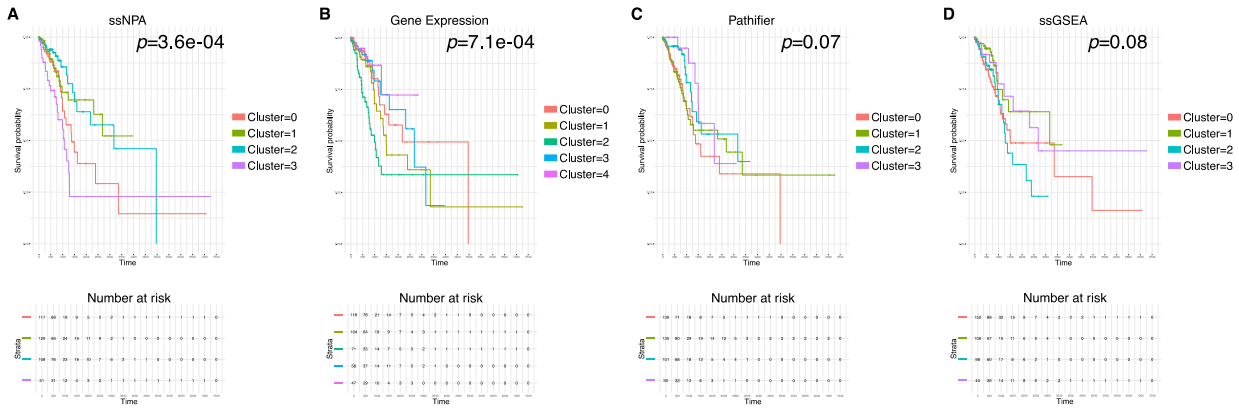

Figure 19 Full lung adenocarcinoma subject survival analysis by cluster as assigned with (A) ssNPA, (B) gene expression, (C) Pathifier, and (D) ssGSEA. ssNPA was used with PD=6. Clustering for all methods was performed with the first six principal components.

3.3.6 ssNPA identifies patient subclusters with differentially deregulated genes previously linked to lung cancer

Similarly to our analysis of breast cancer, ssNPA features can be used not only to separate patients into groups with coherent clinical phenotypes but also investigate the specific network perturbations that underlying differences among clusters. Table 9 lists the top five genes based on their factor loadings for the first six principal components of the ssNPA features.

Table 9 The genes with the top 5 loadings for the first 6 PCs in the LUAD dataset. ssNPA was used with PD=6.

	PC1	PC2	PC3	PC4	PC5	PC6
HOXB9					0.15	
CALCA					0.16	
KRT5						0.16
PRG4	0.17					
CA9	0.18					
VIL1		0.18				
SYT12				0.18		
SPINK1		0.19				
LOC84740				0.19		
PRAME						0.19
TFF3		0.2				
SFTPC	0.22					
FGL1		0.22				
CA9				0.23		
ANXA10			0.24			
SCGB1A1			0.25			
TFF1			0.27			
SFTPC						0.31
GPR110				0.18	0.21	
ITLN1	0.18		0.29			
TFF1		0.22	0.25			0.16
XAGE1D	0.18				0.56	
C20ORF114				0.3	0.2	0.45

Notably, many of these genes have well-documented connections to lung physiology and cancer biology. Carbonic anhydrase IX (CA9) was one of the top genes whose deregulation separated the clusters along the first principal component. CA9 is known to be overexpressed in several types of tumors under hypoxic conditions and, in non-small cell lung cancer, has been linked to worse survival outcomes (59). Trefoil factor 1, 2, and 3 (TFF1, TFF2, and TFF3) were all also found to be important for separating the clusters. In particular, TFF1 was found to be more highly expressed in cluster 1 compared to the other clusters, and its expression was predicted more poorly by the gene network from the normal samples (Figure 20). Patients in cluster 1 have the best prognosis, which is consistent with previous observations that TFF1 deficiency is linked to higher tumor incidence in breast cancer and TFF1-KO mice exhibit increased tumor development in the mammary gland, ovary, and lung (60).

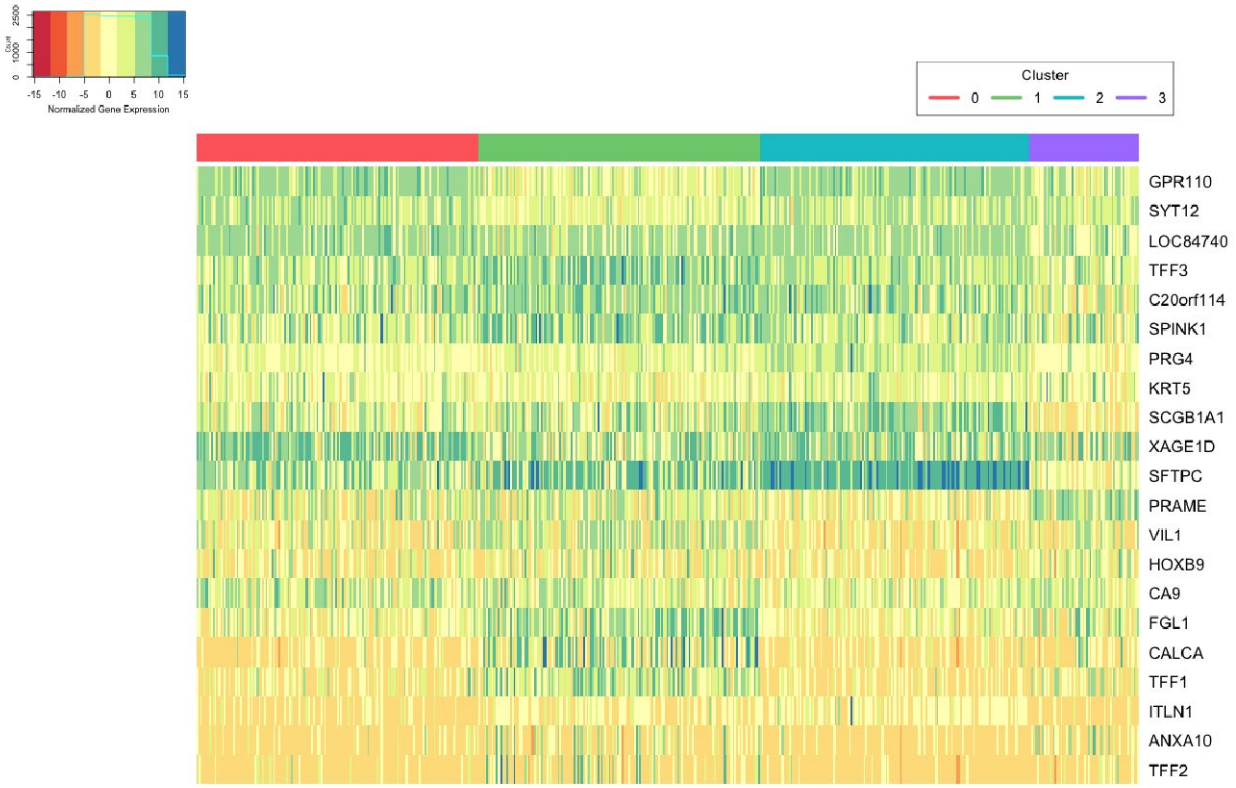

Figure 20 Expression heatmap of the top PCA loading genes that separate the four lung adenocarcinoma clusters. Samples are sorted by cluster.

We further sought to understand how the deregulation of these genes differs among the four subclusters, as well as between normal and cancer samples. Thus, we learned the causal network of every subcluster and compared them to each other as well as the network learned from normal samples. Figure 21 shows selected subnetworks from each of the four clusters, centered around the top genes whose deregulation separates the clusters, and compares them to this subnetwork in the normals. Interestingly, we observed that cluster 3, whose patients experienced the worst survival outcomes, produced the least dense network (3,003 edges compared to 4,518 edges in the normal network). However, each of the other three clusters had additional edges in their graphs compared to the normal network (with 5,303, 5,632, and 5,548 edges, respectively).

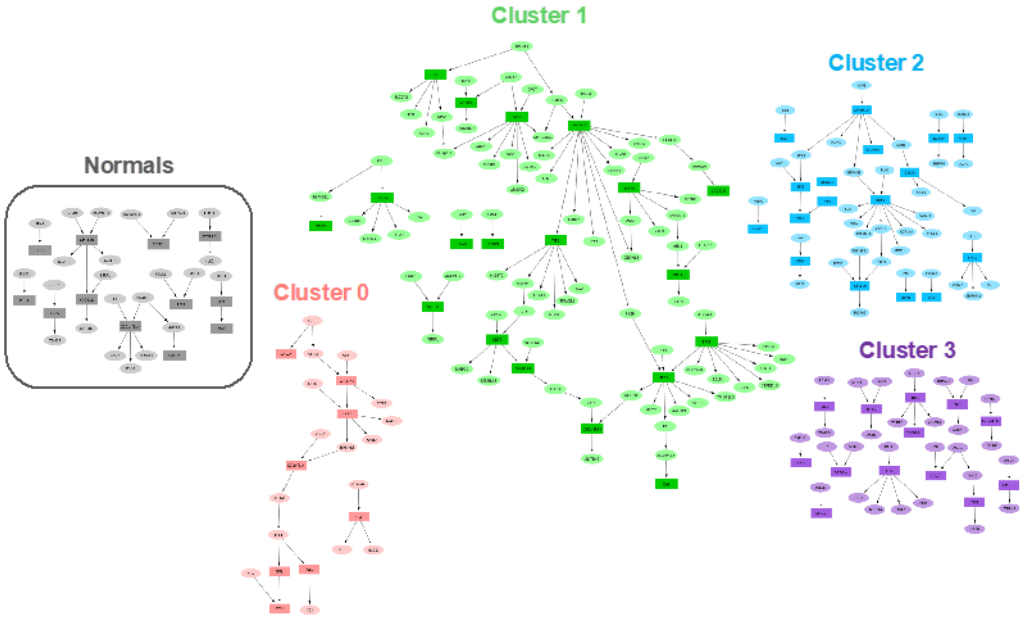

Figure 21 Comparison of lung cancer gene subnetworks of the normal (gray), cluster 0 (red), cluster 1 (green), cluster 2 (blue), and cluster 3 (purple) subjects. Subnetworks highlight the top PCA loading genes (darker color, square nodes) and their first neighbors (lighter color, rounded nodes). Genes with no adjacent edges are not included.

3.4 Discussion

We presented ssNPA, a new method to assess gene network perturbations in each sample. The method first infers the global network from a set of reference samples using causal graph learning. In the following step given a new sample the method calculates its deviation from the reference network at every gene, thus providing information about both the topology and the magnitude of network perturbations. The perturbation feature vector can been used to cluster samples into cell or disease subtypes. We demonstrated the performance of ssNPA by using it to evaluate cluster memberships of datasets with known ground truth; specifically, liver development

cells (time course scRNA-seq data) and TCGA breast and lung cancer data. In the first case, we showed that ssNPA performs better than currently used methods and from simple gene-based clustering on finding the true developmental stage and type of the cell. This showed that network perturbation features can recapitulate the time course data. In this dataset, we found that using one of the middle developmental stages (which are equidistant from both progenitor and fully differentiated extremes) as reference point allows for better results

In the cancer data we identified clusters of patients either with good agreement with known histologically-determined cancer subtypes (breast cancer) or with significant differences in survival (lung adenocarcinoma). Further analysis of the adenocarcinoma clusters gave us some insights on the molecular mechanisms that may affect survival. Both these cases demonstrate the ability of ssNPA to identify disease subtypes, which is the most significant problem in developing personalized medicine strategies, especially in complex diseases.

