NORTHWESTERN UNIVERSITY

Statistical Methods for the Network-Based Analysis of Genomic Data

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Engineering Sciences and Applied Mathematics

By

Sahil D. Shah

EVANSTON, ILLINOIS

September 2019
ABSTRACT

Statistical Methods for the Network-Based Analysis of Genomic Data

Sahil D. Shah

The focus of this thesis is on evaluating, designing, and applying statistical methods that elucidate molecular mechanisms by seeking to understand the pathways that contribute to disease. Chapter 1 introduces the field and motivates the work in this thesis. Chapters 2, 3, and 4 describe original work. Chapter 5 recapitulates our findings in the context of the field.

Chapter 2 outlines a novel evaluation framework for pathway analysis methods. The key idea is that analysis techniques that correctly identify disease-associated pathways should find them across different datasets that are measuring the same underlying conditions. Therefore, we apply eight network-based pathway analysis techniques to ten different ovarian cancer studies that have been curated to ensure comparability and evaluate the methods by their cross-study concordance. This approach allows us to evaluate the method with real (instead of artificial) data.

Chapter 3 presents a new analysis method that integrates expression data and network information in a novel procedure to detect genes that appear to influence nearby genes
with disease-associated dysregulation. Applying our algorithm to real expression data we show that our method is able to identify biologically relevant genes, integrate pathway and expression data, and yield more reproducible results across multiple studies of the same phenotype than competing methods.

Chapter 4 concerns original mouse cell line time-series expression data and the statistical analysis of that data to study the development of acute myeloid leukemia (AML) from severe congenital neutropenia (SCN). From this data, we seek to identify the sources of dysregulation in a mutant GCSF background. The key idea of our approach is to combine our data with an interaction network specific to the context of hematopoiesis. To infer a network that is independent from our data we apply the semi-supervised method iRafNet to data from GCSF responsive cells available through Haemopedia. We then find the genes that are the sources of differential time-course profiles on the network using our method, GeneSurrounder. Together these analyses establish a network-based approach to glean mechanistic insights from transcriptomic data by identifying dysregulated pathways and the sources of dysregulation on those pathways.
Acknowledgments

First and foremost I am incredibly grateful to my advisor, Rosemary Braun, for the opportunity she gave me to train in her lab. Her unwavering support, mentorship and kindness have made my time in her lab a period of tremendous professional and personal growth. I will be using the lessons she taught me for the rest of my career.

I would also like to thank Bill Kath and Danny Abrams for serving on my thesis committee and providing constructive feedback on my thesis. I’d like to thank my labmates, Marta Iwanaszko, Phan Nguyen, Gary Wilk, and Elan Ness-Cohn, and my collaborators, Seth Corey and Marek Kimmel, for working together with me on my research projects.

The faculty, staff, and students of ESAM, Preventive Medicine, NICO, and McCormick have made Northwestern an outstanding place to study and work. In particular, I want to thank the ESAM faculty and my ESAM cohort for their support.

I’d like to thank my family and friends for always listening and offering advice during my studies at Northwestern. Finally, I would like to express my love and gratitude to my parents, Shaila and Divyang Shah, for always believing in me.
Table of Contents

ABSTRACT 3

Acknowledgments 5

Table of Contents 6

List of Tables 8

List of Figures 12

Chapter 1. Introduction 18
  1.1. Modern molecular biology 20
  1.2. Biological networks 21
  1.3. Statistical methods for omics systems biology 23
  1.4. Organization of thesis 28

Chapter 2. Network methods for pathway analysis of genomic data 31
  2.1. Introduction 31
  2.2. Methods 39
  2.3. Results 46
  2.4. Conclusions 57

Chapter 3. Network-based identification of disease genes in expression data 61
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1. Background</td>
<td>61</td>
</tr>
<tr>
<td>3.2. Methods</td>
<td>66</td>
</tr>
<tr>
<td>3.3. Results</td>
<td>76</td>
</tr>
<tr>
<td>3.4. Discussion</td>
<td>84</td>
</tr>
<tr>
<td>3.5. Conclusions</td>
<td>86</td>
</tr>
<tr>
<td>3.6. Supplement</td>
<td>88</td>
</tr>
<tr>
<td>Chapter 4. Identifying transcriptional consequences of impaired GCSF signaling in hematopoiesis</td>
<td>94</td>
</tr>
<tr>
<td>4.1. Introduction</td>
<td>94</td>
</tr>
<tr>
<td>4.2. Methods</td>
<td>97</td>
</tr>
<tr>
<td>4.3. Results</td>
<td>106</td>
</tr>
<tr>
<td>4.4. Discussion</td>
<td>109</td>
</tr>
<tr>
<td>Chapter 5. Conclusions</td>
<td>113</td>
</tr>
<tr>
<td>References</td>
<td>117</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Properties of the various methods, including whether genes are thresholded on $p$-values, restrictions on the type of data and comparisons that can be made, the type of edges used by the model, and the type of null hypothesis tested. 36

2.2 Studies and samples sizes of the data used in this investigation. The data are publicly accessible from GEO [1] and available as part of the curatedOvarianData package. 41

2.3 Package versions and run times (in seconds) for 247 pathways in all ten studies. 47

3.1 Ovarian cancer datasets used in this study: Comparisons were made between low- and high-grade serous ovarian carcinoma using public data. Sample sizes for each group in each dataset are given. The data are publicly accessible and available as part of the curatedOvarianData package [2]. 77

3.2 “Disruptive” disease genes in high-grade ovarian cancer consistently found by GeneSurrounder: At a threshold of $p = 0.05$ and with a diameter of $D = 34$, the Bonferroni corrected threshold is
$-\log_{10}(p) \geq 2.83$. Listed are the genes that pass this threshold in all three studies.

3.3 Correlation between GeneSurrounder results and network/gene statistics: The three columns are the rank correlation between GeneSurrounder results ($p^{GS}$) and network/gene statistics (Degree, Betweenness, and $p^{DE}$) across all genes in each dataset. The Degree and Betweenness are two different network centrality measures. The Degree is the number of connections a node has and the Betweenness is the fraction of shortest paths that passes through the node. $p^{DE}$ is the $p$-value obtained from a standard differential expression $t$-test.

3.4 Cross study concordance of GeneSurrounder results compared to differential expression analysis and LEAN: The columns $p^{GS}$ Cor., $p^{DE}$ Cor., and $p^{LEAN}$ Cor. are the Spearman rank correlations respectively between the results obtained from GeneSurrounder, differential expression analysis, and LEAN for each study pair.

3.5 Bladder cancer datasets used in this study: Comparisons were made between superficial and invasive bladder cancer using public data. Superficial bladder cancer has not grown into the main muscle layer of the bladder and invasive bladder cancer has grown into the main muscle layer of the bladder. Sample sizes for each group in each dataset are given. (GSE19915.GPL5186 originally had 43 superficial samples and 45 invasive samples, but samples with missing data for 25% or more of the
genes were filtered out.) The data are publicly accessible and available as part of the curatedBladderData package [3].

3.6 “Disruptive” disease genes in bladder cancer consistently found by GeneSurrounder: At a threshold of $p = 0.05$ and with a diameter of $D = 34$, the Bonferroni corrected threshold is $-\log_{10}(p) \geq 2.83$.

Listed are the genes that pass this threshold in all three studies.

3.7 Correlation between GeneSurrounder results and network/gene statistics: The three columns are the rank correlation between GeneSurrounder results ($p^{GS}$) and network/gene statistics (Degree, Betweenness, and $p^{DE}$) across all genes in each dataset. The Degree and Betweenness are two different network centrality measures. The Degree is the number of connections a node has and the Betweenness is the fraction of shortest paths that passes through the node. $p^{DE}$ is the $p$-value obtained from a standard differential expression $t$-test.

3.8 Cross study concordance of GeneSurrounder results compared to differential expression analysis and LEAN: The columns $p^{GS}$ Cor., $p^{DE}$ Cor., and $p^{LEAN}$ Cor. are the Spearman rank correlations respectively between the results obtained from GeneSurrounder, differential expression analysis, and LEAN for each study pair.

4.1 Sources of dysregulation found by GeneSurrounder. At a threshold of $p = 0.05$ and with a diameter of $D = 12$, the Bonferroni
corrected threshold is $\log_{10}(p) \geq 2.38$. Listed are the genes that pass this threshold.
List of Figures

1.1 Visual abstract: The focus of this thesis is on evaluating, designing, and applying statistical methods that elucidate molecular mechanisms by seeking to understand the pathways that contribute to disease.

2.1 Fragment of the KEGG cell-cycle pathway (hsa04110). Classical pathway analyses such as GSEA do not take the network topology into account, but rather treat the pathway as a simple list of genes. As a result, changes to a gene such as p53 (red), which has a high degree and a direct influence on a number of other high-degree genes, a large downstream network, and an outgoing connection to a whole other network (the apoptosis pathway) are treated in the same way as changes to a gene such as Bub1 (blue), which has far fewer connections. In contrast, topology-based analyses attempt to incorporate the structure of the network and the relative importance of each gene to the pathway.

2.2 Cross-study concordance for each sub-analysis. For each subcomputation of each method, we show the correlation between pathway p values for all possible study pairs (45 total). Study pairs are ordered along the x axis according to their correlation in gene-level p values, shown
in the top row. Methods are labeled in alternating colors, with the
final/combined \( p \) values denoted in bold. The bottom row of the plot
shows the sum of the sample sizes for each pair of studies, with dark
green being high.

2.3 Cross–study correlations. Each plot displays the cross-study correlation
of the results for each major analysis method. Boxplots within each
frame indicate, for a given method, the distribution of correlations each
study had with the nine other studies. EG, consider the “GSEA.pES”
plot; the blue (leftmost) box indicates the distribution of correlations of
GSEA pathway enrichment score \( p \) values (pES) that study ‘GSE13876’
had with each of the other data sets. The red box indicates the
correlations between ‘GSE14764’ GSEA results and those of other nine
studies, etc. Cross-study correlations of the gene–level statistics are also
shown. Note that the scale on each of the plots is the same.

2.4 Cross–method correlations. Each plot displays the correlation in \( p \)
values amongst different methods applied to each of the data sets.
Here, the boxplots within each frame indicate, for a given study, the
distribution of correlations the results from each method had with the
others. EG, in the top left frame, the blue (left most) box plot indicates
the distribution of correlation between the pathway enrichment score
\( p \) values (pES) vs. the pathway \( p \) values obtained from the other nine
analyses when applied to the GSE13876 data. Note that the scale on
each of the plots is the same.
Number of studies in which a pathway ranks in the top 20% for each analysis. For each subcomputation of each method, we show for each pathway the number of times the pathway was amongst the top 20% most significant in each of the 10 studies. Pathways were only considered significant if they met the 20% cutoff unambiguously; if there were more than 20% of pathways tied for the top spot, none were considered to be meaningfully in the top 20%. Methods are labeled in alternating colors, with the final/combined $p$ values denoted in bold. The number of studies (out of 10 possible) in which the pathway was in the top 20% for that analysis is given by color; black indicates that the method could not give an answer for that pathway (typically a result of gene thresholding leaving no meaningful edges). The $p$ values in the color scale correspond to the probability of that specific overlap assuming 10 Bernoulli trials with $p = 0.2$ success. The 247 pathways are ordered along the $x$ axis by the mean overlap from the final (bolded) analyses, while the bottom row shows the average across all sub-analyses.

$p$-value distributions by method (all pathways, all studies). Depicted are joint and marginal distributions of $-\log_{10}(p)$ values for all pathways in all studies. (Note that higher values are more significant.) In the upper triangle, smoothed scatter plots depict the joint distribution of $-\log_{10}(p)$ for each pair of methods; darker red corresponds to higher density of points. In the lower triangle, Spearman’s rank correlations $\rho$ between the $p$ values obtained from each pair of methods is given, with
positive correlations shown in increasing blue intensity and negative
correlations shown in increasing red intensity (there are no negative
values). Note that because rank correlations provide a measure of
concordance that is independent of the dynamic range of the quantities
being correlated and hence less influenced by outliers, the $\rho$ reported in
the lower triangle may differ from a “by eye” estimate of the correlation
based on the plots in the upper triangle. On the diagonal, the marginal
distributions of $-\log_{10}(p)$ are shown as red histograms, with the
theoretically expected distributions (uniform $p$ under the null) shown as
a black line.

3.1 Overview of GeneSurrounder algorithm. The algorithm
incorporates systems–level information, in the form of a network
model of cellular interactions, with gene expression data to identify
the genes that control disease–associated mechanisms. The algorithm
than identifies “disruptive” genes by assessing the significance of the
combined evidence that (1) a gene has a influence on others in the
network and (2) that its influence is driving disease.

3.2 Procedure for Sphere of Influence. The Sphere of Influence
computation tests if a putative driver gene is more correlated with its
neighbors than a random sample of genes.

3.3 Procedure for Decay of Differential Expression. The Decay of
Differential Expression computation tests if the discordance between
differential expression and distance from the driver gene is greater with the phenotype labels we observe than with a random permutation of the sample labels.

3.4 Illustration of Method. Displayed are the results for the gene MCM2 when our algorithm was applied to Ovarian Cancer Study GSE14764.

(a) shows $-\log_{10}(p_{\text{Sphere}})$ vs the Neighborhood Radius. (b) shows $-\log_{10}(p_{\text{Decay}})$ vs the Neighborhood Radius. (c) shows $-\log_{10}(p_{\text{Combined}})$ vs the Neighborhood Radius. (d) shows the Number of Assayed Genes vs the Neighborhood Radius. In the top three plots, the dashed and dotted lines correspond to a significance level of 0.05 and 0.01 respectively. In the bottom plot, the solid line corresponds to the total number of genes assayed and on the network.

4.1 Workflow for identifying sources of dysregulation amongst GCSF responsive genes

4.2 Displayed is the time course expression profile of OSM for illustrative purposes. Any pattern of observations across four time points can be represented as a combination of linear, quadratic and cubic trends.