In all cases, we compared ssNPA to a classification scheme that uses the gene expression values directly and with ssGSEA and Pathfinder, two known methods for single sample analysis. ssNPA performed better than these methods in all cases, as is evidenced by the greater agreement of the ssNPA-identified clusters to the ground truth and the more significant differences in survival rates in the cancer cases. The difference between using network deregulation features (ssNPA) versus simple gene expression differences is that the former captures not only the gene expression differences but also differences in the topology of the network from the reference samples. The better performance of ssNPA versus ssGSEA and Pathfinder might reflect the fact that the latter depend on prior knowledge that might not be very accurate or might not reflect the particular conditions in the studied dataset. Having said that, we need to emphasize the importance of the selection of the reference group. If the gene networks in the reference group are very different

than those in the sample subgroups (e.g., disease subphenotypes), then ssNPA will not have the power to detect the subgroups because they will all be "equidistant" from the reference set.

In summary, ssNPA is a new method for characterizing single samples of gene expression and offers significant advantages over existing methods. Unlike ssGSEA and Pathifier, it does not require prior pathway knowledge; it is substantially faster than Pathifier; and can be used to produce high quality sample clusters that reflect the underlying mechanisms of the disease condition or phenotype. In the future, ssNPA can be used for analyzing disease data to identify disease subphenotypes and develop personalized intervention strategies.

4.0 Gene expression network-based subtyping according to COPD phenotype predicts genetic mechanism of disease

4.1 Background

Chronic obstructive pulmonary disease (COPD) is a heterogeneous disease related to the narrowing of small airways and emphysema (61, 62). It is the third leading cause of death in the United States (63). Diagnosis of the disease is defined solely by spirometric measures reflecting reduced airflow, specifically a ratio of forced expiratory volume in 1 second (FEV1) over forced vital capacity (FVC) less than 0.70 (64), but the underlying disease mechanism is not well understood, and this definition does nothing to account for the vast heterogeneity observed in COPD cases. The main risk factor of COPD is smoking, but one study still showed that only 50% of even lifelong smokers developed COPD (65). This suggests there is an important genetic component to the disease that is independent of smoking and other environmental exposure impacts. Among people who do develop COPD, there is a lot of variability in the rate of progression of the disease (66), response to treatment (67-69), symptom presentation (70), inflammatory response (71), and changes to lung physiology (72). Therefore, there has been tremendous interest in discovering subtypes of the disease that reflect differences along these axes. Well-characterized subtypes with easily measurable biomarkers would allow for the selection of high-risk COPD populations for potential treatment, patient stratification leading to more highly-powered clinical trials, and enriched signals of rare genetic variants and molecular phenotypes that are risk factors for development of the disease (73).

This type of approach has been relatively successful in asthma (74), but efforts in COPD have proven more difficult. Many previous attempts to subtype COPD have been limited by complications of reproducibility, with the number of subtypes identified largely ranging from 2 to 5, and study design, with cohorts in which women and subjects with mild disease were underrepresented (75). Another study applied a consistent clustering analysis to 10 independent cohorts and found only modest reproducibility across cohorts, but had more success with a continuous PCA-based projection of the subjects (76). The authors suggest that the disease is best represented as a COPD continuum instead of separate and mutually exclusive subtypes. However, this interpretation does not account for the suspected varying genetic basis of COPD and, without clear cut-off points along the continuum, does not offer much utility in practice.

Another limitation to efforts to subtype COPD is that there is a frustrating barrier to validating and interpreting COPD subtypes clustered based on their clinical characteristics, such as spirometric variables and body mass index. Some studies have tried to circumvent this problem by withholding a pre-defined subset of clinical characteristics out at the clustering step and then using those to asses the resulting clusters (77). While it can be possible to find distinct groups of subjects that do separate according to these types of clinical variables, these classifications tell us little to nothing new about how the disease works. Instead, the incorporation of genomic and transcriptomic information can greatly enhance the relevance of COPD subtypes. One previous study identified four COPD clusters based on blood gene expression with a network-based approach (78). Peripheral blood gene expression features are attractive candidates for biomarkers because they are so easily accessible. These clusters of subjects promisingly varied in the severity of their disease, but, because the study relied on microarray gene expression data, discovery was limited to the genes included on those platforms. However, the cost of sequencing is going down,

and the availability of large transcriptomic datasets is increasing, making this approach much more accessible and practical for clinical use.

We hypothesized that our ssNPA method for clustering (described in Chapter 3) would be a useful approach to COPD subtyping based on gene expression data and would improve on previous work in several ways. ssNPA models gene regulatory networks directly with a sophisticated causal algorithm, and the features used for clustering are built intentionally to capture the expression and regulatory differences in the subtypes compared to a reference group, which in this case was a carefully chosen group of former smokers who do not have COPD. This framework contributes to ease of interpretability, even allowing for the identification of specific genes whose deregulation changes among clusters, and can capture information from a complex network of interactions.

4.2 Materials and Methods

4.2.1 COPDGene dataset

COPDGene is a large ongoing longitudinal study that aims to investigate the genetic basis of COPD susceptibility and progression through the observation of over 10,000 subjects over the course of 10 years and counting. The dataset is comprised of a variety of genetic and phenotypic measurements, including genotype, gene expression, protein expression, and a huge array of clinical variables. The study has been previously been described in detail (79). For our analysis, we used a subset of the study consisting of 1,211 subjects for whom peripheral blood mononuclear cell raw count RNA-seq gene expression data were available. These data were collected during

Phase 2 of the study, roughly 5 years after each subject's initial visit. Thus, we restrict our analysis to clinical variable measurements from Phase 2, except for those variables that measure the change between Phase 1 and Phase 2.

4.2.2 Data preprocessing

Several steps were used to process the raw count RNA-seq data in preparation for use with ssNPA. First, the data were filtered with Biomart to keep only the 19,457 protein coding genes (80). Then, the RNA-seq counts were transformed to log2 counts per million through mean-variance modeling by the voom function (Limma v. 3.32.10) (50). Because COPDGene is a large study across multiple centers, the samples were measured in 17 batches. We used the batchdetect function in R (gPCA, v. 1.0) to detect batch effects with guided principal component analysis (81). Before correction, we observed strong batch effects ($p<0.001$), with clear differences among several groups of batches when visualized with the first two principal components of the guided PCA (Figure 22A). In order to correct for these effects, we applied the removeBatchEffect function in R (Limma v 3.32.10) (50). After correction, the guided principal component analysis no longer detected any batch effects ($p=0.545$, Figure 22B). Next, we filtered the data to keep only top 3,000 most variant genes. Finally, because of the overwhelming effects of smoking on gene expression profile (82, 83), we considered only those subjects who were reported as former smokers in both visits. This left us with 617 former smokers with expression measured over 3,000 genes for analysis.

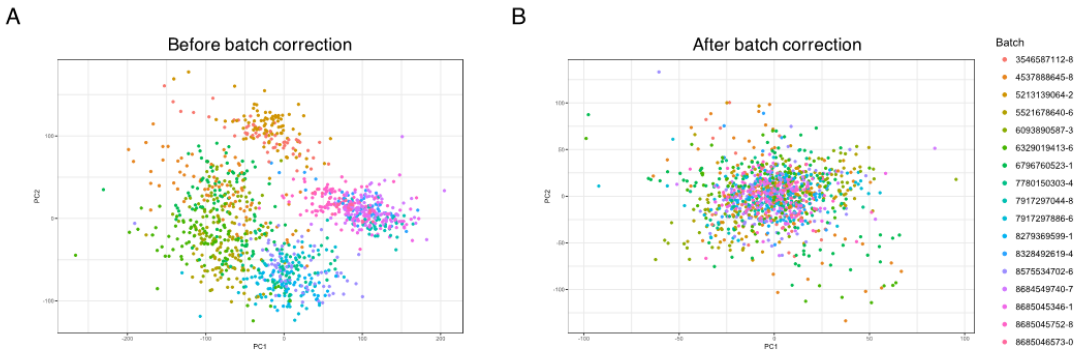

Figure 22 COPDGene RNA-seq data visualized with guided PCA (PC1 and PC2) and colored according to batch (A) before and (B) after batch correction.

4.2.3 Reference subject selection

A reference group of samples that do not have COPD were chosen very conservatively for use with ssNPA. These subjects were selected based on the following criteria: (a) subject was included in Phase 1 and Phase 2 data; (b) there was no missing data for either forced expiratory volume in one second (FEV1) or FEV1/forced vital capacity (FVC); (c) subject was GOLD0 in both visits; (d) subject had percent emphysema (Thirona) less than 5% in both visits; (e) the change in subject's percent predicted FEV1 (FEV1pp) between Phase 1 and Phase 2 visits was greater than -5. After applying these filtering criteria, we were left with 128 reference subjects and 489 COPD subjects.

4.2.4 Single sample network perturbation assessment

We analyzed this COPDGene gene expression dataset with single sample network perturbation assessment (ssNPA), as described in Chapter 3. Briefly, we learned a gene expression

network on the reference subjects who did not have COPD using FGES with a penalty discount PD=4. We trained linear regression models on the reference subjects for every gene to predict its expression based on the expression of the genes in its Markov blanket in the reference network. Next, we calculated the ssNPA features for all the COPD case subjects, where for every gene we recorded the magnitude of the difference between the predicted expression of the gene based on the reference sample network and its actual expression in that sample. Finally, with the COPD subjects represented in this new feature space as our dataset, we clustered them using the first six principal components of the data and visualized the results with a t-SNE plot (47). We expected the ssNPA features would lead to clusters of COPD subjects separated by their relative gene expression network deregulation compared to the reference subjects who do not have COPD.

4.2.5 Cluster annotation

To investigate the clinically relevant differences among the clusters of COPD patients, we compared the values of many clinical variables across the clusters. These included spirometry, radiographic, symptom questionnaire, and peripheral blood cell composition measurements, as well as medical history and comorbidity information. For continuous and ordinal variables, we applied the Kruskal-Wallis test. For discrete and binary variables, a Chi-squared test was used. Multiple comparisons were controlled for with false discovery rate (FDR).

To better understand how the clusters were separated, we considered the magnitude of the PCA loading for each feature. Gene features with the highest loading values in the top principal components correspond to the genes whose deregulation relative to the controls contributes the most to separating the clusters.

4.3 Results and Discussion

We used ssNPA to cluster the COPD subjects. The resulting clusters exhibited different degrees of disease severity and symptom presentation. We sought to understand these differences by investigating the underlying changes in the gene regulatory networks of these subjects and how they can be implicated in the mechanism of the disease.

4.3.1 COPD clusters exhibit different clinical phenotypes

ssNPA separated the 489 COPD subjects into four clusters (Figure 23). The first three clusters were of roughly equal size (33.7% of subjects in cluster 1, 30.5% in cluster 2, and 26.0% in cluster 3). Cluster 4 was the smallest with only 48 subjects (9.8%) and was more clearly separated from the other three clusters in the t-SNE projection.

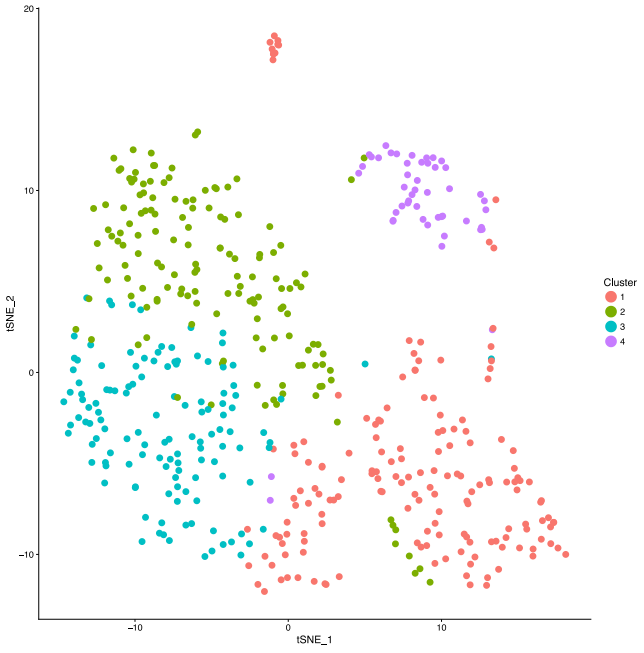

Figure 23 t-SNE plot of the four COPD subject clusters identified by ssNPA.

In order to understand if these clusters were able to capture any meaningful biological differences in the subjects' disease severity or progression, we check how the clusters varied across a set of features that we selected based on their clinical relevance (Table 10). In general we observe that cluster 4 most closely resembles the control group of subjects who do not have COPD, and the subjects in cluster 1 are the most symptomatically affected, with the worst exercise tolerance based on walk distance and lowest FEV1. Clusters 2 and 3 experience more intermediate phenotypes between these two. The variables that were most different among the clusters were largely related to quality of life and greater dyspnea with exertion. These results are significant because only the most basic pulmonary function and emphysema information were taken into account when selecting the reference subjects and the COPD subjects were clustered with their gene expression as the sole input into ssNPA. Thus, our method allowed us to associate transcriptomic signatures directly to variation in phenotypes and high-level symptomatic presentation of the disease. We also noted that, although the control subjects were younger and there were more females than males (on average), age and sex were relatively consistent across the COPD subject clusters and do not seem to be confounding these results.

Table 10 Clinical characteristics of COPD subjects vary across clusters. The variables are sorted by descending significance. P-values were calculated with a Kruskal-Wallis test for continuous and ordinal variables or a Chi-squared test for discrete and binary variables and assess if there differences in variable distribution among clusters. Variable means (standard deviations) are also reported for all COPD subjects overall, each COPD cluster, and all control subjects for comparison. Variables are included with a $p < 0.05$ cut-off and < 5% FDR are shown in bold.

Variable	p-value	FDR	All COPD Subjects	Cluster 1	Cluster 2	Cluster 3	Cluster 4	All Control Subjects
Number of Subjects			489	165	149	127	48	128
% Female			44%	48%	41%	44%	40%	61%
Age			70.1 (7.7)	70.3 (8)	70.5 (7.1)	69 (8)	71.5 (7.5)	66.7 (8.8)
Health status	5.50E-05	**3.96E-03**	3.3 (0.9)	3 (0.9)	3.3 (0.9)	3.5 (0.9)	3.6 (0.9)	3.8 (0.8)
SF-36 general health t-score (normalized)	7.40E-05	**3.96E-03**	46.4 (11.1)	43.7 (11.6)	46.3 (10.4)	49.1 (10.6)	49.5 (10.9)	52.9 (7.7)
SF-36 physical health aggregrate score (normalized)	1.11E-04	**3.96E-03**	43.8 (11)	41.1 (11.5)	43.8 (10.5)	46 (10.3)	47.7 (10.3)	49.5 (9)
SGRQ active score	1.74E-04	**4.65E-03**	36.1 (29.1)	43.4 (29.1)	35.6 (28.4)	30.7 (28.2)	27 (28.4)	16.9 (20.2)
SF-36 physical function t-score (normalized)	3.14E-04	**6.72E-03**	42.2 (12.4)	39 (13)	42.7 (11.6)	44.3 (12.1)	45.8 (11.3)	48.4 (8.8)
CAT8: energy	1.27E-02	**1.36E-02**	1.9 (1.3)	2.1 (1.3)	1.9 (1.3)	1.6 (1.3)	1.6 (1.4)	1.2 (1.2)
FRC/TLC ratio (Thirona)	1.11E-03	**1.45E-02**	0.6 (0.1)	0.6 (0.1)	0.6 (0.1)	0.6 (0.1)	0.6 (0.1)	0.5 (0.1)
Inhaled corticosteroids	1.16E-03	**1.45E-02**	0.1 (0.2)	0.1 (0.3)	0 (0.2)	0 (0.1)	0 (0.2)	0 (0.2)
MMRC dyspnea score	1.28E-03	**1.45E-02**	1.2 (1.4)	1.6 (1.5)	1.2 (1.3)	1 (1.2)	0.8 (1.2)	0.4 (0.9)
Total SGRQ score	1.45E-03	**1.45E-02**	22.6 (19.6)	26.9 (20.6)	21.9 (18.1)	20 (19.4)	16.6 (18)	9.2 (12.5)
6 minute distance walked (feet)	1.49E-03	**1.45E-02**	1303.8 (419.9)	1205 (435.2)	1316.8 (429.9)	1423.5 (358.1)	1279 (414.6)	1520.7 (358.4)
Change in lung density, sponge model adjustment (P1 to P2, Thirona)	2.20E-03	**1.95E-02**	1.1 (11.8)	-1.4 (11.1)	2.7 (11.6)	3 (13.1)	-0.6 (10.2)	0.6 (10.6)
CAT5: limited	2.37E-03	**1.95E-02**	1.3 (1.5)	1.6 (1.7)	1.3 (1.4)	1.1 (1.4)	1 (1.5)	0.5 (1)
CAT4: breathless	2.77E-03	**2.10E-02**	2.3 (1.8)	2.6 (1.8)	2.3 (1.7)	2.2 (1.7)	1.6 (1.7)	1.5 (1.5)
FEV6 (post-Utah)	2.95E-03	**2.10E-02**	2.8 (0.9)	2.6 (0.9)	2.9 (0.9)	2.8 (0.8)	3.1 (1)	3.3 (0.8)
SF-36 role physical t-score (normalized)	3.59E-03	**2.40E-02**	46 (11)	43.9 (11)	46 (11.3)	47.6 (10.6)	48.6 (9.9)	50.7 (8.8)
Pre-bronchodilator FEV6	3.85E-03	**2.42E-02**	2.7 (0.9)	2.5 (0.9)	2.8 (0.9)	2.7 (0.8)	2.9 (1)	3.2 (0.8)
FEV1 (post-Utah)	6.73E-03	**4.00E-02**	1.9 (0.8)	1.8 (0.8)	2 (0.8)	2 (0.8)	2.1 (1)	2.6 (0.6)
Pre-bronchodilator FVC	7.28E-03	**4.04E-02**	2.9 (0.9)	2.8 (0.9)	3 (0.9)	3 (0.8)	3.2 (1)	3.4 (0.8)
SF-36 social functions t-score (normalized)	7.55E-03	**4.04E-02**	50.4 (9.5)	48.5 (10.6)	50.5 (9)	52.6 (7.9)	50.6 (10.1)	53.1 (6.5)
Lymphocytes (K/uL)	8.37E-03	**4.26E-02**	1.9 (0.7)	1.8 (0.7)	1.9 (0.6)	2 (0.6)	1.9 (0.7)	2 (0.7)
SGRQ symptom score	1.10E-02	5.35E-02	25.3 (21.8)	27.9 (21.8)	25.5 (20.7)	24.5 (23)	17.4 (20.8)	11.1 (16.2)
Pre-bronchodilator FEV1	1.15E-02	5.35E-02	1.8 (0.8)	1.7 (0.9)	1.9 (0.8)	1.9 (0.8)	2 (0.9)	2.5 (0.6)
Lymphocyte percentage	1.22E-02	5.44E-02	26.4 (8.5)	25.2 (9.6)	26.3 (7.9)	28.3 (7.8)	26.1 (7.7)	30.9 (8.6)
FEV1 percent predicted (post-Utah)	1.74E-02	7.45E-02	71 (25.6)	66.1 (26.4)	74.2 (24.1)	72.1 (24.6)	75.1 (27.7)	99.1 (11.3)
FVC (post-Utah)	1.91E-02	7.86E-02	3.1 (0.9)	2.9 (1)	3.1 (0.9)	3.1 (0.8)	3.3 (1.1)	3.4 (0.8)
FEF 25-75 (post-Utah)	2.06E-02	8.16E-02	1.3 (1)	1.1 (1)	1.3 (0.9)	1.4 (1.1)	1.3 (1)	2.5 (1)
SGRQ impact score	2.81E-02	1.04E-01	14.1 (16.4)	17.2 (18.3)	13.1 (14.4)	12.7 (16.1)	10.4 (14.7)	4.3 (9.8)
Change in percent emphysema (P1 to P2, Thirona)	2.82E-02	1.04E-01	-0.5 (5.2)	0.7 (5.9)	-0.8 (4.8)	-1.4 (4.7)	-0.8 (4.9)	-0.5 (1.3)
CAT score	2.99E-02	1.07E-01	10.6 (7.5)	12 (8.1)	10.2 (6.8)	9.9 (7.5)	8.7 (6.9)	6.6 (5.7)
CAT1: cough	3.78E-02	1.26E-01	1.6 (1.3)	1.6 (1.3)	1.6 (1.2)	1.6 (1.3)	1.3 (1.2)	1.2 (1)
Oral corticosteroids	3.91E-02	1.26E-01	0 (0.2)	0 (0.2)	0 (0.2)	0 (0)	0 (0)	0 (0.1)
Change in oxygen use (P1 to P2)	3.93E-02	1.26E-01	0.1 (0.3)	0.1 (0.4)	0.1 (0.3)	0 (0.3)	0 (0.2)	0 (0.1)
Basophils (K/uL)	4.00E-02	1.26E-01	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Pre-bronchodilator FEF 25-75	4.22E-02	1.29E-01	1.2 (0.9)	1 (0.9)	1.2 (0.8)	1.2 (1)	1.2 (0.9)	2.1 (0.9)

4.3.2 ssNPA identifies a list of candidate genes deregulated in COPD

After observing these changes in COPD severity and symptom presentation among the clusters we identified, we wanted to understand more about the molecular changes in each cluster that are contributing to these differences. ssNPA clusters subjects according to the relative deregulation of their gene expression networks, so we looked at the gene deregulation features that

had the largest PCA loadings to see which genes were contributing the most to separating the clusters (Table 11). We focused on the top five loadings for each of the first six PCs that were used to cluster the subjects and found that many of the genes came up more than once.