4.3 GCSF responsive genes identified from GCSF treated wild-type expression data Displayed is the dependence on time vs dynamic range for 6920 genes computed from the 16 wild-type samples in the GCSF treated data. The horizontal and vertical lines correspond to a threshold of $p = 0.01$ and $\log_2(\text{Fold Change}) = 1$ respectively.
4.4 **iRafNet Importance Scores** Displayed is the iRafNet Importance Score $p$-values vs iRafNet Importance Scores computed for the 532 GCSF responsive genes on the KEGG network. iRafNet computes separate scores for edge $(i,j)$ and edge $(j,i)$ resulting in $532 \times 531 = 282492$ Importance Scores and $p$-values.

4.5 **Largest Connected Component of Updated GCSF responsive genes KEGG network** Displayed is the network structure of the KEGG network of GCSF responsive genes after adding the 51 most significant edges returned by iRafNet and finding the largest connected component. (a) shows the degree distribution. (b) shows the distribution of shortest path lengths.

4.6 **$p$-values from GeneSurrounder** Displayed are the $p$-values from GeneSurrounder. At a threshold of $p = 0.05$ and with a diameter of $D = 12$, the Bonferroni corrected threshold is $\log_{10}(p) \geq 2.38$.

4.7 **Displayed are the results for the gene Polr3g** (a) shows $-\log_{10}(p_{\text{Sphere}})$ vs the Neighborhood Radius. (b) shows $-\log_{10}(p_{\text{Decay}})$ vs the Neighborhood Radius. (c) shows $-\log_{10}(p_{\text{Combined}})$ vs the Neighborhood Radius. (d) shows the Number of Assayed Genes vs the Neighborhood Radius. In the top three plots, the dashed and dotted lines correspond to a significance level of 0.05 and 0.01 respectively. In the bottom plot, the solid line corresponds to the total number of genes assayed and on the network.
CHAPTER 1

Introduction

Statistical techniques have become essential for teasing out knowledge from rapid, highly parallel measurements of biological molecules. The gathering of this ‘omics’ data became possible two decades ago with the development of new technologies that allow biologists to probe thousands of features in molecular biology simultaneously. As a result, biologists have a large (and growing) volume of data that, when analyzed statistically, can provide insights about the functions and interactions of cellular processes. These insights can be applied to generate new hypotheses, design experiments, and ultimately develop therapeutics for human diseases.

This thesis is concerned with the network-based analysis of transcriptomic (i.e. expression) data, a measurement of the abundance of mRNA, which quantifies the activity of thousands of genes simultaneously and can be used to identify dysregulated genes or sets of genes in a condition. The analytical techniques we consider and develop are ‘network-based’ in the sense that they integrate the information about the network of interactions between genes and gene products with transcriptomic data derived from experiment to exploit the systems-level information contained in these networks. The aims of this thesis are to outline and apply an evaluation framework for pathway analysis methods, present a novel procedure to detect genes that appear to influence genes with disease-associated dysregulation, and study the development of acute myeloide leukemia (AML) from severe congenital neutropenia (SCN) by analyzing time-series expression data (Figure 1.1). The
rest of this chapter introduces the field, reviews past studies, and motivates the work in this thesis.

1. **Evaluate**: Outline and apply an evaluation framework for pathway analysis methods

2. **Develop**: Present a novel procedure to detect genes that appear to influence genes with disease-associated dysregulation

3. **Apply**: Study the development of AML from SCN by analyzing time-series expression data

**Figure 1.1. Visual abstract**: The focus of this thesis is on evaluating, designing, and applying statistical methods that elucidate molecular mechanisms by seeking to understand the pathways that contribute to disease.
1.1. Modern molecular biology

1.1.1. The central dogma

The most fundamental machinery in molecular biology is the “central dogma of molecular biology”: DNA contains information which is “transcribed” to RNA and then “translated” into proteins, each of which has a function in maintaining the cell’s structure and regulation [4]. During transcription, nucleotide sequences are read from DNA and then spliced into messenger RNA (mRNA). During translation, ribosomes then scan the nucleotides from the mRNA and “print” proteins using a genetic code that maps sets of nucleotides to specific amino acids. Since each cell in an organism has the same DNA, differences between cells (e.g. shape, function, and health) arise from differences in the expression level of each gene, that is, the amount of mRNA, and hence protein, produced. As the molecular intermediate between DNA and proteins, mRNA abundance can stand in as a proxy for protein abundance, which is the primary functional output of the central dogma.

1.1.2. “Omics” data

After 20 years of technological development, there are now multiple platforms to measure all of the molecules (10^4 to 10^9 markers) involved in each stage of the central dogma [5]. For every type of molecule, this highly parallel data is appended with an “-omics” suffix. For example, genomics data identifies genomic variants and proteomics data measures protein abundance. In this thesis, we are concerned with the analysis of transcriptomic (i.e. expression) data. Transcriptomic data (10^4 markers) can be measured using gene expression arrays or RNA sequencing. The former quantifies fluorescence intensities and the latter quantifies RNA fragment counts, but both (after preprocessing) result in an
expression matrix of genes by samples that reflects the activity (i.e. mRNA abundance) of a gene in a given sample. The growing volume of omics data is contained in annotation databases that organize the data and make it available for researchers. The goals of analyzing omics data to generate insights for biology can be broadly grouped into three categories: “discovering molecular mechanisms” (finding genes or gene sets that are statistically associated with an outcome of interest), “clustering samples” (finding statistical similarities between samples), and “the prediction of an outcome, such as survival or the efficacy of therapy” [6]. Discovering molecular mechanisms can be approached by either seeking to determine the mechanisms by which genomic loci contribute to disease or seeking to understand the pathways that contribute to disease [7]. The goal of our work in this thesis is to discover molecular mechanisms by seeking to understand the pathways that contribute to disease.

1.2. Biological networks

Genes and their products interact in complex networks called pathways and the topological properties of these networks contain systems-level biological information. To integrate this information into the analysis of transcriptomic data, researchers have organized networks into pathway databases and developed statistical techniques to infer the networks from expression data.

1.2.1. Pathway databases

Pathway databases include KEGG [8], BioGRID [9], and Reactome [10]. The Kyoto Encyclopedia of Genes and Genomes (KEGG) [8] database is a well-established database
that organizes genes into hundreds of individual networks corresponding to biological processes and is accessible through the R statistical computing package. The networks are experimentally derived, which enables existing biological knowledge to be incorporated into transcriptomic data analysis. However, networks from pathway databases may be incomplete or independent from the biological context of the analysis.

1.2.2. Network inference

An alternative to using pathway databases is to infer a context-specific network solely from expression data. These methods, such as ARCANE [11] and the lagged ordered LASSO method [12], are called \textit{de novo} analysis techniques and attempt to identify edges in the network based on patterns of coexpression because correlations in gene expression are assumed to be indicative of a regulatory link. However, they do not take advantage of any existing biological information and are prone to errors resulting from noisy or undersampled data. Semi-supervised techniques, such as iRafNet [13] and the postPLSR method [14], take advantage of known regulatory dependencies when inferring networks. As a result, semi-supervised techniques offer a compromise between pathway databases and \textit{de novo} analysis techniques.

1.2.3. Graph theory definitions

To integrate biological networks into statistical methods for analyzing omics data, the networks are modeled mathematically as graphs. Formally, a graph \( G = (V, E) \) consists of a set \( V \) of vertices and a set \( E \) of edges, where elements of \( E \) are pairs \((u, v)\) of vertices belonging to \( V \). If each edge has an ordering so that \((u, v)\) is distinct from \((v, u)\), the
graph is said to be directed. If the edges are weighted, the graph is said to be a weighted graph. One common representation of a graph is the adjacency matrix $A$, where $A_{ij} = 1$ if $(i, j) \in E$ \cite{15}.

There are many statistics that can be derived for a given graph \cite{15}, but three will be of particular importance in this thesis: degree, betweenness, and diameter. Degree and betweenness captures the “importance” of a vertex $v$ in the network. The degree is the number of edges in $E$ incident upon $v$ and betweenness is the fraction of shortest paths that passes through the vertex $v$. The diameter is the value of the longest distance on the graph, where the distance between two vertices refers to the length of the shortest path between them.

1.3. Statistical methods for omics systems biology

The association between molecular mechanisms and disease is formalized through statistical hypothesis testing in which the goal is to assesses the significance of individual genes or sets of genes.

1.3.1. Gene-level analysis

The most common method for detecting disease-gene association from transcriptomic data without taking into account biological networks is via differential expression analysis, in which each gene is independently tested for significant differences in mean expression between phenotypes \cite{16}. However with $p = 10^4$ genes in a dataset, a hypothesis test for each gene at a significance level of $\alpha$ would result in $\alpha \cdot p$ genes found significant (i.e. associated
with the phenotype) by chance alone. This is referred to as the multiple testing problem [17]. There are several approaches to correct for the multiple testing problem [17]. In this thesis, we use the most conservative approach, the Bonferroni method, which makes each individual hypothesis more stringent by rescaling the significance threshold $\alpha$ by the number of tests $M$ so that a test is rejected at a level of $\frac{\alpha}{M}$.

However, cellular processes are governed by networks of genes and their products that together underlie complex phenotypes. Gene–level analysis can miss these crucial multi-gene interactions. As a result, differential expression analysis can fail to provide mechanistic insights and has also been observed to exhibit poor agreement between different studies of the same conditions [18, 19, 20]. To address these challenges, researchers have developed systems–level analysis techniques.

1.3.2. Over-representation analysis

Over-representation analysis techniques identify pathways (groups of genes) that are enriched for association with the phenotype of interest, where enrichment means that more significant genes than expected by chance alone are on the pathway (given the size of the pathway). These methods, such as hypergeometric tests and Gene Set Enrichment Analysis (GSEA) [21], consist of computing gene-level statistics and then either counting the number of differentially expressed genes on a pathway or aggregating the gene-level statistics into a pathway-level statistic. Both approaches however do not take into account the interactions between the genes on the pathway [19].
When the association between molecular mechanisms and disease is formalized through resampling or permutation testing to assess statistical significance, a major methodological choice is the way in which to resample from the data to generate a null distribution. The “competitive” null permutes the rows (genes), which creates random pathways but holds the sample classes fixed. The “self-contained” null permutes columns (samples), which leaves the pathways fixed (and hence biologically plausible), but disrupts the associations between the genes and the sample classes. The former tests if a pathway is more associated with the phenotype than a random collection of genes (with the same phenotype); the latter tests if a pathway is more associated with the phenotype than it (the same pathway) would be with a random classification of samples.

Over-representation methods ignore information about about the network of interactions between genes on a pathway. The pathway network structure, however, is important to biological function. To incorporate network structure into the analysis of gene sets, researchers have developed network-based approaches.

1.3.3. Network-based approaches

Topology methods, such as SPIA and CePa, also identify groups of genes that are enriched for association, but augment over-representation methods with additional information about the network of interactions between the genes. Topology methods that attempt to find disease associated subnetworks within pathways include jActive-Modules, HotNet, and COSINE. These methods integrate systems-level information with expression data to identify groups of related genes.
While pathway analysis techniques integrate systems–level information with omic data to provide functional interpretations of the dataset, the “significant pathways” identified by such analyses often comprise tens or hundreds of genes, making experimental follow-up challenging. Early efforts to identify precise gene targets while incorporating systems-level information, such as ENDEAVOUR [29] and GeneWanderer [30], required prior knowledge of mechanisms known to be associated with the disease. Later analysis techniques, such as a method that uses the Laplacian kernel [31] and an extension of SPIA [32], addressed this issue and do not require knowledge of disease associated mechanisms to identify precise gene targets. However, the latter recomputes network distances using the Laplacian kernel and finds genes based on “neighboring” differential expression, complicating the interpretation of the disease genes it finds. The former considers each pathway separately, but these pathways have artificial (sometimes overlapping) boundaries resulting in the exclusion of genes that would be found on a global (i.e. union of the individual pathways) network.

Integrating systems-level information from biological networks into the analysis of omics data is a powerful way of finding molecular mechanisms and ultimately generating new hypotheses, designing experiments and developing therapeutics for human diseases [33].

1.3.4. Open questions

One challenge with network-based analysis methods is choosing between many promising techniques that address the same problem of finding disease-associated pathways or gene targets. Unlike network inference methods, which can be tested against readily available
data with known solutions (such as the data suites from the DREAM Challenges [34]),
there is no consensus “gold standard” dataset of disease-associated genes against which
network-based analysis techniques can be benchmarked. One recent review [35] compared
network-based methods, but was limited to discussing their methodological features and
did not benchmark the methods against each other. Though there are no “gold standard”
datasets available for these methods, an evaluation framework to benchmark network-
based analysis techniques may prove useful to the community.

Another challenge with network-based analysis methods is identifying specific genes
that are central to a trait and would therefore be targets for experimental follow-up and
therapeutic interventions. The “significant pathways” found by topology methods eluci-
date disease-associated functions, but often comprise tens or hundreds of genes. The first
methods to find individual genes required prior knowledge of disease-associated mecha-
nisms, which is not always available or complete. Current methods do not require any
prior knowledge to identify specific genes, but these methods use biological networks in
a way that complicates the interpretation of the disease genes they find. For example,
methods that consider pathways separately may only find genes with an impact within
the artificial boundaries of the pathway. Similarly, methods that consider the enrichment
of neighborhoods on a global network for differential expression may only find genes in-
volved in dysregulation with its neighbors. An analysis technique that explicitly seeks
to find the sources of changes on a global network may prove useful in identifying gene
targets and the effect they have on the network.
1.4. Organization of thesis

The work in this thesis seeks to address the above open questions. In Chapter 2, we outline a novel evaluation framework for pathway analysis methods and review eight network-based pathway analysis techniques. The goal of pathway analysis techniques is to identify functional systems that are involved in a given phenotype. In the most common approaches, biological pathways are modeled as simple sets of genes, neglecting the network of interactions comprising the pathway and treating all genes as equally important to the pathway’s function. Recently, a number of new methods have been proposed to integrate pathway topology in the analyses, harnessing existing knowledge and enabling more nuanced models of complex biological systems. However, there is little guidance available to researchers choosing between these methods. In Chapter 2, we discuss eight topology-based methods, comparing their methodological approaches and appropriate use cases. In addition, we present the results of the application of these methods to a curated set of ten gene expression profiling studies using a common set of pathway annotations. We report the computational efficiency of the methods and the consistency of the results across methods and studies to help guide users in choosing a method. We also discuss the challenges and future outlook for improved network analysis methodologies.

In Chapter 3, we present a new analysis method that integrates expression data and network information in a novel procedure to detect genes that appear to influence nearby genes with disease-associated dysregulation. A key challenge of identifying disease-associated genes is analyzing transcriptomic data in the context of regulatory networks, while still identifying precise gene targets for follow-up studies. We introduce GeneSurrounder,
an analysis method that integrates expression data and network information in a novel procedure to detect genes that appear to influence nearby genes with disease-associated dysregulation. Applying GeneSurrounder to real expression data, we show that our method is able to identify biologically relevant genes, integrate pathway and expression data, and yield more reproducible results across multiple studies of the same phenotype than competing methods. Together these findings suggest that GeneSurrounder provides a new avenue for identifying precise gene targets.

In Chapter 4, we consider original mouse cell line time-series expression data and the statistical analysis of that data to study the development of acute myeloid leukemia (AML) from severe congenital neutropenia (SCN). Granulocyte colony stimulating factor (GCSF) is a hematopoietic growth factor naturally produced in humans that stimulates the bone marrow to produce granulocytes. It is administered therapeutically to patients with severe congenital neutropenia (SCN) to avoid life threatening infections. These patients, however, have often been observed to develop acute myeloid leukemia (AML). We hypothesize that a mutant GCSF receptor (GCSFR) triggers system wide perturbations that lead to the clonal evolution to cancer. In this chapter, we present expression data of mouse cell lines with a truncated and wild type receptor treated with GCSF and then assayed at four different time points. From this data, we seek to identify the sources of dysregulation in a mutant GCSF background. The key idea of our approach is to combine our data with an interaction network specific to the context of hematopoiesis. To infer a network that is independent from our data we apply the semi-supervised method iRafNet to data from GCSF responsive cells available through Haemopedia. We then find the genes that are the sources of differential time-course profiles on the network using
GeneSurrounder. Our findings suggest promising new gene targets for follow-up studies and could help to identify prognostic markers and therapeutic targets for patients with SCN at risk of developing AML.
CHAPTER 2

Network methods for pathway analysis of genomic data

2.1. Introduction

Modern high-throughput (HT) technologies enable researchers to make comprehensive measurements of the molecular state of biological samples and have yielded a wealth of information regarding the association of genes with specific phenotypes. However, the complex and adaptive nature of living systems presents a significant challenge to deriving accurate and predictive mechanistic models from genomic data. Because cellular processes are governed by networks of molecular interactions, critical alterations to these systems may arise at different points yet result in similar phenotypes. At the same time, the adaptability and robustness of living systems enables variations to be tolerated. Typical gene-level analyses of HT data, such as tests of differential expression, are unable capture these effects. As a result, there has been growing interest in systems-level analyses of genomic data.

Pathway analysis techniques, which aim to examine HT data in the context of mechanistically related gene sets, have been enabled by the growth of databases describing functional networks of interactions. These include KEGG [36], BioCarta [37], Reactome [38], the NCI Pathway Interaction Database (NCI-PID) [39], and InnateDB [40], amongst others. To address the challenge of querying these databases using a common

\[1\text{This chapter has been adapted from [20].}\]
framework, markup languages such as KGML (used by KEGG) and BioPAX have been
developed to describe pathways using a consistent format. In particular, the Biological
Pathway Exchange (BioPAX) project now provides a unified view of the data from many
of the above sources [41], including NCI-PID, Reactome, BioCarta, and WikiPathways.

Over the past decade, a number of pathway analysis methods have been developed
to integrate this information with data derived from genomic studies [19]. These meth-
ods can be broadly grouped into two categories. The first category comprises analyses
designed to identify pathways in which significant genes are overrepresented. A compre-
hensive review of these methods was recently published Khatri &al [19]; examples include
hypergeometric tests, Gene Set Enrichment Analysis (GSEA) [42], and Ingenuity Path-
way Analysis [43]. The second category of approaches use dimension reduction algorithms
to summarize the variation of the genes across the pathway and test for pathway–level
differences without relying on single–gene association statistics. Examples include GPC-
Score [44], Pathifier [45], and PDM [46]. In contrast to enrichment analyses like GSEA,
these methods are capable of identifying differences at the systems–level that would be in-
detectable by methods that rely upon single–gene association statistics, such as differences
in the coordination of the expression of two genes.

Despite these advances, the majority of these methods treat pathways as simple lists
of genes, neglecting the network of interactions codified in pathway databases (cf. Fig 2.1)
despite the fact that the importance of pathway network structure to biological function
has long been appreciated. In [47], the authors presented systematic mathematical anal-
ysis of the topology of metabolic networks of 43 organisms representing all three domains
Figure 2.1. Fragment of the KEGG cell-cycle pathway (hsa04110). Classical pathway analyses such as GSEA do not take the network topology into account, but rather treat the pathway as a simple list of genes. As a result, changes to a gene such as p53 (red), which has a high degree and a direct influence on a number of other high–degree genes, a large downstream network, and an outgoing connection to a whole other network (the apoptosis pathway) are treated in the same way as changes to a gene such as Bub1 (blue), which has far fewer connections. In contrast, topology–based analyses attempt to incorporate the structure of the network and the relative importance of each gene to the pathway.

of life, and found that despite significant variation in the pathway components, these networks share common mathematical properties which enhance error-tolerance. In [48], the authors compared the lethality of mutations in yeast with the positions of the affected protein in known pathways, and found that the biological necessity of the protein was well modeled by its connectivity in the network.

To incorporate known interaction network topology with traditional pathway analyses, a multitude of approaches have been proposed for overlaying gene–specific data (either the raw data itself or \( p \)-values derived from gene–level statistical tests) onto pathway networks [49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68]. However, to date few comparisons between them have been made. One recent
review [35] attempted to compare three network–based analyses (SPIA, PARADIGM, and PathOlogist); unfortunately, the comparison was stymied by the methods’ disparate implementations and reference databases, and yielded inconclusive results. To address this gap, we review eight topology–based pathway analysis methods [42, 49, 69, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60]. While this is not a comprehensive review of all such algorithms, the methods we consider have the common feature of being implemented in R and permitting the user to provide the pathway models (such as those obtained from KEGG), thus allowing them to be compared directly without the issues encountered in [35]. Their free availability from CRAN and BioConductor [70] also makes them the most popular network analysis methods in the bioinformatics and computational biology research community.

We briefly describe the features and limitations of these methods. In addition, we also provide a comparison of these methods applied to a curated set of gene expression data from 10 ovarian cancer studies [71], using a common set of 247 KEGG pathways. Using this data, we were able to evaluate both the computational efficiency of the methods and the consistency of the results between the methods and across the studies. The suite of data and scripts used in this comparison are available from our website, enabling researchers to compare updated versions and new network analysis packages using this common framework.
2.1.1. Overview of Network Pathway Analysis Methods and Methodological Considerations

We consider 8 popular network-based pathway analysis methods. In the following discussion, we consider a pathway to be a network of genes (nodes), where edges represent a biochemical interaction. These edges may be directed (e.g., gene $i$ induces gene $j$, but not vice-versa) or signed (e.g., $+1$ for activation or induction, $-1$ for repression or inhibition). Also, while these methods were designed with differential expression analyses in mind, it should be noted that most are flexible enough to accommodate other statistics or data types (e.g., allele frequencies and $\chi^2$ statistics from GWAS). A summary of the key features of the methods is given in Table 2.1. For comparison, we also consider GSEA [42], a commonly–used pathway analysis algorithm that does not incorporate network topology. An outline of these methods is provided in our paper [20]. For details, see the respective publication of each method.

In Table 2.1, the design features of each method are listed. A major distinction between the methods is the need for gene $p$-value thresholding. Threshold–free analyses are generally considered preferable to those that require gene–level significance thresholds, since the threshold introduces an arbitrary choice and excludes the full spectrum of data from the analysis; for this reason, GSEA is preferred to hypergeometric enrichment tests [42, 19] for non-network pathway analyses. Of the network–based methods, SPIA, PathNet, and NEA have the common drawback of requiring thresholding on gene–level significance; in contrast, CePa and ROT/pe provide threshold–free options, and the DEGraph and TopologyGSA analyses are threshold–free by design. However, it should be noted that the cutoff–free ROT/pe analysis differs substantially from the cutoff based
Table 2.1. Properties of the various methods, including whether genes are thresholded on p-values, restrictions on the type of data and comparisons that can be made, the type of edges used by the model, and the type of null hypothesis tested. Notes: 1 “Expression data” is used here to denote any data that meets the assumptions used in gene expression testing, i.e., that the data is continuous and normally distributed. Other data meeting these assumptions can also be used, but methods which have this restriction cannot accept SNP data, etc. 2 “Signed” edges have signs assigned based on the interaction type in the pathway reference graph, distinguishing activating/inducing edges (+1) from repressing/inhibiting edges (-1). 3 “Null type” refers to the type of null hypothesis tested (cf. [19]. “Competitive” null hypotheses compare the pathway of interest to randomly generated pathways, without permuting sample labels, i.e., testing that the pathway statistic is more significant than a random pathway given the sample labels. “Self-contained” null hypotheses compare the statistic for the pathway to that obtained by randomly permuting sample labels, i.e., testing that the pathway is more strongly associated with a particular phenotypic attribute than expected by chance given the genes and topology of the pathway. The self-contained null is considered to be a stronger test [19]. 4 CePa-GSA is not inherently restricted to 2-sample tests of differential expression methodologically, however the present implementation will only carry out the non-thresholded “GSA” type analysis for 2-sample tests of differential expression. As currently implemented, other types of analyses (e.g., using SNP data or modeling survival) must be carried out in CePa using the “ORA” analysis, which requires gene p-value thresholding and can only test the competitive null hypothesis3.

<table>
<thead>
<tr>
<th>Method</th>
<th>Gene p-value thresholding</th>
<th>Expression(^1)</th>
<th>2-sample restriction</th>
<th>Directed</th>
<th>Signed(^2)</th>
<th>Null type(^3)</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSEA</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>N/A</td>
<td>N/A</td>
<td>self-contained</td>
<td>[42]</td>
</tr>
<tr>
<td>SPIA</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>competitive</td>
<td>[49, 69, 50]</td>
</tr>
<tr>
<td>ROT/pe</td>
<td>optimal</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>competitive</td>
<td>[50, 51, 52]</td>
</tr>
<tr>
<td>PathNet</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>competitive</td>
<td>[53]</td>
</tr>
<tr>
<td>NEA</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>competitive</td>
<td>[54, 55]</td>
</tr>
<tr>
<td>CePa-ORA</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>competitive</td>
<td>[56, 57]</td>
</tr>
<tr>
<td>CePa-GSA</td>
<td>no</td>
<td>written(^4)</td>
<td>written(^4)</td>
<td>as</td>
<td>yes</td>
<td>self-contained</td>
<td>[56, 57]</td>
</tr>
<tr>
<td>DEGraph</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>optional</td>
<td>self-contained</td>
<td>[58]</td>
</tr>
<tr>
<td>TopologyGSA</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>self-contained</td>
<td>[59, 60]</td>
</tr>
</tbody>
</table>

\(^1\) “Expression data” is used here to denote any data that meets the assumptions used in gene expression testing, i.e., that the data is continuous and normally distributed. Other data meeting these assumptions can also be used, but methods which have this restriction cannot accept SNP data, etc.

\(^2\) “Signed” edges have signs assigned based on the interaction type in the pathway reference graph, distinguishing activating/inducing edges (+1) from repressing/inhibiting edges (-1).

\(^3\) “Null type” refers to the type of null hypothesis tested (cf. [19]). “Competitive” null hypotheses compare the pathway of interest to randomly generated pathways, without permuting sample labels, i.e., testing that the pathway statistic is more significant than a random pathway given the sample labels. “Self-contained” null hypotheses compare the statistic for the pathway to that obtained by randomly permuting sample labels, i.e., testing that the pathway is more strongly associated with a particular phenotypic attribute than expected by chance given the genes and topology of the pathway. The self-contained null is considered to be a stronger test [19].

\(^4\) CePa-GSA is not inherently restricted to 2-sample tests of differential expression methodologically, however the present implementation will only carry out the non-thresholded “GSA” type analysis for 2-sample tests of differential expression. As currently implemented, other types of analyses (e.g., using SNP data or modeling survival) must be carried out in CePa using the “ORA” analysis, which requires gene p-value thresholding and can only test the competitive null hypothesis.
ROT/pe analysis. Specifically, ROT/pe with a cutoff computes both the downstream perturbation analysis (pPert) and a hypergeometric over-representation analysis (pORA) which are then combined to measure the impact of differential expression on a pathway, while in the cutoff–free analysis, only the downstream perturbations pPert are considered. This limitation could be overcome simply by using a GSEA-like analysis in place of the hypergeometric test, enabling both pORA and pPert to be computed with or without a threshold.

The second major methodological distinction is in the type of hypothesis being tested by the methods. “Competitive” null hypotheses compare the pathway of interest to a random pathway while holding the sample classes (or gene–level associations) fixed, whereas “self-contained” null hypotheses test if the pathway is more strongly associated with a particular phenotypic attribute than expected by chance given the genes and topology of the pathway [19]. The two tests represent two different conditional probabilities (“competitive” being conditioned on the sample labels, allowing the definition of the pathway to vary; the “self–contained” being conditioned on the pathway definition, but allowing the sample classes to vary), and may thus give different results. The “self–contained” null is considered superior since it is both better justified biologically (“competitive” permutations tests will create physiologically unrealistic pathways) and directly answers the question of whether a particular pathway is associated with the phenotype of interest. Unfortunately, though, methods testing the “self–contained” null tend also to be limited in the type of data that can be used: DEGraph, TopologyGSA, and CePa-GSA are limited to two–sample comparisons of continuous data, making them unsuitable for survival analysis or application to GWAS SNP data. However, while this limitation is inherent
to the distributional assumptions made in DEGraph (which uses Hotelling’s $T^2$ test) and TopologyGSA (which uses a Gaussian network model), it is only an implementation limitation in CePa-GSA rather than a methodological constraint. A revision of the CePa package with a more flexible interface would provide a threshold–free, “self–contained” network analysis tool that could be applied to a broad variety of studies.