Table 11 The genes with the top 5 loadings for the first 6 PCs used for clustering the COPD subjects.

		PC1	PC2	PC3	PC4	PC5	PC6
ADGRG7	ENSG00000144820		0.61	0.22	0.45	0.34	0.18
RHD	ENSG00000187010	0.06	0.26	0.55	0.13		0.43
DSP	ENSG00000096696			0.18		0.17	0.46
FMOD	ENSG00000122176		0.12	0.17	0.13		
KRT77	ENSG00000189182					0.11	0.15
GABRB2	ENSG00000145864			0.14	0.11		
SIGLEC14	ENSG00000254415						0.18
GSTM1	ENSG00000134184	0.12					
MDGA1	ENSG00000112139				0.11		
BTNL3	ENSG00000168903	0.11					
HEPHL1	ENSG00000181333					0.10	
OVCH1	ENSG00000187950		0.10				
MAOA	ENSG00000189221					0.10	
RNF182	ENSG00000180537		0.10				
ADARB2	ENSG00000185736	0.08					
SLC44A5	ENSG00000137968	0.07					

Several of these genes have previously been noted as having a role in COPD. Desmoplakin (DSP) was recently identified as one of 22 genes containing a top coding variant (rs2076295) in a COPD genome-wide association study (GWAS) over 15,256 COPD cases and 47,936 controls (84). This locus also colocalized with an expression quantitative trait locus (eQTL) from another lung tissue dataset that included subjects with COPD (85). DSP is a desmosomal protein that plays an essential role in cell-cell linkages, especially in epidermis and cardiac muscle (86, 87). DSP variants have also been associated with idiopathic pulmonary fibrosis (88), although these variants may be protective against COPD (84).

Adhesion G Protein-Coupled Receptor G7 (ADGRG7; alias GPR128) is another gene whose deregulation was important for separating the clusters. ADGRG7 encodes the G protein-

coupled receptor 128 (GPR128) protein. Differential nasal expression of this gene has been linked to atopic asthma in children (89), which is known to be a risk factor for later development of COPD (90-92). The genomic region containing GPR128 shares synteny with the homologous region of the mouse genome that has been implicated in cigarette smoke-induced emphysema susceptibility in mice (93). Fibromodulin (FMOD) is a protein involved in extracellular matrix organization and has been connected within a module to a large number of other extracellular matrix pathway components in a COPD network (94). Extacellular matrix remodeling is common in all lung compartments of patients with mild and moderate COPD and likely plays an important role in airflow obstruction (95).

Another gene we identified, gluthathione S-transferase μ 1(GSTM1), belongs to a family of enzymes that are linked to lung disease, likely through their utility in detoxifying electrophilic compounds, such as cigarette smoke and environmental toxins (96). A homozygous GSTM1-null genotype has been linked to lung cancer pathogenesis (97, 98), emphysema (99, 100), and COPD susceptibility (101, 102). A different study found the GSTM1-null phenotype to be a risk factor decreased lung function in smokers living near coal mines (103), which alludes to an interesting interaction between the genetic underpinnings of COPD and environmental exposures.

The list of genes we have identified provide an interesting look into the molecular mechanism of susceptibility, such as the role of environmental toxin processing, and progression, including pathways involved in extracellular matrix organization. Several of the genes on the list such as keratin 77 (KRT77) have not been specifically cited for an association with COPD, but they code for important structural proteins and could clearly play a role in airway remodeling.

4.4 Conclusions

We have shown that ssNPA identified clusters of COPD subjects that correspond to clinically relevant variations of the disease, reflecting both severity and symptoms. Furthermore, the feature vectors used for clustering themselves can give us mechanistic insight into the disease that specifically relates the COPD cases to a control group of subjects who do not have COPD. Additionally, we identified a set of genes whose deregulation is responsible for separating the clusters. Many of these genes have previously-described connections to COPD that are further bolstered by this work. Our results also provide strong evidence for the role of a number of novel genes in COPD. The network learning and gene selection were completely unbiased, using no prior knowledge of disease mechanism or biology pathways. Finally, ssNPA is a flexible framework that can handle a variety of data types. As the data become available through COPDGene and other studies, future work could incorporate genetic variant, epigenetic, proteomic, or metabolomic variables into the network learning and feature calculations that would provide a multi-layered, more complete picture of the molecular pathology and heterogeneity of COPD.

5.0 Causal network modeling applications to chronic lung disease

5.1 Background

Causal modeling is a powerful tool for learning networks that can improve understanding of complex diseases through classification and prediction. This methodology has been applied to wide variety of diseases-related applications, including lung cancer (104), breast cancer (105, 106), coronary artery disease (107), and age-related diseases (108). We focused on two specific applications to chronic lung diseases, modeling lung function decline in COPD and cell type-specific gene interactions in IPF.

COPD is a disease of airflow limitation from airway or alveolar abnormalities and typically caused by smoking or other environmental exposures (64). The disease produces a huge healthcare burden worldwide (109), but still little is understood about the mechanism of disease and the factors responsible for its progression. Additionally, COPD is known to interact with a variety of other diseases (comorbidities) that can be causally related (110-112), through the same risk factors contributing to both diseases or the presence of one disease worsening the effects of the other (113). The rate of lung function decline, typically measured by change in FEV1, varies widely among patients (114). Several attempts have been made to build regression models for longitudinal lung function decline, although these studies have relied on very few basic clinical variables (115) or simple logistic regression analysis to model a very short period of decline (116). Our analysis of the Pittsburgh SCCOR cohort, however, combined a rich dataset of well-characterized subjects with both baseline and 2-year follow-up measurements with a causal modeling approach that

improves on basic regression analysis by modeling the conditional (in)dependencies of all the variables directly.

Idiopathic pulmonary fibrosis (IPF) is another chronic and progressive lung disease. Patients experience a mean survival time of three years after diagnosis (117). IPF is serious disease characterized by fibrosis of the lungs and involves the interactions of many different cell types, including alveolar epithelial cells (118, 119), fibroblasts (120, 121), and macrophages (122), among others. However, the specifics of the roles of each of these cells in IPF pathogenesis are poorly understood. Recently, single cell RNA-seq (scRNA-seq) experiments have made huge strides toward resolving some of the molecular complexities of varying cell populations and transcriptomes in IPF (123, 124). While these studies are contributing an incredible wealth of information about the molecular mechanism of IPF, much of their potential has yet to be unlocked, with current analyses limited mostly to standard differential gene expression and simple pathway enrichment. We present an analysis of a newly collected IPF scRNA-seq dataset that uses a causal network to focus in on the interactions among specific genes that are differently expressed in several important IPF cell types.

5.2 Materials and Methods

5.2.1 SCCOR dataset

The Pittsburgh Specialized Center of Clinically Oriented Research (SCCOR) cohort consists of 747 subjects that were recruited from a larger, less characterized community-based tobacco-exposed cohort, enriched for subjects with visual emphysema. The dataset included

clinical data from 385 COPD patients that had a completed baseline and 2-year follow up visit. Dataset acquisition included semi-quantitative visual and quantitative MDCT chest radiographic analyses, pre- and post-bronchodilator lung function testing including spirometry, body plethysmography, impulse oscillometry and diffusing capacity, extensive symptom, demographic, environmental exposure and health outcome data, incremental shuttle exercise testing, and blood circulating proteins. There were 281 variables measured in total. We used these data to identify which factors measured in visit 1 are directly linked to lung function decline, as observed two years later at visit 2.

5.2.2 IPF scRNA-seq dataset

The single cell RNA-seq (scRNA-seq) dataset included cells from tissue samples collected in both the right upper and lower lung lobes from two lungs that were normal control lungs that were rejected for transplant and two IPF lungs that were removed during transplantation surgery. The processed gene expression dataset that we obtained contained single cell samples from the following cell type groups: secreted phosphoprotein 1-positive (SPP1+) macrophage (4,489 cells), fibroblast (2,270 cells), alveolar type I (AT1, 87 cells), club/clara and goblet (1,781 cells), keratin 5-positive (KRT5+) basal epithelial (797 cells), and alveolar type II (AT2, 733 cells). These cells were characterized over the top genes that were differentially expressed between normal lungs and IPF lungs in each cell type (100 genes per cell type). Genes that were differentially expressed in more than one of these groups were excluded, resulting in 394 gene expression variables measured over 10,157 cells.

5.2.3 Causal modeling

5.2.3.1 MGM-PCS

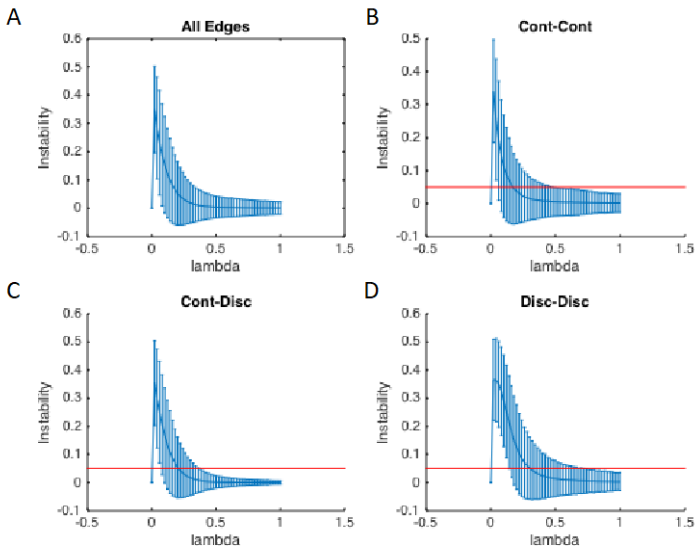

Figure 24 StEPS results for edge sparsity parameter selection. Panel (A) illustrates the edge instablity behavior over all edges types. However, separate sparsity parameters were chosen for (B) continuous-continuous edges, (C) continuous-discrete edges, and (D) discrete-discrete edges. Error bars show standard deviation of edge instability.

We applied MGM-PCS as described by (36) for the causal learning over the COPD dataset. Briefly, this approach first learns a mixed graphical model (MGM) over mixed data containing both discrete and continuous variables. The result is an undirected graph. In the next step, causal directions are assigned to the edges of the graph with a PC-stable, a methodology that relies on conditional independence testing (125). A unique feature of this implementation of MGM learning is the separate sparsity penalties, λ, assigned to each edge type (those that connect continuous to continuous variables (λ_{CC}), continuous to discrete variables (λ_{CD}), and discrete to discrete variables (λ_{DD}). We used the StEPS approach (36) to independently set these three parameters, choosing the highest value of λ for each edge type such that edge instability was greater than 0.05 (Figure 24).

As a result, we set $\lambda_{CC} = 0.16$, $\lambda_{CD} = 0.2$, and $\lambda_{DD} = 0.3$. PC-stable was applied with the parameter $\alpha = 0.25$ for the conditional independence testing.

5.2.3.2 FGES

In order to identify the causal relationships among the expression levels of the selected differentially expressed genes in the different cell types of the IPF dataset, we used the fast greedy equivalence search (FGES) algorithm (126-128). FGES performs a greedy search in which edges are added between nodes until no additional edge increases the Bayesian Information Criterion (BIC) score, and then edges are removed until no additional removal increases the score. The variables in the dataset were transformed using nonparanormal (129, 130) to relax the normality assumption of the network learning algorithm.

The penalty discount parameter, *c*, for the BIC scoring function was chosen by a method analogous to Stability Approach to Regularization Selection (StARS) (131) as described in (36). For a dataset with *n* samples by *d* variables and given *c*, we drew 40 random subsamples of size $\left\lfloor \frac{n}{2} \right\rfloor$ according to complementary pairs stability selection (132). A network was constructed for each subsample using FGES as implemented in Tetrad (http://www.phil.cmu.edu/tetrad/). For a given *c*, $\hat{\theta}_{ij}(c)$ was defined as the fraction of subsample networks in which an edge between node *i* and node *j* appeared. The edge instability was then calculated according to $\hat{\xi}_{ij}(c) = 2\hat{\theta}_{ij}(c)\left(1 - \hat{\theta}_{ij}(c)\right)$. The total instability of the network for a given penalty discount was calculated as the average instability over all edges, $\widehat{D}(c) = \frac{\sum_{i<j} \hat{\xi}_{ij}(c)}{\binom{d}{2}}$. A penalty discount value of 6 was chosen based on the total instability and sparsity in the resulting network (Figure 25).

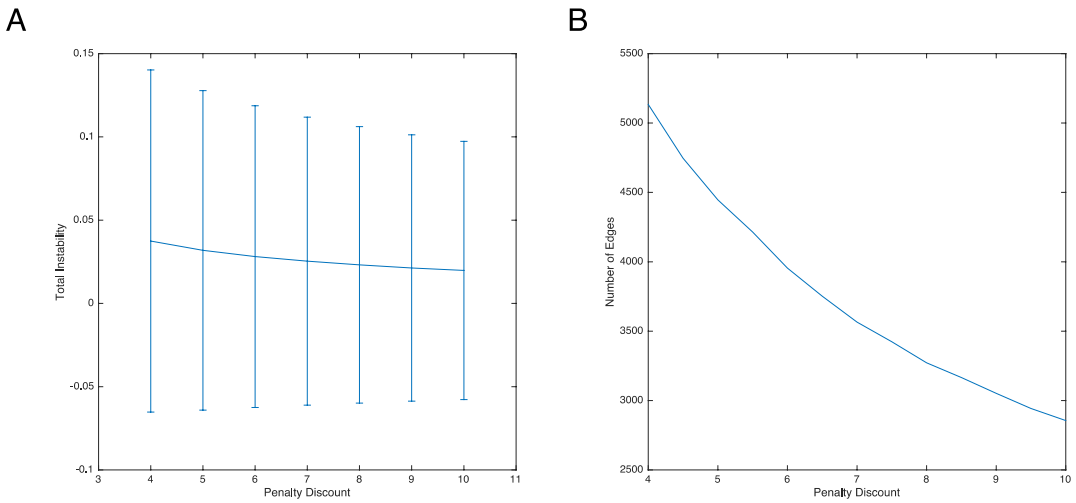

Figure 25 (A) StARS total instability versus FGES penalty discount. Error bars show standard deviation of edge instability. (B) Number of edges in the network versus FGES penalty discount.

5.3 Results and Discussion

5.3.1 Baseline factor prediction of lung function decline in COPD

We applied MGM-PCS to the clinical SCCOR dataset to identify the baseline variables that are directly (causally) linked to the 2-year lung function decline in COPD patients. Given the substantial variation in longitudinal decline in lung function, identification of baseline subject attributes that are connected to disease progression is useful for developing prediction models and offers mechanistic insight and helps to identify risk factors of progression, which could be used to develop personalized approaches to disease management or treatment. The SCCOR dataset we used included 281 variables that recorded a variety of clinical, environmental, psychological, and patients' history data in visit 1 (baseline). We ran MGM-PCS on this dataset, and we added a variable measuring the lung function decline between visit 1 and visit 2 as measured by the change

in FEV1 ("FEV1 progression"). The first and second neighbors around this target variable are shown in Figure 26.

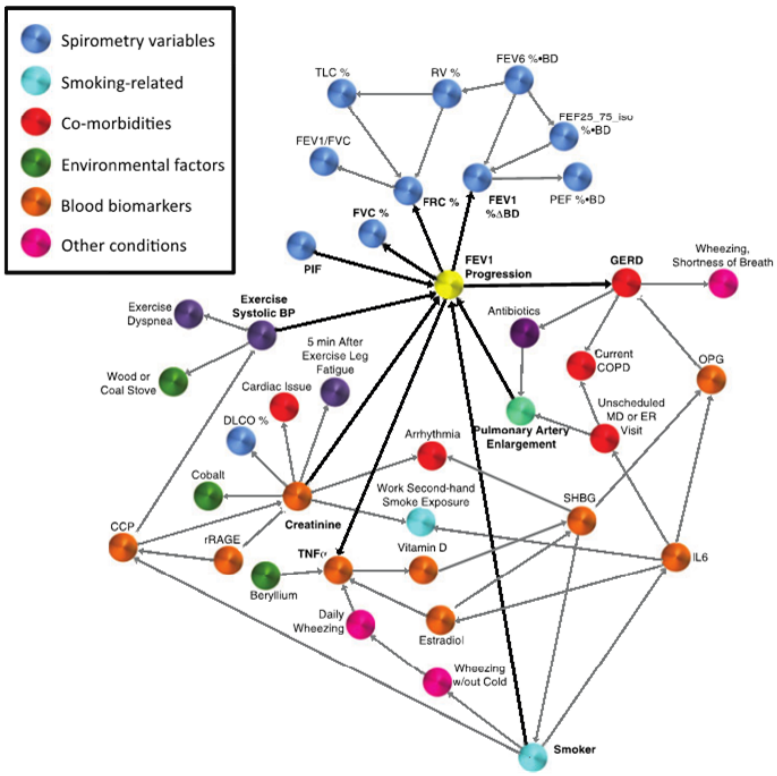

Figure 26 First and second neighbors of 2-year lung function decline, measured as FEV1 Progression. The variables that most influence the FEV1 progression are smoking status, creatinine and TNFα blood levels, pulmonary artery enlargement, history of GERD, systolic BP after exercise, and four spirometry variables (% change in FEV1 before and after bronchodilators, best % predicted FVC, best % predicted FRC, and PIF).

This network offers face validity by identifying variables as direct connector that are expected to be associated with lung function decline, such as "Smoker" and bronchodilator reversibility, "FEV1 %ΔBD" (133, 134). Some of the other connections we observe are more novel. Notably, three of the first neighbors of FEV1 progression are markers of non-pulmonary co-morbidities. "Creatinine" is a biomarker of renal dysfunction; "Exercise Systolic BP" is the systolic blood pressure at the end of 6 minute walking exercise; and "GERD" is history of

gastroesophageal reflux disease. While each has been previously linked with COPD or its exacerbations (135-138), such a dominant association with lung function decline is not well described. Such associations, however, are consistent with a systems biology mechanistic model of COPD, where activity and interaction in multiple organs rather than a single organ centric view better define the potential underlying mechanisms and impact on the patient (139).

Creatinine, for example, is directly connected to FEV1 decline and is also a hub in our network. The connections within this hub may offer further insights into the mechanistic associations between renal and lung disease. Renal dysfunction and elevated creatinine levels have been associated with pulmonary emphysema severity, which is supported by the direct connection to "DLCO" (135), a marker of parenchymal emphysema or pulmonary vascular dysfunction. Further, recent studies propose a mechanistic link between emphysema and renal dysfunction through RAGE (140-143), the receptor of which (sRAGE) is a first neighbor of creatinine in our network. The creatinine hub is further linked to a number of other important variables and confounders, including the blood biomarker CCP (Clara cell protein) whose association to COPD has been previously reported (144). In fact, the interaction between CCP and RAGE identified in our network provides incentive to explore relationships between these molecular pathways. Other direct links to creatinine include "Cardiac issue" and "Arrhythmia", attributes form the subjects' medical history, may be indicator of a common vascular mechanistic systems link. The direct line of TNFα, another blood biomarker, with disease progression is of both prognostic and mechanistic interest. TNFα is a representative biomarker for TH1 inflammatory pathways commonly linked with COPD (145). In fact, TNF modulation has been tested as a therapy in COPD, but with mixed results (146).

Another direct connector to FEV1 progression, “Exercise Systolic BP” may also reflect the vascular/endothelial processes common to pulmonary and systemic processes. The common linkage of CCP between this and the other first neighbor, creatinine, is of further interest. We note, though, that the causal direction might be predicted incorrectly in these associations. “GERD”, the final comorbidity variable that is a first neighbor of lung function decline, is of potential interest in either causal direction, as gastroesophageal reflux has known potential impacts on lung function, and lung function decline associated with lung hyperinflation can alter transdiaphragmatic pressure gradients leading to reflux. The direct connection of “Pulmonary Artery Enlargement” with FEV1 progression is of particular interest given the second neighbor links of this measure with indicators of COPD exacerbation in the past year, “Unscheduled MD or ER Visit” and “Antibiotics.” Previous work has connected pulmonary arterial enlargement to COPD (147).

Finally, three other pulmonary physiology variables are linked directly to COPD progression: “FRC%”, functional residual capacity; “FVC%”, forced vital capacity; and “PIF”, peak inspiratory flow rate. All of these measures are directly or indirectly linked to air trapping and lung hyperinflation but are independently measured attributes. To our knowledge, the direct association of these measures with FEV1 decline has not previously been defined.

These results are significant, not only because this combination of factors can determine and predict COPD progression, but because for the first time we are able to build a causal network of COD that combines heterogeneous types of information, such as measurements of lung function, symptoms, systemic comorbidities, and blood biomarkers, with environmental exposures, such as ongoing tobacco exposure. Other environmental or psychological variables, while not linked to COPD progression directly, were part of the larger network. A variable describing whether the patient has been diagnosed with depression or is on anti-depression medication, for example, was

linked to pack years of smoking. The associations found in this network or particularly notable in that they extend previous work describing the important link between non-pulmonary organ comorbidities and lung function impairment, supporting the systems biology paradigm in understanding lung disease activity (136, 139).