2.1.2. Evaluating the Performance of Network Analysis Methods

These network analysis methods have different features that make them better suited to some use cases than others. Nevertheless, for many common analyses, most of these approaches could be applied, and the user is faced with a choice between several promising methods. Unfortunately, benchmark tests to systematically evaluate the performance of network analysis methods remain lacking, limiting the community’s ability to compare methods.

The development of a systematic evaluation framework faces a number of challenges. First, the methods themselves have highly disparate implementations, often using different databases and pathway semantics, making them difficult to compare in a consistent way. For example, PathOlogist \cite{67,66} treats pathways as bipartite graphs of genes and interactions in contrast to gene–mode networks and is restricted in its implementation to pathways from the NCI/PID \cite{39} database. Secondly, it is not clear what the “gold standard” for these methods should be. Unlike machine learning and network inference problems which are readily tested against simulated benchmark data with known solutions (such as the in-silico data suites used in the DREAM Challenges \cite{34}), there is no agreement on what the “correct” results of these analyses would be.
To provide intuition regarding the performance of the methods reviewed here, we systematically applied them to a curated suite of gene expression data from 10 ovarian cancer studies [71] using a common set of pathway definitions obtained from the KEGG database. The goal of these tests was not only to supply provisional guidance about the relative performance of the methods, but also to suggest a strategy for a testing framework for pathway analysis methods.

2.2. Methods

Our approach is motivated by the observation that systems-level analyses improve the concordance of results between different studies of the same phenotype [72, 18]. Although multiple studies of the same phenotype may yield very different lists of significant genes, pathway analyses tend to show much greater agreement. This effect is not unexpected, considering the complexity of biological systems [73] and the noisiness of HT data; individual disease-associated genes may be detected in some studies but miss the significance threshold in others. However, if a pathway is functionally related to the disease, we may reasonably expect to detect its association across multiple studies, even if the specific genes contributing to its significance vary from one study to the next.

This observation leads to the following conjectures: If a specific pathway is functionally related to a particular phenotype, we expect that some manifestation of its involvement will be present in the data for all studies of that disease, and that an accurate and sensitive network analysis approach will detect those signals consistently across the studies. A poor network analysis method, on the other hand, will yield results that are strongly influenced by noise in the data, and hence will detect pathways that are particular to
each study rather than the common biological signal. On the basis of this conjecture, we use the cross–study concordance of each method’s results to measure its ability to detect a common (and presumably “true”) signal in each of the studies.

2.2.1. Ovarian Cancer Data

For the purposes of our analysis, we used gene expression and clinical information from curatedOvarianData [71], an expert–curated collection of uniformly prepared microarray data and documented clinical metadata from 23 ovarian cancer studies totalling 2970 patients. The curatedOvarianData project was designed to facilitate gene expression meta-analysis as well as software development. By providing a consistent representation of data that has been processed to ensure comparability between studies, the package enables users to immediately analyze the data without needing to reconcile different microarray technologies, study designs, expression preprocessing methods, or clinical data formats.

Because several of the methods under consideration were limited to two–sample comparisons, we selected data sets with sample classes that could be meaningfully dichotomized. Since the vast majority of the samples came from patients with stage III cancers, tumor/normal and stage–based comparisons were not feasible; instead, we chose to compare low– and high–grade ovarian serous carcinomas. These grades have distinct histological features, molecular characteristics, and clinical outcomes [74, 75]. Low–grade serous carcinomas typically evolve slowly from adenofibromas, acquiring over time frequent mutations to KRAS, BRAF, or ERBB2 genes, but not TP53 mutations. In contrast, high–grade serous carcinomas are characterized by TP53 mutations, often without mutations to KRAS, BRAF, or ERBB2. They arise from unknown precursor lesions, progress rapidly,
and have worse clinical outcomes. For this analysis, we selected studies with a minimum of 15 high- and low-grade serous carcinomas and 1000 genes assayed, keeping only the patients who fell into those categories and who had survival data. 10 of the 23 available studies met these criteria. The study accession numbers and sample counts are given in Table 2.2.

<table>
<thead>
<tr>
<th>Study Accession No.</th>
<th>grade</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N(low)</td>
<td>N(high)</td>
</tr>
<tr>
<td>GSE13876_eset</td>
<td>59</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>GSE14764_eset</td>
<td>24</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td>GSE17260_eset</td>
<td>67</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td>GSE30161_eset</td>
<td>19</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>GSE32062_GPL6480_eset</td>
<td>131</td>
<td>129</td>
<td></td>
</tr>
<tr>
<td>GSE32063_eset</td>
<td>23</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>GSE9891_eset</td>
<td>103</td>
<td>154</td>
<td></td>
</tr>
<tr>
<td>PMID17290060_eset</td>
<td>57</td>
<td>57</td>
<td></td>
</tr>
<tr>
<td>PMID19318476_eset</td>
<td>17</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>TCGA_eset</td>
<td>75</td>
<td>470</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2. Studies and samples sizes of the data used in this investigation. The data are publicly accessible from GEO [1] and available as part of the curatedOvarianData package.

The microarray data was filtered to keep only the genes common to all 10 studies; no other filtering was done. This resulted in 7680 genes common to all 10 studies.

To obtain the gene–level statistics required by several of the analyses, the R limma package [76] was used. log₂ fold changes were used for the magnitude of differential expression when required; the significance of the association was quantified using the p value for the empirical Bayes estimated t statistic [76]. Where thresholds for significance were needed, the 0.05 most significant genes were selected. (NB, this corresponds to the
0.05 quantile of significance, not \( p = 0.05 \). Because the studies varied considerably in their sample sizes, and hence power, we chose to use a quantile–based threshold rather than a \( p \) value threshold to render them comparable. While the \( p \) value for the 0.05 quantile varied from study to study, in all cases this corresponded to \( p \ll 0.05 \).

### 2.2.2. Network Models

In order to ensure that each of the eight methods tested used a common, comparable set of pathway definitions, we created the pathway annotation objects required for each method by hand from a fresh download of the KEGG pathway database \[36\]. The KEGGgraph R package \[77\] was used to obtain the pathway KGML files for 247 human pathways. The KGML files were first processed into R \texttt{graphNEL} objects for use by ROntoTools/PathwayExpress (ROT/pe), TopologyGSA, and DEGraph. The \texttt{graphNEL} objects were then used to generate the lists of genes, edges, and adjacency matrices variously required by GSEA, PathNet, NEA, and CePa. The \texttt{path.info} data used by SPIA was also generated from the KEGG \texttt{graphNEL} objects and written to disk as required by SPIA. In this way, we ensured that the set of pathways considered by each method would be directly comparable to each other.

### 2.2.3. Application

Each of the methods shown in Table \[2.1\] was applied to all 10 data sets described in Table \[2.2\] for all 247 KEGG pathways in our database. Where permutation tests were required, 1000 permutations were used. Several of the methods have options that permit
different styles of analysis, which we also explored. The details of our application is given below:

**GSEA:**

As a point of reference, the non–networked GSEA was applied as described in [78], using the gene \( p \)-values obtained from limma as specified in [20]. Significance was tested using 1000 permutations of the sample classes.

**SPIA:**

SPIA was applied using an 0.05–quantile threshold for significance as specified in [20]. A number of the pathways had no edges considered by SPIA (which only considers directed edges, cf. [49, 69]), and so were preemptively excluded from the analysis by the package. The overrepresentation and perturbation \( p \)-values were combined using Stouffer’s normal–inverse method.

**ROT/pe, cutoff:**

ROT/pe permits analysis with and without a \( p \) value threshold [50, 51, 52]; we applied both. Here, we used the same threshold as in SPIA, meaning that the results should be roughly comparable to SPIA for the \( p \).ORA overrepresentation analysis. In contrast to SPIA, however, \( p \).Pert is now weighted by the gene’s significance, rather than treating all significant genes equally. Genes not meeting the significance threshold are excluded from \( p \).Pert with 0 weight, similar to the exclusion in SPIA. \( p \).Pert and \( p \).ORA were combined as in SPIA.

**ROT/pe, no cut:**

We performed ROT/pe without a significance threshold. Because the hypergeometric test cannot be performed without setting a cut-off, only \( p \)Pert is reported.
In contrast to the thresholded analysis, pPert now involves data from all the genes, although those with low significance will have low weighing.

**PathNet:**
PathNet was carried out as specified in [20] and in [53]. PathNet returns both the PathNet $p$ value combining the “direct” and “indirect” evidence, along with the simple hypergeometric $p$ value. The quantity of interest is $p_{\text{PathNet}}$.

**NEA:**
NEA was carried out as specified in [20] and in [54, 55], using the same gene thresholds and number of permutations as in the other studies.

**CePa-ORA:**
Like ROT/pe, CePa also has options to perform the analyses with or without setting a gene–significance threshold [56, 57]. We performed both; here, we use the same thresholds used in the other analyses to carry out CePa-ORA. As specified in [20], CePa will report the significance using a variety of network centrality measures. Because there is no clear choice of which one is correct, we chose to combine all six measures into a single $p$ value for CePa-ORA using Stouffer’s method.

**CePa-GSA:**
We also performed the non-thresholded CePa-GSA. The analysis differs from CePa-ORA not only in the number of genes considered, but also in the type of hypothesis test performed. While CePa-ORA tests a “competitive” hypothesis [19] comparing pathways to random subsets of genes while holding fixed the
sample labels (and hence the gene–level statistics), CePa-GSA tests the “self–
contained” null hypothesis by permuting the sample labels while holding the
pathway definition fixed (see Table 2.1). The two tests are orthogonal to each
other, and we do not anticipate that the results of CePa-GSA will necessarily be
the same as those for CePa-ORA. As in CePa-ORA, we chose to combine the six
p’s into one p value for CePa-GSA using Stouffer’s method.

DEGraph:

DEGraph also presents several alternative analysis approaches, specifically, whether
or not the network should be signed (corresponding to inversely–related nodes)
or unsigned for the purposes of computing the smoothing vector \[58\]. We per-
formed both the signed and unsigned analyses. For each of these, we also had to
make a decision regarding how to handle pathways with more than one connected
component, and hence more than one p-value. We tried both simply taking the p
value for the largest connected component as the p value for the pathway, as well
as combining the p values for all the components using Stouffer’s method. DE-
Graph will report both the non-networked \(p(T^2)\) as well as the graph–smoothed
\(p(T^2)_{\text{graph}}\); the latter is the primary quantity of interest.

TopologyGSA:

Finally, we attempted to apply TopologyGSA \[59, 60\] as implemented to our
data. In principle, TopologyGSA reports p values for both differential variance
and differential mean expression across the pathway submodules, both of which
are of interest.
2.3. Results

2.3.1. Computational efficiency

The computational time to complete each of the analyses on a desktop machine (3.4 GHz quad-core Intel Core i7 iMac with 16GB RAM) is given in Table 2.3. Each of the methods shown in Table 2.1 was applied to all 10 data sets described in Table 2.2 for 247 pathways. Where permutation tests were required, 1000 permutations were used. All jobs completed the calculation in under an hour per study with the exception of NEA, which required $\sim 2.5$ hrs/study, and TopologyGSA, which failed to complete even the first analysis when it was finally halted after 100hrs (>4 days). Interestingly, ROT/pe, which is a weighted modification of the same computation carried out in SPIA, required less time than SPIA (and, additionally, was able to treat more pathways), which we attribute to code improvements by the authors of both methods \cite{49, 69, 50, 51, 52} and made ROT/pe amongst the fastest of the methods we tested. Nevertheless, with most methods taking only a few minutes per study, the differences in computational cost between them are minor.

With the exception of TopologyGSA and DEGraph, the major computational cost is due to permutation testing. (In DEGraph, the computation of the smoothing vector scales as $\mathcal{O}(n^3)$ where $n$ is the number of genes in the pathway graph, and so can be cumbersome for very large pathways; however, DEGraph does not require permutation tests.) Permutation testing is trivially parallelizable, and the development of parallel R libraries such as snow \cite{79} facilitates development of packages that can be run on clusters. Yet, of the packages considered here, only CePa provides a parallelized implementation.
The extreme computational cost of TopologyGSA was due to a step in the algorithm that involves identifying the maximal clique in a moralized graph, a problem that scales exponentially with the network size and is thus intractable for even modest pathways. For further discussion, see [20].

2.3.2. Analyses

We posit that if a pathway is functionally related to a particular phenotype (here, high vs. low grade ovarian cancer), we expect that a manifestation of its involvement will be present in the data for all studies of that disease, and that an accurate and sensitive network analysis approach will detect those signals consistently across the studies while a poor network analysis method will yield results that are strongly influenced by noise in

<table>
<thead>
<tr>
<th>Method</th>
<th>Package Version</th>
<th>time (sec), 247 pathways, all 10 studies</th>
<th>avg runtime per study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>user</td>
<td>system</td>
</tr>
<tr>
<td>SPIA</td>
<td>SPIA_2.14.0</td>
<td>1229.803</td>
<td>123.825</td>
</tr>
<tr>
<td>ROT/pe, cutoff</td>
<td>ROntoTools_1.4.0</td>
<td>948.702</td>
<td>9.497</td>
</tr>
<tr>
<td>ROT/pe, no cut</td>
<td>ROntoTools_1.4.0</td>
<td>1054.795</td>
<td>16.861</td>
</tr>
<tr>
<td>PathNet</td>
<td>PathNet_1.2.0</td>
<td>747.983</td>
<td>9.572</td>
</tr>
<tr>
<td>NEA</td>
<td>neaGUI_1.0.0</td>
<td>78099.148</td>
<td>5280.723</td>
</tr>
<tr>
<td>CePa-ORA</td>
<td>CePa_0.5</td>
<td>3217.099</td>
<td>21.367</td>
</tr>
<tr>
<td>CePa-GSA</td>
<td>CePa_0.5</td>
<td>29127.875</td>
<td>512.618</td>
</tr>
<tr>
<td>DEGraph (s,u)</td>
<td>DEGraph_1.14.0</td>
<td>2467.075</td>
<td>43.743</td>
</tr>
<tr>
<td>TopologyGSA</td>
<td>topologyGSA_1.4.3</td>
<td>[N/A: job halted after 100hrs]</td>
<td>&gt;360000.000</td>
</tr>
</tbody>
</table>

Table 2.3. Package versions and run times (in seconds) for 247 pathways in all ten studies. Average times (in h/m/s) for a single study (i.e., 1/10th wallclock) are also noted. Times for computation (“user”) and kernel system calls (“system”) are given; the total CPU time consumed is the sum of these. The computations were carried out on a lightly-loaded 3.4 GHz quad-core Intel Core i7 iMac with 16GB RAM running R version 3.0.3 under OS X 10.8.5 (Darwin 12.5.0). Note that the times for DEGraph include both the signed and unsigned graph analyses. 1000 random permutations were used for the methods that perform resampling (all except DEGraph).
the data and will vary strongly from one study to another. Based on this intuition, we consider the cross–study concordance of each method’s results to measure its ability to detect a common (and presumably “true”) signal in each of the studies.

**Cross–study concordance.** For each method tested, we examined the correlation of \( p \)-values obtained for the 247 pathways between all 45 possible pairs of the 10 studies listed in Table 2.2. For reference, we also began by examining the correlation in gene–level statistics between the studies. Because each method has a different range/resolution of possible \( p \) values owing to the different uses of parametric and nonparametric tests, we use the nonparametric Spearman’s rank correlation as a measure of concordance.

Figure 2.2 depicts the correlation of results for all all 45 study pairs. For completeness, we show the cross–study correlations for each method’s subanalyses, in addition to the “final” combined result. This includes the \( p \)ORA and \( p \)Pert values for SPIA and ROT/pe (both with and without a cutoff) that are combined to form the final \( p \)Comb value for the pathway; the standard \( p \)Hyper hypergeometric test along with the network edge–based hypergeometric test from PathNet; the results for all six centrality measures considered by CePa ORA and GSA analyses, which are then combined into \( p \)Comb; and the DEGraph results for the standard Hotelling \( T^2 \) and the network–smoothed \( pT2graph \) for the largest connected component (lcc) and the full pathway (\( p \)’s for all connected components combined) for both the signed and unsigned analyses. The “final” results for each analysis are denoted by bolding. To aid in the interpretation of Fig. 2.2, Fig. 2.3 presents a summary of the cross–study correlations for each of the “final” results. Here, the distribution of correlations for each of the 10 studies with respect to all other studies is shown for each of the methods.
Figure 2.2. Cross–study concordance for each sub-analysis. For each sub-computation of each method, we show the correlation between pathway p values for all possible study pairs (45 total). Study pairs are ordered along the x axis according to their correlation in gene–level p values, shown in the top row. Methods are labeled in alternating colors, with the final/combined p values denoted in bold. The bottom row of the plot shows the sum of the sample sizes for each pair of studies, with dark green being high.

As can be seen in the top row of Fig. 2.2 and first panel of Fig. 2.3, the correlation of gene–level statistics is often poor; we obtained a maximum rank correlation $\rho = 0.20$ for the gene p-values, with a median of $\sim 0.02$. This lack of correspondence even amongst studies of the same phenotype has been observed in other investigations of microarray data [18]. Also as previously observed in other studies [72, 18], we found that the
Figure 2.3. Cross–study correlations. Each plot displays the cross-study correlation of the results for each major analysis method. Boxplots within each frame indicate, for a given method, the distribution of correlations each study had with the nine other studies. EG, consider the “GSEA.pES” plot; the blue (leftmost) box indicates the distribution of correlations of GSEA pathway enrichment score $p$ values ($pES$) that study ‘GSE13876’ had with each of the other data sets. The red box indicates the correlations between ‘GSE14764’ GSEA results and those of other nine studies, etc. Cross-study correlations of the gene–level statistics are also shown. Note that the scale on each of the plots is the same.

correlations were generally improved in pathway analyses, both in terms of the number of positively–correlated study pairs (fewer blue cells in Fig. 2.2) and in the magnitude of the positive correlations (cf Figs. 2.2,2.3). Considering that each study is interrogating gene expression in the same phenotypes, we ideally desire that the correlations between all study pairs should be positive, and indeed many were, with the CePa GSA analyses being the strongest and most consistent (cf Fig. 2.3).
In addition, we also observe that the concordance for the network–based analyses was always stronger than that exhibited by non–network analyses. In Fig. 2.2 this can be seen in the darker blue cells of GSEA and both SPIA and ROT/pe pORA sub-analysis values, all of which measure enrichment without considering the network topology. In Fig. 2.3 this difference is manifest by the lower medians and tails for the GSEA boxplot versus the others.

It is also instructive to consider the cross–study correlations for each method in the context of the gene–level concordance and the sample size of the studies. The study pairs depicted in Fig. 2.2 are sorted in order of gene–level concordance. Generally, we expect that studies which have greater similarity at the gene level will also exhibit greater similarity in the pathway statistics, and in most cases this pattern holds, with lower correlations at the left end of the plot. Qualitative departures from this trend can be seen in the SPIA, NEA, and CePa–ORA betweenness results. Moreover, we expect that two large studies will exhibit greater correlation than two smaller studies, on the basis of the intuition that two large studies are better powered to distinguish subtle but biologically meaningful, and hence consistent, effects from noise. In Fig. 2.2 the sum of the sample sizes for each study pair is given in the bottom row, and the pattern of study sizes can be seen to be most strongly manifest in the various CePa analyses.

Consistency between methods. A similar intuition regarding the concordance of findings may be had regarding the results of the various methods within a given study. That is, despite the differences between the methods, we might reasonably expect that if a pathway is strongly affected in a particular study, it will be detected by more than one of these methods, i.e., that for a given study the results from SPIA, ROT/pe, NEA,
Figure 2.4. Cross–method correlations. Each plot displays the correlation in p values amongst different methods applied to each of the data sets. Here, the boxplots within each frame indicate, for a given study, the distribution of correlations the results from each method had with the others. EG, in the top left frame, the blue (left most) box plot indicates the distribution of correlation between the pathway enrichment score p values (pES) vs. the pathway p values obtained from the other nine analyses when applied to the GSE13876 data. Note that the scale on each of the plots is the same.

etc. will be correlated. Figure 2.4 shows, for each study, the distribution of pathway p–value correlations that each method had with all other methods. We use Spearman’s rank correlation as a measure of concordance to account for methodological differences influencing the dynamic range of p values. NEA analyses tended to give results that disagree with methods obtained from the other analyses (cf Fig. 2.4), while the other methods are fairly comparable.
While the above comparisons consider correlations for the whole range of \( p \) values returned by each method, we now consider how consistent the selection of “significant” pathways is for the various methods. The heatmap in Figure 2.5 depicts the number of times that a pathway was ranked in the top 20% for the 10 studies listed in Table 2.2 that a pathway was ranked in the top 20%. Results from all subanalyses of each method are given, including the \( p \)ORA and \( p \)Pert values for SPIA and ROT/pe (with and without a cutoff) that are combined to form the final \( p \)Comb value for the pathway; the standard \( p \)Hyper hypergeometric test and the network edge–based hypergeometric test from PathNet; the results for all six centrality measures considered by CePa ORA and GSA analyses that are then combined into \( p \)Comb; and DEGraph results for the standard Hotelling \( T^2 \) test \( p \)T2 and the network–smoothed \( p \)T2graph for the largest connected component (lcc) and the full pathway (\( p \)’s for all connected components combined) for both the signed and unsigned analyses. The “final” results for each analysis are denoted by bolding. The pathways are sorted by the average across the bolded analyses, such that (eg) CePa–GSA contributes to that average once as opposed to seven times. The probabilities for the counts under the null hypothesis are also shown; our expectation is that we will see, by chance, counts of 0–4 with 95% probability, while counts > 5 should occur only once by chance amongst the 247 pathways. On the other hand, pathways for which the null hypothesis is indeed false should exhibit high counts more frequently, corresponding to their detection in multiple studies. Generally, this pattern appears to qualitatively hold, and the pathways which are deemed significant in > 5 studies tend to detected consistently by most of the methods (with the exception of NEA). This suggests that the results are being
driven by commonalities across studies, and that those common patterns are detectable by many of the methods considered.

Distribution of results. A more detailed understanding of these patterns emerges from looking at the concordance between methods for all pathways in all studies, shown in Fig. 2.6. In the upper triangle of plots, the joint distributions of $-\log_{10} p$ values are reported; the corresponding correlation coefficients are shown in the lower triangle. Note that rank correlation coefficients are reported (and so may differ from “by eye” estimates). On the diagonal, histograms of $-\log_{10} p$ are plotted in red; the black lines show the expected distribution of $-\log_{10} p$ corresponding to the uniform distribution of $p$ values expected under the null. Strikingly, it can be seen from the histograms in Fig. 2.6 that the $p$ values obtained from NEA are strongly biased toward highly significant $p$ values; indeed, over half of all NEA $p$Z values fall $\leq 10^{-3}$, whereas we expect that the proportion would be $\sim 1/1000$. This causes an extremely large fraction of pathways to be deemed statistically significant even after adjusting for multiple hypotheses in NEA, making it difficult to discriminate truly significant pathways using the current implementation. Such severely skewed $p$-value distributions are generally attributable to an incorrect null model.

The other methods follow the theoretical $p$ value distributions relatively well, with a slight deviation observed in CePa–GSA.

---

1In the case of NEA, we believe there may be a simple remedy for the incorrect null model. Specifically, we note that the NEA package does not distinguish genes that are assayed and deemed non-significant from genes those that are simply not assayed. By treating non-assayed genes as insignificant, the proportion of significant genes is significantly reduced, lowering the probability of significant edges in the resampled graphs.
Figure 2.5. Number of studies in which a pathway ranks in the top 20% for each analysis. For each subcomputation of each method, we show for each pathway the number of times the pathway was amongst the top 20% most significant in each of the 10 studies. Pathways were only considered significant if they met the 20% cutoff unambiguously; if there were more than 20% of pathways tied for the top spot, none were considered to be meaningfully in the top 20%. Methods are labeled in alternating colors, with the final/combined \( p \) values denoted in bold. The number of studies (out of 10 possible) in which the pathway was in the top 20% for that analysis is given by color; black indicates that the method could not give an answer for that pathway (typically a result of gene thresholding leaving no meaningful edges). The \( p \) values in the color scale correspond to the probability of that specific overlap assuming 10 Bernoulli trials with \( p = 0.2 \) success. The 247 pathways are ordered along the \( x \) axis by the mean overlap from the final (bolded) analyses, while the bottom row shows the average across all sub-analyses.
Figure 2.6. p-value distributions by method (all pathways, all studies). Depicted are joint and marginal distributions of $-\log_{10}(p)$ values for all pathways in all studies. (Note that higher values are more significant.) In the upper triangle, smoothed scatter plots depict the joint distribution of $-\log_{10}(p)$ for each pair of methods; darker red corresponds to higher density of points. In the lower triangle, Spearman’s rank correlations $\rho$ between the $p$ values obtained from each pair of methods is given, with positive correlations shown in increasing blue intensity and negative correlations shown in increasing red intensity (there are no negative values). Note that because rank correlations provide a measure of concordance that is independent of the dynamic range of the quantities being correlated and hence less influenced by outliers, the $\rho$ reported in the lower triangle may differ from a “by eye” estimate of the correlation based on the plots in the upper triangle. On the diagonal, the marginal distributions of $-\log_{10}(p)$ are shown as red histograms, with the theoretically expected distributions (uniform $p$ under the null) shown as a black line.
2.4. Conclusions

New network–based methods have garnered increasing interest as tools to analyze complex genomic datasets at the systems level. Despite the development of a number of promising tools, however, there is little guidance available to researchers for choosing between the methods. In this review, we sought to compare all the network analysis methods available in R/BioConductor at the time of this writing. In addition to discussing their methodological and implementation features, we also proposed and applied a novel means to compare their performance using a suite of curated microarray data-sets and a set of updated KEGG mappings developed to enable consistent pathway models for each method. The results of our tests clearly indicated the benefits and limitations of each approach. The tests also revealed idiosyncracies that would have been unnoticed except in comparison; for example, our comparisons revealed a bug in the previous version of the ROT/pe computation, which led us to suggest a fix that has now been implemented in the current version (reviewed here).

In addition to providing guidance about the features of the methods (Table 2.1), the efficiency of the computations (Table 2.3), and the consistency of the results (Figs. 2.2–2.6), our review also suggests a number of directions for future methodological development.

Most notably, there is a need for benchmark and testing standards against which network analysis methods should be tested. We used the consistency of the results across a set of comparable studies, but this approach is plagued by a serious limitation: namely, we have no way to assess whether the “consistent” results are consistent owing to biological commonalities amongst ovarian cancers or due to a fluke of the microarray data, since the
set is homogenous with respect to the disease type. A more insightful analysis could be obtained by the development of a database of diverse studies that are all curated to the same standards, just as was done for the curated ovarian data [71]. While diverse datasets are readily obtained, the work required to ensure that they are all comparable is non-trivial (and was beyond the scope of this paper); however, such data would be immensely useful to the research community. Relatedly, agreement on a common pathway representation format such as BioPAX [41] and developers’ adoption of a consistent API accepting these pathway files would aid comparison between these methods without requiring that the pathways be prepared by the user in different ways.

Secondly, we note that the most significant methodological distinctions between the packages involve a choice between using the preferred “self-contained” null hypothesis versus having the flexibility to apply the method in contexts other than two-sample differential expression studies. We recommend using methods that test the self-contained null (both for statistical and biological reasons), but at present none of these packages are able to test, for example, a self-contained hypothesis that a pathway is significantly associated with survival. This compromise could be easily resolved by further development of CePa–GSA allowing the user indicates to the function the statistical test (or model to be fit) rather than assuming that a two-sample \( t \)-test is desired. In the case of DEGraph, which uses Hotelling’s two-sample \( T^2 \) statistic to compare the graph “smoothed” gene expressions in two phenotypes, such an extension is less obvious but would be a valuable addition to DEGraph’s functionality.