5.3.2 Differentially expressed gene connectivity across cell types in IPF

We built a causal network model observe the direct relationships between differentially expressed genes for SPP1 macrophages, fibroblasts, and various epithelial cell types. The resulting network included 394 differentially expressed genes and 3,951 edges (connections between these genes), and the overall relationships among the genes DE in the different cell types are summarized in Figure 27. SPP1 macrophages and fibroblasts were the most densely connected gene groups with 944 edges between them, which suggests a strong interplay between these two cell types in IPF. There were also relatively strong connections between clara/club/goblet cells and both AT2 (107 edges) and AT1 cells (102 edges). Because there were a different number of unique DE genes for each cell type, we considered edge density out of the total number of possible edges between two cell type groups. We also observed that the genes for the clara/club and goblet cell type were especially densely connected among each other with 26.6% of all possible edges among these genes appearing in the network compared to SPP1 macrophage (15.3%), fibroblast (16.2%), AT1 (10.1%), AT2 (15.4%), and KRT5 basal (15.1%).

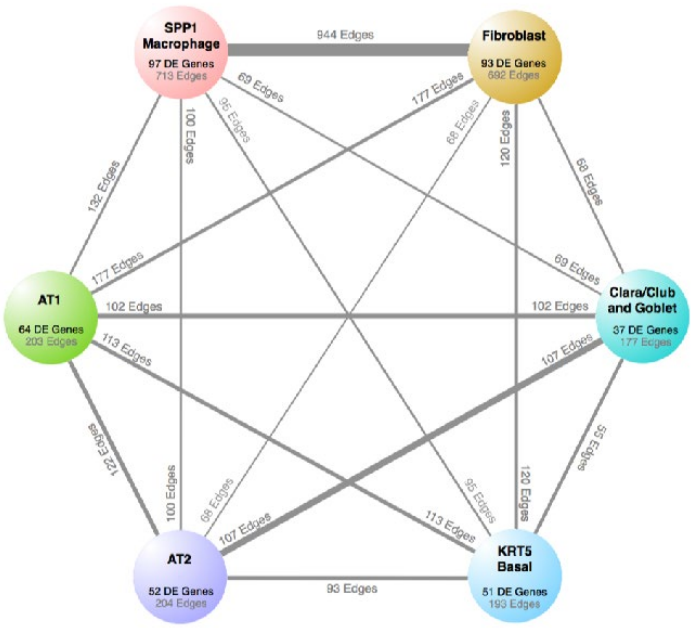

Figure 27 Causal analysis of IPF and control lung gene expression. Summary nodes represent the collection of genes that were differentially expressed in a given cell type. Each node is labeled with the number of differentially expressed (DE) genes that node represents as well as the number of edges among the genes represented by that node. Lines connecting nodes in this figure represent the set of edges that connect pairs of genes that were differentially expressed in different cell types. These are labeled with the number of edges in the network that span the two cell types, and the line weights are proportional to the percentage of edges that appear in the network relative to the total number of possible edges.

While the overall structure of the gene network can give us a big picture view of the interactions among cell types, we also explored smaller subnetworks centered around certain genes of interest. Specifically, we investigated the first neighbors of SPP1 and MERTK (Figure 28A), which were differentially expressed in marcophages and first neighbors of each other, and COMP (Figure 28B), which was the most highly upregulated gene in fibroblasts. Some of the first neighbors of SPP1 and MERTK where genes that were downregulated in IPF fibroblasts including complement factor D (CFD) and selenoprotein P (SEPP1), a gene also downregulated in prostatic

fibroblasts by TGFβ (148). The only fibroblast first neighbor of MERTK/SPP1 macrophages was versican (VCAN), a protein associated with early fibroblastic foci (149). VCAN was upregulated in IPF fibroblasts. Several epithelial cell genes were also first neighbors of MERTK/SPP1. Fibrinogen gamma chain (FGG) was strongly upregulated in AT2 cells from the upper but not the lower lobe, and serpin family A member 1 (SERPINA1) was upregulated in AT2 cells from both upper and lower lobes. These may represent genes involved in AT2 cell injury.

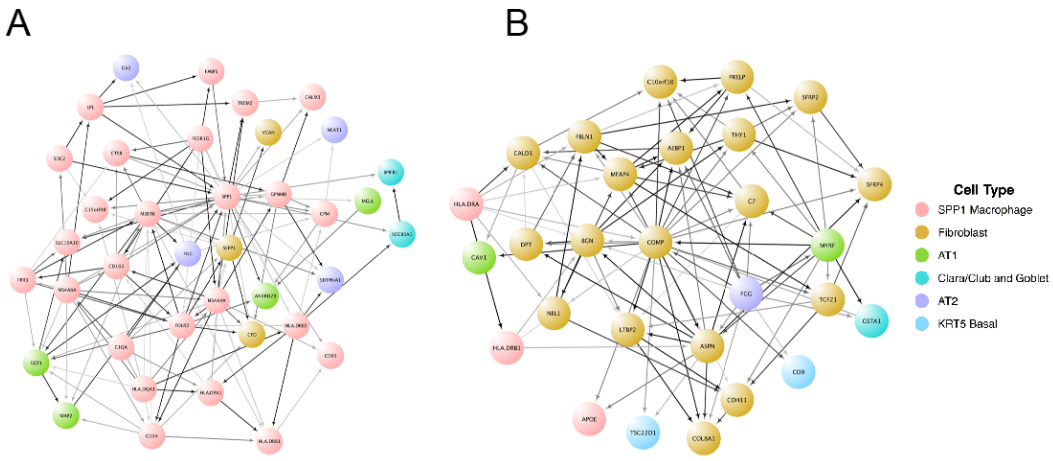

Figure 28 Subnetworks highlight the first neighbors of (A) MERTK and SPP1 which are DE in SPP1 macrophages and (B) COMP which is DE in fibroblasts. Node colors indicate the cell type in which the gene is DE, and darker colored edges correspond to more stable connections.

FGG is also a first neighbor of COMP and is expressed in liver as well as fetal lung. This suggests it might be re-expressed under AT2 stress or regeneration. Although deletion of SERPINE1 increased plasmin reduces lung fibrosis, deletion of fibrinogen alpha gene does not affect bleomycin induced pulmonary fibrosis (150). The effect of FGG deletion has not been explored.

5.4 Conclusions

These two examples illustrate the utility of causal network modeling in understanding chronic lung diseases. We have shown how adaptable these methods are to two different lung disease datasets that vary in their scale, composition, and scope. The MGM-PCS approach is appropriate for mixed data types, including continuous and discrete variables, which makes it particularly well-suited for clinical data. For example, in the SCCOR COPD dataset we explored, many of the patient history and environmental exposure variables are discrete. However, we are able to model them jointly with other continuous variables like the spirometry measurements and blood biomarker concentrations. Combining these rich heterogeneous data with MGM-PCS allowed us to take a comprehensive clinical and systems level look at the factors underlying lung function decline in COPD. We were able to identify a variety of baseline factors, including a mixture of both strongly established and novel features, that influenced 2-year lung function decline in these subjects. These factors are useful both for their insight into how the disease works and, more directly, can be used as input for predicting lung function decline. This work highlights new connections that could help improve prognosis of the disease progression and hopefully lead to better treatments to mitigate lung function decline.

Our approach to investigating IPF employed a similar causal modeling approach, however the dataset for insight on a completely different scale. In the case of COPD, we looked at the disease as a function of body systems, comorbidities, and environmental interactions. For IPF, on the other hand, we took an extremely fine-grained approach. Our data were measures on the level of expression of hundreds of genes in individual cells, which allowed us to better understand how the various cell types interact in the lung in the context of IPF. The scRNA-seq gene expression variables are all continuous so they were well-suited to using FGES for causal modeling.

Taken together, these results emphasize the promise of causal modeling methods for understanding complex diseases. This general approach is flexible and fast, can be adapted for a huge variety of data, and result in easily interpretable models that can provide real insight into extremely complicated systems.

6.0 Conclusions and Future Work

In this dissertation, we have focused on improving clustering in high-throughput biological datasets by developing a variety of new features that are specifically tailored to reflect the biological properties of the systems we are trying to understand. We started by proposing new features for representing transcription factor binding sites that capture both the DNA sequence composition of the binding region and the TF-DNA binding strength and observed that these new scaled sequence features aided clustering for improved DNA binding motif discovery. Next, we presented a new method, ssNPA, and demonstrated how causal network learning algorithms could be used to build features that capture the complex interactions of variables within biological systems such as gene regulatory networks and cluster samples based on how these networks are deregulated in different subtypes. We validated this method in a murine liver cell development dataset and with transcriptomic datasets comparing breast cancer and lung adenocarcinoma tumor samples to normal tissue. Then we used ssNPA to describe new subtypes of COPD that were based on their relative gene network deregulation compared to normal samples. Finally, we applied causal network modeling techniques to two datasets of chronic lung diseases, exploring the systems biology of lung function decline in COPD at the body systems level and cell type interactions in IPF at the scale of the gene expression in single cells.

This work presents many opportunities for further development. Primarily, ssNPA has a lot of potential to be extended and applied to more complex problems. As it is designed, any type of continuous data can be used as input although we have focused here on gene expression variables. We first want to explore the inclusion of genetic variant data in addition to gene expression. Genotypes can reasonable be modeled as ordinal variables (0, 1, or 2 copies of the

variant), so we should be able to use them as input into FGES. However, with the incorporation of a larger variety of data and heterogeneous data types, such as methylation or clinical measurements, we can adapt ssNPA to use a different causal learning algorithm such as MGM-PCS, which can appropriately handle mixed data types. As we broaden the types of variables we add into the network model and our dataset increase in their dimensions, especially those derived from high-throughput and genome-wide experiments, feature selection will become an increasingly important task.

Additionally, there is still a lot of opportunity with the application of ssNPA to COPD. The COPDGene study has already collected an abundant compendium of -omics data on its long-term cohort, including full genome sequencing on all of its subjects and large-scale proteomics measurements. A "central dogma" analysis that combines genotype information with both RNA and protein expression levels in a single network is possible with ssNPA and could more fully elucidate the variations in disease mechanism experienced by different subtypes of subjects. The COPDGene research program is currently entering into its Phase 3 data collection phase, and as the work continues, the dataset will be increasingly enriched. Of course, COPD is not the only complex disease of interest. There has been a trend to highly collaborative, multi-center research programs that tackle a single disease from a variety of angles, and ssNPA could be easily applied to any number of these studies.

Bibliography

1. Gerstein MB*, et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489(7414):91-100.
2. mod EC*, et al.* (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* 330(6012):1787-1797.
3. Yip KY*, et al.* (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* 13(9):R48.
4. Boyle AP*, et al.* (2011) High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res* 21(3):456-464.
5. Hesselberth JR*, et al.* (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods* 6(4):283-289.
6. Giresi PG, Kim J, McDaniell RM, Iyer VR, & Lieb JD (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* 17(6):877-885.
7. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, & Greenleaf WJ (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10(12):1213-1218.
8. Rhee HS & Pugh BF (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* 147(6):1408-1419.
9. He Q, Johnston J, & Zeitlinger J (2015) ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat Biotechnol* 33(4):395-401.
10. Kundaje A*, et al.* (2012) Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res* 22(9):1735-1747.
11. Nair NU, Kumar S, Moret BM, & Bucher P (2014) Probabilistic partitioning methods to find significant patterns in ChIP-Seq data. *Bioinformatics* 30(17):2406-2413.
12. Ye T*, et al.* (2011) seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res* 39(6):e35.