Relatedly, we note that care must be exercised when constructing null models for the pathways for the “competitive” tests. An easy check of whether or not the null model is
correct is to examine the distribution of $p$ values across a large set of pathways; strong deviation from the expected uniform distribution of $p$ values is indicative of an incorrect null model. However, this rough assessment will only reveal egregious flaws. In many network biology methods, null graphs are generated by simply resampling node or edge properties. This destroys the correlation structure in the data (as [42] discussed) as well as the expected assortativity of gene expression in the pathway, yielding excessively conservative null models. There is thus a need to develop methods that can produce null graphs that are more biologically plausible.

Finally, we observe that a common drawback to all of these methods is their reliance upon single–gene statistical tests. As a result, while all of these methods are able to articulate differences in gene expression that have an impact at the pathway level, they cannot detect differentially regulated pathways when there are no detectable marginal effects at the gene level. An alternative approach would be to overlay the gene expression data itself onto the network (instead of using statistics corresponding to the gene’s differential expression), obtain a summary statistic for the network as a whole, and compare those. This approach has proved powerful in a non–network context [44, 72], where it was able to detect pathways in which non-linear patterns of gene expression were associated with phenotype. While network extensions have been proposed [65], R implementations remain lacking.

Network analysis is rapidly becoming a valuable tool for harnessing existing biological information to yield mechanistic, systems–level insights from HT data. A number of promising methods have been developed, and we have found that most yield more consistent results (as measured by cross–study concordance) than both gene–level analyses
and non-network pathway analyses (GSEA [42]). Nevertheless, challenges remain, and further work in this area has the potential to significantly improve the systems-based analysis of HT data, facilitating better understanding of the structure and function of the complex networks that coordinate living processes.
CHAPTER 3

Network-based identification of disease genes in expression data

3.1. Background

The advent of high-throughput transcription profiling technologies has enabled identification of genes and pathways associated with disease, providing new avenues for precision medicine. A key challenge is to analyze this data in the context of the regulatory networks that control cellular processes, while still obtaining insights that can be used to design new diagnostic and therapeutic interventions. It is thus necessary to develop methods that analyze omic data in the context of the full network of interactions, while still providing specific, targetable gene-level findings.

The most common method for detecting gene-association is via differential expression analysis, in which each gene is independently tested for significant differences in mean expression between phenotypes [16]. However, while differential expression analysis can identify specific (and hence targetable) disease-associated genes, it does not take into consideration the network of molecular interactions that govern cellular function, limiting the mechanistic insights that can be derived from the data. As a result, this analysis can miss crucial multi-gene interactions that underlie complex phenotypes. Since biological systems are complex and expression data is typically noisy, the multi-gene mechanisms that underlie a disease may be detectable across multiple studies, but the individual genes

---

1This chapter has been adapted from [80]. R code for the proposed algorithm is available at [https://github.com/sahildshahi/gene-surrounder](https://github.com/sahildshahi/gene-surrounder).
contributing to those mechanisms may vary from one study to the next. As a result, differential expression analysis can exhibit poor agreement between different studies of the same conditions [18, 19, 20].

Maps of experimentally derived molecular interaction networks contained in pathway databases and the growth of analysis techniques that infer context-specific interaction networks have enabled the development of methods that integrate systems level information with expression data. KEGG [8], for example, is a well-established pathway database that organizes genes into hundreds of individual networks corresponding to biological processes. One use of interaction networks has been to identify groups of related genes underlying a biological mechanism. By incorporating systems-level information, these pathway analysis techniques can capture multi-gene interactions, yielding mechanistically interpretable results that are more reliable than single-gene analyses [18, 19, 20]. Pathway analysis techniques can be broadly grouped into three categories: ‘functional scoring methods’, ‘topology methods’, and ‘active modules tools.’ Functional scoring methods, such as GSEA [21], identify groups of genes that are enriched for association with the phenotype of interest. Topology methods, such as SPIA [23] and CePa [24, 25], also identify groups of genes that are enriched for association, but augment functional scoring methods with additional information about the network of interactions between the genes. Active modules tools, such as jActiveModules [26], HotNet [27], and COSINE [28], attempt to find disease associated subnetworks within pathways. These methods integrate systems-level information with expression data to identify groups of related genes.

While pathway analysis techniques integrate systems-level information with omic data to provide functional interpretations of the dataset, the “significant pathways” identified
by such analyses often comprise tens or hundreds of genes, making experimental follow-up challenging. Additionally, boundaries between pathways are often arbitrary, thus potentially neglecting key interactions. Moreover, many techniques rely on user-settable parameters and ad-hoc heuristics that depend on network size, limiting their interpretability and reliability [20, 81]. Together, these issues point to the need for analysis techniques that integrate network and omics data to identify precise gene targets for follow-up studies.

Early efforts to identify precise gene targets while incorporating systems-level information include ENDEAVOUR [29] and GeneWanderer [30]. ENDEAVOUR takes in as input various data sources (such as literature abstracts and protein-protein interactions) and prioritizes genes based on their similarity to genes known to be involved in the disease. GeneWanderer uses protein-protein interaction networks and identifies gene targets based on distance to known disease genes on the network. However, these methods require knowledge of mechanisms known to be associated with the disease. Later analysis techniques – such as a method that uses the Laplacian kernel [31], an extension of SPIA [32], and nDGE [82] – addressed this issue and do not require knowledge of disease associated mechanisms to identify precise gene targets. The first method uses a protein association network, recomputes distances using the Laplacian kernel, and finds disease genes based on “neighboring” differential expression. Since the distances are recomputed, the neighbors could include genes that are not neighbors on the original network. In other words, this method uses indirect interactions instead of direct interactions, complicating the interpretation. In the extension of SPIA [32], disease genes are found by propagating changes in expression along the edges of the individual pathway so that each gene is
scored for disease-association according to its own change in expression combined with the change in expression of its upstream neighbors. Since each pathway is considered separately and the pathways have artificial (sometimes overlapping) boundaries, an individual pathway could exclude genes that are on a global network (i.e. union of the individual pathways). nDGE takes in as input expression profiles and for a putative disease gene class conditionally identifies its co-regulated and actively co-regulated neighbors. While powerful, each of these is limited in its treatment of the networks. These methods either do not consider direct interactions between genes on a global network ([31] uses indirect interactions based on the Laplacian kernel and [32] considers each KEGG pathway separately) or infer interactions based on correlations (e.g., [32]). Thus due to the limitations of the previously described techniques, an analysis technique that takes into account direct interactions between genes globally may prove useful in identifying targets and the effect they have on the network.

Most recently, LEAN [83] was developed to use direct interactions on a global interaction network and find disease genes by scoring the differential expression of “local subnetworks.” LEAN scores each gene for disease-association according to the enrichment of its immediate neighbors. Thus, LEAN’s algorithm restricts its focus to a local subnetwork that only considers nearest neighbors. As a result, LEAN only identifies genes based on the changes in expression of a given gene’s local subnetwork, but cannot determine whether that gene is actually the source of changes in its neighborhood or on the network; an important characteristic when looking for potential targetable disease genes for use in precision medicine.
The goal of the present work is to combine pathway network information with gene expression data to determine the degree to which a gene is a source of dysregulation on the network. We present a novel analysis technique, GeneSurrounder, that takes into account the complex structure of interaction networks to identify specific disease-associated genes from expression data. The key idea of our method is to score genes based on the evidence that they influence the dysregulation of their neighbors on the network in a manner that impacts cell function. In this way, the genes returned by our method may be considered sources of “disruption” on the network and therefore candidate targets for therapeutics. We thus seek to identify genes with two defining characteristics: 

(i) they appear to influence other genes nearby in the network, as evidenced by strongly correlated expression with nearby genes; and (ii) their dysregulation is associated with disease, as evidenced by a pattern of (progressively weaker) differential expression centered about that gene. By finding these genes, our method identifies candidate genes that are “disruptive” to the mechanisms underlying a given phenotype and does so without any reliance on user-set parameters or arbitrary pathway boundaries.

In this manuscript, we describe the GeneSurrounder algorithm and apply it to data from three independent studies of ovarian cancer to demonstrate its use, evaluate the reproducibility of its results, and demonstrate the methodological and biological validity of our approach. In order to evaluate the algorithm, we evaluate its cross-study concordance, i.e., its consistency across different data sets measuring the same phenotype. We compare the cross-study concordance of GeneSurrounder’s results to that of standard differential expression analysis, and find that genes identified as sources of pathway disruption by
GeneSurrounder are more consistently identified across the various studies than are differentially expressed genes. We also compare our method to LEAN, and show that genes identified by GeneSurrounder are more consistent across studies than both LEAN and differential expression analysis. We demonstrate that our method represents an integration of pathway and expression data to yield results that are not solely driven by either alone and find that it identifies genes associated with ovarian cancer. Together, these results suggest that GeneSurrounder reproducibly detects functionally-relevant genes by integrating gene expression and network data. Our novel analysis approach complements existing gene– and pathway–based analysis strategies to identify specific genes that control disease–associated pathways, providing a new strategy for identifying promising therapeutic targets.

3.2. Methods

Our goal is to identify candidate disease genes by analyzing gene expression data in the context of interaction networks to discover genes that drive the behavior of pathways associated with disease. We thus seek to identify genes with two defining characteristics: (i) they appear to influence other genes nearby in the network, as evidenced by strongly-correlated expression with nearby genes; and (ii) their dysregulation is associated with disease, as evidenced by a pattern of differential expression centered about that gene. Since the ‘extent’ of dysregulation of a given gene on a global gene network is not known a priori, we score the gene separately for every neighborhood size on the network (i.e. genes one ‘hop’ away, genes up to two ‘hops’ away, etc) and then return the results for the
highest scoring neighborhood. Genes with significantly high-scoring neighborhoods may then be prioritized for follow-up experiments.

To this end, the GeneSurrounder method consists of two tests that are run independently of each other (Figure 3.1) and then combined, for every neighborhood size on the network. To determine if the putative disease gene is a “disruptive” candidate disease gene meeting both criteria, the results for the highest scoring neighborhood are returned. To prioritize genes, our method is applied exhaustively to each assayed gene in a transcriptomic data set, and the results from each gene’s highest scoring neighborhood are compared to rank the genes.

Figure 3.1. **Overview of GeneSurrounder algorithm.** The algorithm incorporates systems–level information, in the form of a network model of cellular interactions, with gene expression data to identify the genes that control disease–associated mechanisms. The algorithm than identifies “disruptive” genes by assessing the significance of the combined evidence that (1) a gene has a influence on others in the network and (2) that its influence is driving disease.
The algorithm takes as input gene expression data and a network model of cellular interactions derived from a pathway database. In order to consider the full scope of a gene’s interactions and avoid artificially imposed pathway boundaries, we create a global KEGG network by merging the individual pathways so that the links which are in at least one KEGG pathway will be part of the new global network (i.e., the graph union of all pathways). We then consider the largest connected component of the resulting network in our algorithm. Using this global network and gene expression data, we compute evidence for each of the above criteria as follows.

**Does a gene appear to influence its neighbors in the network? Evidence of “Sphere of Influence”**

If a gene is a source of regulatory control or disruption, we may expect to see that its behavior is correlated with that of its neighbors. The first step, dubbed “Sphere of Influence,” assesses if a candidate gene $i$ meets this criterion by testing if gene $i$ is more strongly correlated with its network neighbors than would be expected by chance (Figure 3.2), compared to a random sample of genes. The first step, therefore, of the Sphere of Influence procedure is to calculate the Spearman rank correlation $\rho_{ij}$ between gene $i$ and every other gene $j$ assayed and on the network. From this set of correlations, we calculate the observed total (absolute) correlation between gene $i$ and its neighbors within a neighborhood of radius $r$,

\begin{equation}
C_i(r) = \sum_{\{j : d_{ij} \leq r\}} |\rho_{ij}|,
\end{equation}

where $d_{ij}$ indicates the geodesic distance of gene $j$ from gene $i$ on the network.
Figure 3.2. Procedure for Sphere of Influence. The Sphere of Influence computation tests if a putative driver gene is more correlated with its neighbors than a random sample of genes.

In order to compute the distribution of total correlation under the null hypothesis that it is drawn from a random sample of genes, we re-sample with replacement from the set of correlations between gene $i$ and every other gene $j$ and recompute Equation 3.1. This procedure effectively redistributes the gene–gene correlations about the network, enabling a comparison of gene $i$’s influence in the true network neighborhood to its influence on a random selection of genes. This step tests the so-called “competitive null” described in [19]; that is, whether gene $i$ has a greater correlation with genes in its neighborhood than would be expected from a random set of genes.

The null distribution of the total absolute correlation for gene $i$ as a function of the neighborhood radius is computed using $10^3$ re-samplings, and the observed total
absolute correlation is compared to the re-sampled null distribution, yielding a “Sphere of Influence” $p$–value at each neighborhood radius for gene $i$, $p_i^S(r)$, that quantifies whether $i$ is more correlated with its neighbors than expected by chance, evidence that it may be an influential gene.

Does the gene’s neighborhood exhibit an association with phenotype? Evidence of “Decay of Differential Expression”

The previous step tests whether gene $i$ is strongly correlated with its network neighbors, independent of phenotype. If a gene is a source of disease-associated disruption, we may expect that it and its neighbors will exhibit differential expression. We thus now turn our attention to whether the gene and its neighbors also exhibit an association with the phenotype of interest. In particular, if a gene $i$ is a source of dysregulation that drives the phenotype, we would expect that gene $i$ and its close neighbors will be differentially expressed, while genes farther away in the network will exhibit weaker differential expression. In other words, we expect a pattern of differential expression that is strongly localized about $i$ and decays as one moves farther from it in the network.

Hence, the second calculation, “Decay of Differential Expression,” tests whether the magnitude of differential expression of other genes $j$ in the neighborhood is inversely related to the distance $d_{ij}$ of gene $j$ from gene $i$ (Figure 3.3).

In order to do this, we must first compute a gene–level statistic $g_j$ that quantifies the magnitude of $j$’s association with the outcome of interest. We then quantify the “decay of differential expression” with the Kendall $\tau_B$ rank correlation coefficient between the differential expression and distance from gene $i$. 
Figure 3.3. **Procedure for Decay of Differential Expression.** The Decay of Differential Expression computation tests if the discordance between differential expression and distance from the driver gene is greater with the phenotype labels we observe than with a random permutation of the sample labels.

The observed discordance is

\[
D_i(r) = \tau_B \left( \{g_j : d_{ij} \leq r\}, \{d_{ij} : d_{ij} \leq r\} \right),
\]

where \(d_{ij}\) is the geodesic distance between gene \(j\) and gene \(i\).

To assess the statistical significance of \(D_i(r)\), we randomly permute the phenotype labels and recompute the gene–level association statistics \(g_j\) under the null hypothesis that the genes are not meaningfully associated with the phenotype. We then use the permuted \(g_j^\ast\) to recompute \(D_i^\ast\) according to Equation 3.2. A set of \(10^3\) such re-computations forms...
a reference distribution against which we compare the observed $D_i$ to obtain a $p$ value $p_i^D(r)$ as the fraction of $D_i^* < D_i$.

It should be noted here that while $p_i^S(r)$ (above) was obtained by randomly permuting genes, $p_i^D(r)$ is obtained by permuting the class labels. An important feature of the latter is that it preserves correlations between genes that were found in the $p_i^S(r)$ calculation. In consequence, the null models, and hence the interpretations, of the two tests differ. $p_i^S(r)$ quantifies whether the neighborhood surrounding gene $i$ is more strongly correlated with it than a random set of genes would be (independent of phenotype), testing the so-called “competitive null” [19]. In contrast, $p_i^D(r)$ assesses whether the neighborhood surrounding gene $i$ is more strongly associated with the phenotype of interest than those same genes would be with randomly-assigned phenotype labels (preserving the organization of genes in the network), thus testing the so-called “self-contained null” [19]. That is, it tests whether a specific set of genes in a neighborhood is more strongly associated with the phenotype of interest than the same set of genes would be for a random phenotype.

Because these two procedures permute orthogonal axes (genes vs. samples), they provide two independent tests with independent interpretations: $p_i^S(r)$ tests whether gene $i$ influences its neighbors, and $p_i^D(r)$ tests whether that neighborhood is associated with disease. We then combine these independent pieces of evidence into a single assessment, as described below.

**Combined Evidence**

At this point in our algorithm, the Sphere of Influence and Decay of Differential Expression procedures have been run independently of each other, but neither component is sufficient
by itself to determine if putative disease gene $i$ is in fact a “disruptive” candidate disease gene meeting both criteria. Therefore, the last step our method performs is to combine the $p$-values outputted by each component ($p_i^S(r)$ and $p_i^D(r)$) using Fisher’s method \[84,\]

\[
X^2 = -2(\ln(p_i^S(r)) + \ln(p_i^D(r))).
\]

$X^2$ follows a $\chi^2$ distribution with 4 degrees of freedom, which can be used compute $p_i^{\text{Comb}}(r)$, the combined evidence that gene $i$ is a “disruptive” gene.

**Neighborhood Size**

Above we described the Sphere of Influence and Decay of Differential Expression procedures for a fixed radius ($r$), but different genes may have different extents of influence on the network, and this extent is not known \textit{a priori}. Therefore, we have devised our analysis technique to apply the Sphere of Influence, Decay of Differential Expression, and Combined Evidence calculations to the neighborhood of every radius (up to $D$ the diameter of the network). The $p$-value our method outputs for each gene ($p_i^{GS}$), therefore, is the smallest $p_i^{\text{Comb}}(r)$ across all distances.

To evaluate the significance of $p_i^{GS}$, we then apply a Bonferroni correction to the significance threshold to conservatively adjust for the multiple hypothesis tests that we perform when applying our method to the neighborhoods of each radius. Since the number of neighborhoods (and therefore number of tests) is determined by the diameter of the network, we scale the significance threshold by the diameter of the network to determine whether a gene was significantly found to be “disruptive” in the data. Adjustment for the multiplicity of genes tested is achieved through permutation as previously described \[85,\]
this has the important benefit of preserving the biologically-relevant dependency structure between genes [86, 21].

Example of GeneSurrounder steps applied to an example gene

To illustrate the components of the GeneSurrounder computation, we present the results for each component of our algorithm as applied to gene MCM2 using data from one study of high-vs-low grade ovarian cancer [2] (GEO accession GSE14764). In Figure 3.4 each of the first three plots (from top to bottom) displays the $-\log_{10}(p)$ from the Sphere of Influence, Decay of Differential Expression and Combined components of our method. Since we compute these values as a function of network neighborhood size surrounding that gene, the $p$-values are plotted against the neighborhood radius (i.e. radius of geodesic distance from the putative “disruptive” disease disease gene MCM2.)

Figure 3.4A (Sphere of Influence) illustrates the dilution of influence with distance and the effect that the size (i.e. number of assayed genes) of a neighborhood has on the decrease of influence. The putative disease gene in this example, MCM2, has significant influence in neighborhoods near to it, but this influence falls off and stays non-significant at far-away distances. The largest difference occurs between a radius of 5 and 6, where the number of assayed genes within the neighborhood (Figure 3.4D) increases sharply, contributing to the dilution of MCM2’s influence.

Figure 3.4B (Decay of Differential Expression) indicates a significant concentration of differential expression for neighborhoods with radii of 4–6. We observe that small neighborhoods immediately near a putative disease gene are not big enough to detect a decaying pattern of differential expression, such that the localized differential expression
Figure 3.4. Illustration of Method. Displayed are the results for the gene MCM2 when our algorithm was applied to Ovarian Cancer Study GSE14764. (a) shows $-\log_{10}(p_{\text{Sphere}})$ vs the Neighborhood Radius. (b) shows $-\log_{10}(p_{\text{Decay}})$ vs the Neighborhood Radius. (c) shows $-\log_{10}(p_{\text{Combined}})$ vs the Neighborhood Radius. (d) shows the Number of Assayed Genes vs the Neighborhood Radius. In the top three plots, the dashed and dotted lines correspond to a significance level of 0.05 and 0.01 respectively. In the bottom plot, the solid line corresponds to the total number of genes assayed and on the network.
is only detectable at with a radius of at least 4. At the other end, big neighborhoods are too diverse to exhibit a consistent decay of differential expression; like the sphere of influence, the significance of the decay of differential expression flattens out at large distances.

Figure 3.4C illustrates the results of combining the results for each neighborhood. The p-value our method outputs for each gene is the most significant $p^\text{Comb}(r)$ across all neighborhood radii; for $MCM2$ in this study, this occurs at a neighborhood radius of 4 with $p_{GS} = 1.48e-05$. Since our method returns the smallest $p^\text{Comb}(r)$ for each gene (equivalently, the largest $-\log_{10} p^\text{Comb}(r)$) and the smallest $p^\text{Comb}(r)$ of $MCM2$ is highly significant, $MCM2$ would be identified as a central candidate disease gene of high grade ovarian cancer. From a biological standpoint, this finding is sensible: $MCM2$ is a DNA replication factor, and therefore likely plays a role in the aggressive proliferation associated with high-grade ovarian carcinoma.

3.3. Results

Application to Ovarian Cancer Data with Global KEGG Network Model

We apply our algorithm to three gene expression data sets of high-vs-low grade ovarian cancer from the publicly available and curated collection ‘curatedOvarianData’ [2] to illustrate the components of the GeneSurrounder method and evaluate its performance. In order to test our algorithm, we evaluate its cross-study concordance, i.e., its consistency across different data sets that are measuring the same conditions, as previously described [20]. The intuition underlying this approach is that methods that detect true
biological signals should find them across different data sets measuring the same conditions. To test this we use data from three independent studies of gene expression in high and low grade ovarian cancer tumors (Table 3.1). The data were obtained from the Bioconductor package `curatedOvarianData` [2], a project designed to facilitate meta-analysis by providing data that has been harmonized to ensure that clinical measurements (such as grade) are comparable across studies. Gene expression data was preprocessed by the original authors using established normalization techniques, and no further preprocessing was required. Following our previous work [20], we confine our analysis to genes assayed in common across all datasets and which appear in the KEGG network; a total of 2709 genes meet these criteria.

<table>
<thead>
<tr>
<th>GEO Accession No.</th>
<th>N(low-grade)</th>
<th>N(high-grade)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE14764</td>
<td>24</td>
<td>44</td>
</tr>
<tr>
<td>GSE17260</td>
<td>67</td>
<td>43</td>
</tr>
<tr>
<td>GSE9891</td>
<td>103</td>
<td>154</td>
</tr>
</tbody>
</table>

Table 3.1. **Ovarian cancer datasets used in this study**: Comparisons were made between low- and high-grade serous ovarian carcinoma using public data. Sample sizes for each group in each dataset are given. The data are publicly accessible and available as part of the curatedOvarianData package [2].

Our method combines two independent sources of information — the gene expression data and a pathway network model — to detect the disruptive genes of the phenotype under consideration. We use the same global network model for each study, which we have constructed from KEGG pathways [8]. The KEGG database organizes experimentally derived pathway information into individual networks of functionally related molecules. In the KEGG representation, the nodes (i.e. vertices) are genes or gene products, and the
links (i.e. edges) are cellular interactions. We create a global KEGG network to avoid the artificial boundaries between individual pathways by taking the graph union of the individual pathways, i.e. merging the pathways so that the links which are in at least one KEGG pathway will be part of the new global network. We then consider the largest connected component of the resulting network in our algorithm. This global network has \( N = 4867 \) nodes, \( L = 42874 \) edges, and a diameter \( D = 34 \). Of the \( N = 4867 \) nodes, 2709 of them are also amongst the 7680 genes assayed in all three ovarian cancer studies.

We apply our method to each of the ovarian cancer studies with the global gene network to calculate the combined evidence \( p_i^{\text{Comb}}(r) \) for each of the 2709 genes \( i \) that are assayed and on the network. A table of the full results is provided as an additional file in [80]. With the results from each of the three ovarian cancer data sets, we evaluate not only the cross-study concordance of our analysis technique, but also its ability to identify biologically relevant genes and truly integrate pathway and expression data.

### Disruptive genes found by GeneSurrounder are associated with ovarian cancer

To evaluate GeneSurrounder’s ability to identify biologically relevant genes, we compare our results in all three ovarian cancer studies (Table 3.2) to existing biological knowledge. Applying GeneSurrounder to the 2709 common genes between studies that were assayed and on the network, we generated three distinct ranked lists of genes for each study based on the computed \( p_i^{\text{GS}} \) value. To compare these results to existing biological knowledge, we consider genes that pass our Bonferroni corrected threshold (at significance level \( \alpha = 0.05 \) and with a diameter of \( D = 34 \), our Bonferroni corrected threshold is \( -\log_{10}(p) \geq 2.83 \)) in all three studies (Table 3.2).
We used the DOSE R package [87] to analyze the enrichment of these genes with Disease Ontology (DO) terms [88]. We found that the genes that pass our Bonferroni corrected threshold in at least one ovarian cancer study were significantly enriched with the DO term “ovarian cancer” (DOID:2394) ($p = 0.0000177$). Furthermore, amongst these genes are three families of protein coding genes, *CDC* (involved in the cell division cycle), *MCM*, and *ORC* (both involved in DNA replication), with biological functions that support their role in ovarian cancer.

<table>
<thead>
<tr>
<th>Gene</th>
<th>GSE14764</th>
<th>GSE17260</th>
<th>GSE9891</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>ADRB3</em></td>
<td>3.033</td>
<td>2.933</td>
<td>3.554</td>
</tr>
<tr>
<td><em>AURKA</em></td>
<td>2.865</td>
<td>3.383</td>
<td>3.716</td>
</tr>
<tr>
<td><em>CDC45</em></td>
<td>4.270</td>
<td>3.741</td>
<td>4.830</td>
</tr>
<tr>
<td><em>CDC7</em></td>
<td>4.386</td>
<td>3.769</td>
<td>4.830</td>
</tr>
<tr>
<td><em>DBF4</em></td>
<td>4.270</td>
<td>3.769</td>
<td>4.830</td>
</tr>
<tr>
<td><em>IL7</em></td>
<td>3.055</td>
<td>2.898</td>
<td>2.910</td>
</tr>
<tr>
<td><em>ITGAM</em></td>
<td>2.961</td>
<td>3.024</td>
<td>3.094</td>
</tr>
<tr>
<td><em>MCM2</em></td>
<td>4.830</td>
<td>3.372</td>
<td>4.830</td>
</tr>
<tr>
<td><em>MCM3</em></td>
<td>4.830</td>
<td>3.383</td>
<td>4.830</td>
</tr>
<tr>
<td><em>MCM4</em></td>
<td>4.830</td>
<td>3.394</td>
<td>4.830</td>
</tr>
<tr>
<td><em>MCM5</em></td>
<td>4.830</td>
<td>3.372</td>
<td>4.830</td>
</tr>
<tr>
<td><em>MCM6</em></td>
<td>4.830</td>
<td>3.428</td>
<td>4.830</td>
</tr>
<tr>
<td><em>ORC4</em></td>
<td>4.386</td>
<td>3.172</td>
<td>4.830</td>
</tr>
<tr>
<td><em>ORC6</em></td>
<td>4.386</td>
<td>3.691</td>
<td>4.830</td>
</tr>
<tr>
<td><em>TTK</em></td>
<td>2.904</td>
<td>3.089</td>
<td>4.830</td>
</tr>
</tbody>
</table>

Table 3.2. “Disruptive” disease genes in high-grade ovarian cancer consistently found by GeneSurrounder: At a threshold of $p = 0.05$ and with a diameter of $D = 34$, the Bonferroni corrected threshold is $-\log_{10}(p) \geq 2.83$. Listed are the genes that pass this threshold in all three studies.
To further compare our results to existing biological knowledge, we found evidence in the literature that $\text{CDC7}$, $\text{ORC6L}$, and $\text{DBF4}$ are associated specifically with ovarian cancer [89, 90, 91]. The inclusion of $\text{CDC45}$ suggests the possibility that it is also associated with ovarian cancer. $\text{CDC7}$ encodes for a cell division cycle protein and has been found to both predict survival and be a powerful anticancer target in ovarian cancer [89]. $\text{ORC6L}$ encodes for a origin recognition complex that is crucial for the initiation of DNA replication and has been found to highly expressed in ovarian cancer [90]. $\text{DBF4}$ encodes for a protein that activates the kinase activity of $\text{CDC7}$ and was found to be associated with ovarian cancer [91]. The finding of these genes from studies of high-vs-low grade ovarian cancers suggests the possibility that they are not only involved in ovarian cancer but, more specifically, drive high grade ovarian cancer. A table of the full results is provided as an additional file in [80].