13. Narlikar L (2013) MuMoD: a Bayesian approach to detect multiple modes of protein-DNA binding from genome-wide ChIP data. *Nucleic Acids Res* 41(1):21-32.
14. Marcais G & Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27(6):764-770.
15. Portales-Casamar E*, et al.* (2010) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res* 38(Database issue):D105-110.
16. Zhang Y*, et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9(9):R137.
17. Guo Y, Mahony S, & Gifford DK (2012) High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. *PLoS Comput Biol* 8(8):e1002638.
18. Heinz S*, et al.* (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38(4):576-589.
19. Gordan R*, et al.* (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep* 3(4):1093-1104.
20. Huang C, Chan JA, & Schuurmans C (2014) Proneural bHLH genes in development and disease. *Curr Top Dev Biol* 110:75-127.
21. Schneider DS, Hudson KL, Lin TY, & Anderson KV (1991) Dominant and recessive mutations define functional domains of Toll, a transmembrane protein required for dorsal-ventral polarity in the Drosophila embryo. *Genes Dev* 5(5):797-807.
22. Matsukage A, Hirose F, Yoo MA, & Yamaguchi M (2008) The DRE/DREF transcriptional regulatory system: a master key for cell proliferation. *Biochim Biophys Acta* 1779(2):81-89.
23. Nitta KR*, et al.* (2015) Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife* 4.
24. Cummings BB*, et al.* (2017) Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci Transl Med* 9(386).

25. Kremer LS, *et al.* (2017) Genetic diagnosis of Mendelian disorders via RNA sequencing. *Nat Commun* 8:15824.
26. Chen Y, *et al.* (2018) Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. *Cell Res* 28(9):879-896.
27. Villani AC, *et al.* (2017) Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* 356(6335).
28. Zhao T, *et al.* (2018) Single-Cell RNA-Seq Reveals Dynamic Early Embryonic-like Programs during Chemical Reprogramming. *Cell Stem Cell* 23(1):31-45 e37.
29. Barbie DA, *et al.* (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 462(7269):108-112.
30. Tomfohr J, Lu J, & Kepler TB (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics* 6:225.
31. Hanzelmann S, Castelo R, & Guinney J (2013) GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14:7.
32. Drier Y, Sheffer M, & Domany E (2013) Pathway-based personalized analysis of cancer. *Proc Natl Acad Sci U S A* 110(16):6388-6393.
33. Mohammadi S, Ravindra V, Gleich DF, & Grama A (2018) A geometric approach to characterize the functional identity of single cells. *Nat Commun* 9(1):1516.
34. Friedman N (2004) Inferring cellular networks using probabilistic graphical models. *Science* 303(5659):799-805.
35. Sachs K, Perez O, Pe'er D, Lauffenburger DA, & Nolan GP (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308:523-529.
36. Sedgewick AJ, Shi I, Donovan RM, & Benos PV (2016) Learning mixed graphical models with separate sparsity parameters and stability-based model selection. *BMC Bioinformatics* 17 Suppl 5:175.
37. Huang GT, Tsamardinos I, Raghu V, Kaminski N, & Benos PV (2015) T-ReCS: stable selection of dynamically formed groups of features with application to prediction of clinical outcomes. *Pac Symp Biocomput*:431-442.
38. Raghu VK, *et al.* (2018) Biomarker identification for statin sensitivity of cancer cell lines. *Biochem Biophys Res Commun* 495(1):659-665.

39. Raghu VK, Poon A, & Benos PV (2018) Evaluation of Causal Structure Learning Methods on Mixed Data Types. in *Proceedings of 2018 ACM SIGKDD Workshop on Causal Disocvery* (PMLR, Proceedings of Machine Learning Research), pp 48--65.
40. Raghu VK*, et al.* (2018) Comparison of strategies for scalable causal discovery of latent variable models from mixed data. *Int J Data Sci Anal* 6(1):33-45.
41. Sedgewick AJ*, et al.* (2018) Mixed Graphical Models for Integrative Causal Analysis with Application to Chronic Lung Disease Diagnosis and Prognosis. *Bioinformatics* 34:accepted.
42. Yang L*, et al.* (2017) A single-cell transcriptomic analysis reveals precise pathways and regulatory mechanisms underlying hepatoblast differentiation. *Hepatology* 66(5):1387-1401.
43. Cancer Genome Atlas N (2012) Comprehensive molecular portraits of human breast tumours. *Nature* 490(7418):61-70.
44. Cancer Genome Atlas Research N (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489(7417):519-525.
45. Butler A, Hoffman P, Smibert P, Papalexi E, & Satija R (2018) Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* 36(5):411-420.
46. Waltman LvE, N. J. (2013) A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B* 86:471.
47. van der Maaten LJPH, G. E. (2008) Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research* 9:2579-2605.
48. Subramanian A*, et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102(43):15545-15550.
49. Ramsey J, Glymour M, Sanchez-Romero R, & Glymour C (2017) A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *Int J Data Sci Anal* 3(2):121-129.
50. Ritchie ME*, et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43(7):e47.

51. Friedman J, Hastie T, & Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 33(1):1-22.
52. Kanehisa M & Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1):27-30.
53. Tozbikian G*, et al.* (2014) Mesothelin expression in triple negative breast carcinomas correlates significantly with basal-like phenotype, distant metastases and decreased survival. *PLoS One* 9(12):e114900.
54. Wang Y, Wang L, Li D, Wang HB, & Chen QF (2012) Mesothelin promotes invasion and metastasis in breast cancer cells. *J Int Med Res* 40(6):2109-2116.
55. Conley SJ*, et al.* (2016) HER2 drives Mucin-like 1 to control proliferation in breast cancer cells. *Oncogene* 35(32):4225-4234.
56. Celis JE*, et al.* (2008) 15-prostaglandin dehydrogenase expression alone or in combination with ACSM1 defines a subgroup of the apocrine molecular subtype of breast carcinoma. *Mol Cell Proteomics* 7(10):1795-1809.
57. Naderi A, Chia KM, & Liu J (2011) Synergy between inhibitors of androgen receptor and MEK has therapeutic implications in estrogen receptor-negative breast cancer. *Breast Cancer Res* 13(2):R36.
58. Gangadharan A*, et al.* (2017) Protein calorie malnutrition, nutritional intervention and personalized cancer care. *Oncotarget* 8(14):24009-24030.
59. Giatromanolaki A*, et al.* (2001) Expression of hypoxia-inducible carbonic anhydrase-9 relates to angiogenic pathways and independently to poor outcome in non-small cell lung cancer. *Cancer Res* 61(21):7992-7998.
60. Buache E*, et al.* (2011) Deficiency in trefoil factor 1 (TFF1) increases tumorigenicity of human breast cancer cells and mammary tumor development in TFF1-knockout mice. *Oncogene* 30(29):3261-3273.
61. Burrows B, Niden A, Fletcher CM, & Jones N (1964) Clinical types of chronic obstructive lung disease in London and in Chicago: a study of one hundred patients. *American Review of Respiratory Disease* 90(1):14-27.
62. Burrows B, Fletcher C, Heard B, Jones N, & Wootliff J (1966) The emphysematous and bronchial types of chronic airways obstruction: a clinicopathological study of patients in London and Chicago. *The Lancet* 287(7442):830-835.

63. Miniño AM & Murphy SL (2012) Death in the united states, 2010.
64. Vogelmeier CF*, et al.* (2017) Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report. GOLD executive summary. *American journal of respiratory and critical care medicine* 195(5):557-582.
65. Lundbäck B*, et al.* (2003) Not 15 but 50% of smokers develop COPD?—report from the obstructive lung disease in Northern Sweden studies. *Respiratory medicine* 97(2):115-122.
66. Casanova C*, et al.* (2011) The progression of chronic obstructive pulmonary disease is heterogeneous: the experience of the BODE cohort. *American journal of respiratory and critical care medicine* 184(9):1015-1021.
67. Renkema TE, Schouten JP, Koëter GH, & Postma DS (1996) Effects of long-term treatment with corticosteroids in COPD. *Chest* 109(5):1156-1162.
68. Antus B, Barta I, Horvath I, & Csiszer E (2010) Relationship between exhaled nitric oxide and treatment response in COPD patients with exacerbations. *Respirology* 15(3):472-477.
69. Cushen B*, et al.* (2018) Clinical and exacerbation characteristics may predict treatment response in acute exacerbations of COPD. (Eur Respiratory Soc).
70. Johnson KM*, et al.* (2018) Heterogeneity in the respiratory symptoms of patients with mild-to-moderate COPD. *International journal of chronic obstructive pulmonary disease* 13:3983.
71. Barnes PJ, Shapiro S, & Pauwels R (2003) Chronic obstructive pulmonary disease: molecular and cellularmechanisms. *European Respiratory Journal* 22(4):672-688.
72. O'donnell DE, Neder JA, & Elbehairy AF (2016) Physiological impairment in mild COPD. *Respirology* 21(2):211-223.
73. Rennard SI & Vestbo J (2008) The many "small COPDs": COPD should be an orphan disease. *Chest* 134(3):623-627.
74. Moore WC*, et al.* (2010) Identification of asthma phenotypes using cluster analysis in the Severe Asthma Research Program. *American journal of respiratory and critical care medicine* 181(4):315-323.
75. Pinto LM*, et al.* (2015) Derivation and validation of clinical phenotypes for COPD: a systematic review. *Respiratory research* 16(1):50.
76. Castaldi PJ*, et al.* (2017) Do COPD subtypes really exist? COPD heterogeneity and clustering in 10 independent cohorts. *Thorax* 72(11):998-1006.

77. Castaldi PJ, *et al.* (2014) Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax* 69(5):416-423.
78. Chang Y, *et al.* (2016) COPD subtypes identified by network-based clustering of blood gene expression. *Genomics* 107(2-3):51-58.
79. Regan EA, *et al.* (2011) Genetic epidemiology of COPD (COPDGene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease* 7(1):32-43.
80. Smedley D, *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res* 43(W1):W589-598.
81. Reese SE, *et al.* (2013) A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal component analysis. *Bioinformatics* 29(22):2877-2883.
82. Vink JM, *et al.* (2017) Differential gene expression patterns between smokers and non-smokers: cause or consequence? *Addiction biology* 22(2):550-560.
83. Huan T, *et al.* (2016) A whole-blood transcriptome meta-analysis identifies gene expression signatures of cigarette smoking. *Human molecular genetics* 25(21):4611-4623.
84. Hobbs BD, *et al.* (2017) Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat Genet* 49(3):426-432.
85. Hao K, *et al.* (2012) Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet* 8(11):e1003029.
86. Vasioukhin V, Bowers E, Bauer C, Degenstein L, & Fuchs E (2001) Desmoplakin is essential in epidermal sheet formation. *Nature cell biology* 3(12):1076.
87. Norman M, *et al.* (2005) Novel mutation in desmoplakin causes arrhythmogenic left ventricular cardiomyopathy. *Circulation* 112(5):636-642.
88. Mathai SK, *et al.* (2016) Desmoplakin variants are associated with idiopathic pulmonary fibrosis. *American journal of respiratory and critical care medicine* 193(10):1151-1160.
89. Poole A, *et al.* (2014) Dissecting childhood asthma with nasal transcriptomics distinguishes subphenotypes of disease. *J Allergy Clin Immunol* 133(3):670-678 e612.
90. McGeachie MJ (2017) Childhood asthma is a risk factor for the development of chronic obstructive pulmonary disease. *Current opinion in allergy and clinical immunology* 17(2):104.

91. Hayden LP, *et al.* (2018) Childhood asthma is associated with COPD and known asthma variants in COPDGene: a genome-wide association study. *Respiratory research* 19(1):209.
92. Tai A, *et al.* (2014) The association between childhood asthma and adult chronic obstructive pulmonary disease. *Thorax* 69(9):805-810.
93. Radder JE, *et al.* (2017) Variable susceptibility to cigarette smoke–induced emphysema in 34 inbred strains of mice implicates Abi3bp in emphysema susceptibility. *American journal of respiratory cell and molecular biology* 57(3):367-375.
94. Sharma A, *et al.* (2018) Integration of Molecular Interactome and Targeted Interaction Analysis to Identify a COPD Disease Network Module. *Scientific reports* 8(1):14439.
95. Annoni R, *et al.* (2012) Extracellular matrix composition in COPD. *European Respiratory Journal* 40(6):1362-1373.
96. Strange RC, Spiteri MA, Ramachandran S, & Fryer AA (2001) Glutathione-S-transferase family of enzymes. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 482(1-2):21-26.
97. Seidegard J, Pero RW, Miller DG, & Beattie EJ (1986) A glutathione transferase in human leukocytes as a marker for the susceptibility to lung cancer. *Carcinogenesis* 7(5):751-753.
98. Hirvonen A, Husgafvel-Pursiainen K, Anttila S, & Vainio H (1993) The GSTM1 null genotype as a potential risk modifier for squamous cell carcinoma of the lung. *Carcinogenesis* 14(7):1479-1481.
99. Harrison D, Cantlay A, Rae F, Lamb D, & Smith C (1997) Frequency of glutathione S-transferase M1 deletion in smokers with emphysema and lung cancer. *Human & experimental toxicology* 16(7):356-360.
100. Lakhdar R, *et al.* (2010) Association of GSTM1 and GSTT1 polymorphisms with chronic obstructive pulmonary disease in a Tunisian population. *Biochemical genetics* 48(7-8):647-657.
101. Cheng SL, Yu CJ, Chen CJ, & Yang PC (2004) Genetic polymorphism of epoxide hydrolase and glutathione S-transferase in COPD. *European Respiratory Journal* 23(6):818-824.
102. Young RP, Hopkins RJ, Hay BA, & Gamble GD (2011) GSTM1 null genotype in COPD and lung cancer: evidence of a modifier or confounding effect? *The application of clinical genetics* 4:137.

103. Dey T*, et al.* (2014) Role of glutathione S transferase polymorphism in COPD with special reference to peoples living in the vicinity of the open cast coal mine of Assam. *PloS one* 9(5):e96739.
104. Raghu VK*, et al.* (2019) Feasibility of lung cancer prediction from low-dose CT scan and smoking factors using causal models. *Thorax*:thoraxjnl-2018-212638.
105. Zapater-Moros A*, et al.* (2018) Probabilistic graphical models relate immune status with response to neoadjuvant chemotherapy in breast cancer. *Oncotarget* 9(45):27586.
106. Trilla-Fuertes L*, et al.* (2018) Molecular characterization of breast cancer cell response to metabolic drugs. *Oncotarget* 9(11):9645.
107. Trainor PJ*, et al.* (2018) Inferring metabolite interactomes via molecular structure informed Bayesian graphical model selection with an application to coronary artery disease. *bioRxiv*:386409.
108. Zierer J*, et al.* (2016) Exploring the molecular basis of age-related disease comorbidities using a multi-omics graphical model. *Scientific reports* 6:37646.
109. Abubakar I, Tillmann T, & Banerjee A (2015) Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 385(9963):117-171.
110. Barnes P & Celli B (2009) Systemic manifestations and comorbidities of COPD. *European respiratory journal* 33(5):1165-1185.
111. Mannino DM, Thorn D, Swensen A, & Holguin F (2008) Prevalence and outcomes of diabetes, hypertension and cardiovascular disease in COPD. *European Respiratory Journal* 32(4):962-969.
112. Sin DD, Anthonisen NR, Soriano JB, & Agusti A (2006) Mortality in COPD: role of comorbidities. *European Respiratory Journal* 28(6):1245-1257.
113. Fabbri L, Luppi F, Beghé B, & Rabe K (2008) Complex chronic comorbidities of COPD. *European Respiratory Journal* 31(1):204-212.
114. Lange P*, et al.* (2015) Lung-function trajectories leading to chronic obstructive pulmonary disease. *New England Journal of Medicine* 373(2):111-122.
115. Zafari Z*, et al.* (2016) Individualized prediction of lung-function decline in chronic obstructive pulmonary disease. *CMAJ* 188(14):1004-1011.

116. Chen S, *et al.* (2017) Risk factors for FEV1 decline in mild COPD and high-risk populations. *International journal of chronic obstructive pulmonary disease* 12:435.
117. Blackwell TS, *et al.* (2014) Future directions in idiopathic pulmonary fibrosis research. An NHLBI workshop report. *American journal of respiratory and critical care medicine* 189(2):214-222.
118. Sisson TH, *et al.* (2010) Targeted injury of type II alveolar epithelial cells induces pulmonary fibrosis. *American journal of respiratory and critical care medicine* 181(3):254-263.
119. Pan L, *et al.* (2001) Type II alveolar epithelial cells and interstitial fibroblasts express connective tissue growth factor in IPF. *European Respiratory Journal* 17(6):1220-1227.
120. Álvarez D, *et al.* (2017) IPF lung fibroblasts have a senescent phenotype. *American Journal of Physiology-Lung Cellular and Molecular Physiology* 313(6):L1164-L1173.
121. Ramos C, *et al.* (2001) Fibroblasts from idiopathic pulmonary fibrosis and normal lungs differ in growth rate, apoptosis, and tissue inhibitor of metalloproteinases expression. *American journal of respiratory cell and molecular biology* 24(5):591-598.
122. Brittan M, *et al.* (2016) A Unique Human Alveolar Macrophage Polarization Phenotype Is Associated With Disease Progression In Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med* 193:A6603.
123. Reyfman PA, *et al.* (2018) Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *American journal of respiratory and critical care medicine* (ja).
124. Xu Y, *et al.* (2016) Single-cell RNA sequencing identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis. *JCI insight* 1(20).
125. Colombo D & Maathuis MH (2014) Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research* 15(1):3741-3782.
126. Chickering DM (2002) Optimal structure identification with greedy search. *Journal of machine learning research* 3(Nov):507-554.
127. Meek C (1997) Graphical Models: Selecting causal and statistical models. (PhD thesis, Carnegie Mellon University).
128. Ramsey J, Glymour M, Sanchez-Romero R, & Glymour C (2017) A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional

graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics* 3(2):121-129.

129. Liu H, Lafferty J, & Wasserman L (2009) The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research* 10(Oct):2295-2328.
130. Zhao T, Liu H, Roeder K, Lafferty J, & Wasserman L (2012) The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research* 13(Apr):1059-1062.
131. Liu H, Roeder K, & Wasserman L (2010) Stability approach to regularization selection (stars) for high dimensional graphical models. *Advances in neural information processing systems*, pp 1432-1440.
132. Shah RD & Samworth RJ (2013) Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75(1):55-80.
133. Anthonisen NR, Connett JE, & Murray RP (2002) Smoking and lung function of Lung Health Study participants after 11 years. *American journal of respiratory and critical care medicine* 166(5):675-679.
134. Tashkin DP*, et al.* (1996) Methacholine reactivity predicts changes in lung function over time in smokers with early chronic obstructive pulmonary disease. The Lung Health Study Research Group. *American journal of respiratory and critical care medicine* 153(6):1802-1811.
135. Chandra D*, et al.* (2012) The relationship between pulmonary emphysema and kidney function in smokers. *Chest* 142(3):655-662.
136. Divo M*, et al.* (2012) Comorbidities and risk of mortality in patients with chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine* 186(2):155-161.
137. Hokanson J*, et al.* (2013) Airway-predominant COPD is associated with diabetes and the metabolic syndrome. *Am J Respir Crit Care Med* 187:A2897.
138. Zhao H, Martinez F, Criner G, & Ramos F (2014) Gastroesophageal reflux disease and chronic obstructive pulmonary disease in spiromics. *Arbor* 1001:4.

139. Agusti A, Sobradillo P, & Celli B (2011) Addressing the complexity of chronic obstructive pulmonary disease: from phenotypes and biomarkers to scale-free networks, systems biology, and P4 medicine. *American journal of respiratory and critical care medicine* 183(9):1129-1137.

140. Chandra D, Palevsky P, & Sciurba FC (2017) EnRAGEed Kidneys in Chronic Obstructive Pulmonary Disease? *Am J Respir Crit Care Med* 195(11):1411-1413.

141. Polverino F*, et al.* (2017) A Pilot Study Linking Endothelial Injury in Lungs and Kidneys in Chronic Obstructive Pulmonary Disease. *Am J Respir Crit Care Med* 195(11):1464-1476.

142. Sukkar MB*, et al.* (2012) RAGE: a new frontier in chronic airways disease. *Br J Pharmacol* 167(6):1161-1176.

143. Yonchuk JG*, et al.* (2015) Circulating soluble receptor for advanced glycation end products (sRAGE) as a biomarker of emphysema and the RAGE axis in the lung. *Am J Respir Crit Care Med* 192(7):785-792.

144. Lomas DA*, et al.* (2008) Evaluation of serum CC-16 as a biomarker for COPD in the ECLIPSE cohort. *Thorax* 63(12):1058-1063.

145. Hodge G, Nairn J, Holmes M, Reynolds PN, & Hodge S (2007) Increased intracellular T helper 1 proinflammatory cytokine production in peripheral blood, bronchoalveolar lavage and intraepithelial T cells of COPD subjects. *Clin Exp Immunol* 150(1):22-29.

146. Rennard SI*, et al.* (2007) The safety and efficacy of infliximab in moderate to severe chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 175(9):926-934.

147. Wells JM*, et al.* (2012) Pulmonary arterial enlargement and acute exacerbations of COPD. *N Engl J Med* 367(10):913-921.

148. Sampson N*, et al.* (2011) ROS signaling by NOX4 drives fibroblast-to-myofibroblast differentiation in the diseased prostatic stroma. *Molecular endocrinology* 25(3):503-515.

149. Bensadoun ES, Burke AK, Hogg JC, & Roberts CR (1996) Proteoglycan deposition in pulmonary fibrosis. *American journal of respiratory and critical care medicine* 154(6):1819-1828.

150. Hattori N*, et al.* (2000) Bleomycin-induced pulmonary fibrosis in fibrinogen-null mice. *The Journal of clinical investigation* 106(11):1341-1350.