**GeneSurrounder results represent a true integration of pathway and expression data**

The method that we have developed combines gene expression data with an independent network model. To investigate whether our results are driven solely by either the network or the expression data or represent a true integration of biological knowledge (the pathway networks) and experimental data, we consider the association between our results, the centrality, and the differential expression for each gene. If the results were driven solely by the network, the evidence a gene is a disruptive gene would correlate strongly with its centrality in the network. We therefore calculate the correlation between our results and two different measures of centrality. If the results were driven solely by the expression
data, the evidence a gene is a disruptive gene would correlate strongly with its differential expression. We therefore calculate the correlation between our results and the differential expression for each of the studies. The results are given in Table 3.3. We find that for each of the studies, the correlations are small (on the order of 0.01), confirming that GeneSurrounder is not driven solely by network features or the expression data, but rather represents a true integration of biological knowledge (the pathway networks) with experimental data.

<table>
<thead>
<tr>
<th>Network/Gene Statistic</th>
<th>GSE14764</th>
<th>GSE17260</th>
<th>GSE9891</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree Cor.</td>
<td>0.044</td>
<td>0.070</td>
<td>0.038</td>
</tr>
<tr>
<td>Betweenness Cor.</td>
<td>0.047</td>
<td>0.059</td>
<td>0.030</td>
</tr>
<tr>
<td>$p_{DE}$ Cor.</td>
<td>0.060</td>
<td>0.103</td>
<td>$-0.051$</td>
</tr>
</tbody>
</table>

Table 3.3. **Correlation between GeneSurrounder results and network/gene statistics**: The three columns are the rank correlation between GeneSurrounder results ($p^{GS}$) and network/gene statistics (Degree, Betweenness, and $p^{DE}$) across all genes in each dataset. The Degree and Betweenness are two different network centrality measures. The Degree is the number of connections a node has and the Betweenness is the fraction of shortest paths that passes through the node. $p^{DE}$ is the $p$-value obtained from a standard differential expression $t$-test.

**GeneSurrounder findings are more concordant than differential expression analysis**

The intuition underlying evaluating cross-study concordance is that methods that detect true biological signals should find them across different data sets measuring the same conditions. To investigate the cross-study concordance of our analysis technique (i.e. its consistency across different data sets measuring the same conditions), we consider each
pair of the three studies and calculate the correlation between our results. As a point of reference, we also calculate the correlation between the gene level statistics obtained using the customary $t$-test for differential expression. The results are given in Table 3.4.

As mentioned earlier, methods that do not take into account systems-level information tend to have poor agreement between studies because the individual genes contributing to disease-associated mechanisms can vary from one study to the next. Indeed, we find that the cross-study concordance of differential expression results is remarkably low (Table 3.4). By contrast our method is 3.51—8.55 times more consistent than differential expression analysis. This cross-study concordance suggests that our method reliably detects biological effects reproducibly across studies.

<table>
<thead>
<tr>
<th>Ovarian Cancer Study Pair</th>
<th>$p^{GS}$ Cor.</th>
<th>$p^{DE}$ Cor.</th>
<th>$p^{LEAN}$ Cor.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE14764 - GSE17260</td>
<td>0.342</td>
<td>0.040</td>
<td>0.056</td>
</tr>
<tr>
<td>GSE14764 - GSE9891</td>
<td>0.436</td>
<td>0.056</td>
<td>0.130</td>
</tr>
<tr>
<td>GSE17260 - GSE9891</td>
<td>0.485</td>
<td>0.138</td>
<td>0.290</td>
</tr>
</tbody>
</table>

Table 3.4. Cross study concordance of GeneSurrounder results compared to differential expression analysis and LEAN: The columns $p^{GS}$ Cor., $p^{DE}$ Cor., and $p^{LEAN}$ Cor. are the Spearman rank correlations respectively between the results obtained from GeneSurrounder, differential expression analysis, and LEAN for each study pair.

GeneSurrounder findings are more concordant than LEAN

We also compare GeneSurrounder to LEAN, a recent method that also attempts to integrate gene expression and network data to identify significant genes. In contrast to our method, LEAN considers only the immediate neighborhood (i.e. at a radius of one) and assesses the enrichment of significant genes. To compare the performance of our analysis
technique to LEAN, we compare their respective cross-study concordances. To ensure comparability between our method and LEAN, we use the same network and expression data for inputs to LEAN that we used for GeneSurrounder. Again, we consider each pair of the three studies and calculate the correlation between our results and the correlation between results of LEAN [83] (which is available as an R package on CRAN). The results are given in Table 3.4. We found that while LEAN is more consistent than the differential expression analysis, GeneSurrounder is more consistent than LEAN. That is, the list of “disruptive” genes detected by GeneSurrounder are more reproducible across studies than both differentially expressed genes and the results from LEAN.

**Application to Bladder Cancer Data with Global KEGG Network Model**

As a further demonstration of our method, we apply our algorithm to three bladder cancer gene expression data sets from the publicly available and curated collection ‘curatedBladderData’ [3] (Table S1). Bladder tumor samples in each data set are classified as either superficial (no invasion of the main muscle layer) or invasive (tumor growth into the main muscle layer), and we compare samples between these two groups. As in our application to ovarian cancer, the bladder cancer data was downloaded using the Bioconductor package ‘curatedBladderData’ without further processing. The same global KEGG network we created to analyze the ovarian cancer data was used to analyze the bladder cancer data. In order to compare results obtained from each of the bladder cancer datasets and compute cross-study concordance, we restricted each analysis (GeneSurrounder, differential expression, and LEAN) to the set of 2205 genes that were assayed in all three studies and were in the KEGG network. These were then filtered further to exclude both genes
and samples with > 25% missing data in any study. After mapping gene symbols in the three bladder data sets to KEGG identifiers and filtering out genes with missing data, 1757 genes remained in common to all three bladder cancer studies.

We apply GeneSurrounder to these 1757 genes to identify genes passing the Bonferroni corrected threshold (at a threshold of \( p = 0.05 \) and with a diameter of \( D = 34 \), our Bonferroni corrected threshold is \(-\log_{10}(p) \geq 2.83\)). A table of the full results is provided as an additional file. Several genes are identified as statistically significant in all three studies (Table S2); their functional roles in cell migration and adhesion (a mechanism required for the progression of tumors from “superficial” to “invasive”) further supports the ability of GeneSurrounder to detect biologically relevant signals.

As with the ovarian cancer data, we also evaluate our method’s correlation with network features and its cross-study concordance. We confirm that GeneSurrounder is not driven solely by network features or the bladder cancer expression data, but represents an integration of both (Table S3). We also confirm that GeneSurrounder yields more reproducible results than competing analyses (Table S4). While concordance values for all analysis methods were generally lower in the bladder cancer studies than in the ovarian cancer studies, we nevertheless find that GeneSurrounder is still more concordant than both differential expression analysis and LEAN. A more detailed description of these results, including discussion of significant genes, is provided in the Section 3.6.

3.4. Discussion

In this manuscript, we have developed and presented a new analysis technique, GeneSurrounder, that integrates a network model with expression data to identify individual
genes that can be targeted therapeutically. Our analysis technique identifies “disruptive”
genes — genes that impact pathway networks in a disease associated manner. The algo-
algorithm consists of two tests that are run independently of each other and then combined.
The first test, Sphere of Influence, calculates the evidence that a putative disease gene
is correlated with its neighbors, and the second test, Decay of Differential Expression,
calculates the evidence that the neighbors of a putative disease gene are differentially
expressed (with the magnitude of differential expression decreasing with distance).

We applied our algorithm to three gene expression data sets of high-vs-low grade
ovarian cancer and combined each of them with the same global network model that we
constructed from KEGG pathways. With the results from each of the three ovarian cancer
data sets, we evaluated our analysis technique. By applying our method to three different
data sets measuring the same conditions, we were able to show that it yields consistent (i.e.
concordant) results across studies, suggesting its ability to detect biologically meaningful
associations that are reproducible across studies. We also compare our results to existing
biological knowledge and find that our method identifies biologically relevant genes. To
show that our method truly integrates pathway and expression data, we compare the
results from our method to the results from a single gene analysis and the centrality of
the genes in the network. Our positive results along these three dimensions of our analysis
technique suggest that our method is a promising new strategy for identifying the genes
that control disease.

As discussed in the Background section of this chapter, pathway analysis techniques
such as GSEA, jActiveModules, and COSINE use interaction networks
and expression data to find groups of related disease-associated genes. GeneSurrounder,
to make experimental follow-up easier, identifies precise gene targets rather than groups of tens or hundreds of genes. Efforts to identify individual genes, as our method does, have either required prior biological knowledge (as in ENDEAVOUR [29] and GeneWanderer [30]) or have not used direct interactions on a global network (as in [31], an extension of SPIA [32], and nDGE [82]). Our analysis technique addresses these shortcomings by using the shortest direct distance on a global network and not requiring any prior biological knowledge. LEAN [83] considers interactions on a global interaction network and is closest to our method in this respect, but restricts its focus to nearest neighbors on the network and does not determine whether a putative disease gene is the source of change on the network.

3.5. Conclusions

The key innovation of GeneSurrounder is the combination of pathway network information with gene expression data to determine the degree to which a gene is a source of dysregulation on the network. GeneSurrounder employs a novel strategy by finding genes that both appear to influence nearby genes and cause dysregulation associated with the disease. Because GeneSurrounder considers every neighborhood size around a putative gene, it is able to identify disease genes that may have broad effects on the regulatory network (beyond nearest neighbors). GeneSurrounder thus provides a new avenue for identifying disease-associated genes by detecting genes that appear to be sources of change and could therefore be promising therapeutic targets.

While our method performs well in practice, there are limitations that bear consideration. We note that the the network model that we use, KEGG, is not phenotype-specific
(as are most pathway databases) and we therefore have to assume that the network does not change between conditions. Additionally, because KEGG (and other pathway databases) may not be complete, genes that are not annotated in any pathway cannot be considered in a GeneSurrounder analysis. Furthermore, as implemented our algorithm calculates geodesic distances between genes without taking into account the direction or type of interactions. However, we note that our approach as presented here could easily be modified to take in as input other kinds of networks (including context-specific computationally derived networks) and/or considering edge directionality by changing the gene-gene distance matrix that the Sphere of Influence and Decay of Differential computations use.

GeneSurrounder can be potentially generalized to other types of data. For instance, one might envision applying it to other kinds of omic data. For example, GeneSurrounder could potentially be extended to use genomic sequence data to identify epistatic interactions, evidenced by gene neighborhoods that have a high level of correlations in their genetic variants. Our method could also possibly be generalized for time-series gene expression data by either changing the gene-level statistics used by the algorithm or applying it separately to time points.

GeneSurrounder thus provides means to prioritize genes that are sources of disruption for a disease in the context of gene regulatory networks. By prioritizing genes in this way, our method provides insights into disease mechanisms and suggests diagnostic and therapeutic targets. Our method can be used to help biologists select among tens or hundreds of genes for further validation. Furthermore, it can be generalized to other kinds of networks (including context-specific networks) and omic data. This approach can
not only be used directly to prioritize promising targets, but also suggests new strategies for integrating systems level information with omic data to identify, validate, and target disease mechanisms. We have made the implementation of our method available to researchers on GitHub at [http://github.com/sahildshahl/gene-surrounder](http://github.com/sahildshahl/gene-surrounder) with the aim of furthering our understanding of statistical techniques to identify disease-associated genes.

### 3.6. Supplement

**GeneSurrounder Analysis of curatedBladderData**

In this supplement, we detail the results of applying GeneSurrounder to data from the curatedBladderData package (Table 3.5). The analysis parallels that of the ovarian cancer data described in the main manuscript. We apply our algorithm to three gene expression data sets of superficial-vs-invasive bladder cancer from the publicly available and curated collection ‘curatedBladderData’ [3] (Table S1). The three gene expression data sets and the KEGG network model have 2205 genes in common. After mapping gene symbols in the three bladder data sets to KEGG identifiers and filtering out genes with missing values in 25% or more of the samples in any study, 1757 genes remained in common to all three bladder cancer studies.

**Disruptive genes found by GeneSurrounder are associated with invasive bladder cancer**

To evaluate GeneSurrounder’s ability to identify biologically relevant genes, we compare our results in all three bladder cancer studies (Table 3.6) to existing biological knowledge.
Applying GeneSurrounder to the 1757 common genes between studies that were assayed and on the network, we generated three distinct ranked lists of genes for each study based on the computed $p_i^{GS}$ value. To compare these results to existing biological knowledge, we consider genes that pass our Bonferroni corrected threshold (at significance level $\alpha = 0.05$ and with a diameter of $D = 34$, our Bonferroni corrected threshold is $\log_{10}(p) \geq 2.83$).

<table>
<thead>
<tr>
<th>GEO Accession No.</th>
<th>N(superficial)</th>
<th>N(invasive)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE13507</td>
<td>103</td>
<td>62</td>
</tr>
<tr>
<td>GSE19915.GPL5186</td>
<td>38</td>
<td>41</td>
</tr>
<tr>
<td>GSE32894</td>
<td>213</td>
<td>93</td>
</tr>
</tbody>
</table>

Table 3.5. **Bladder cancer datasets used in this study**: Comparisons were made between superficial and invasive bladder cancer using public data. Superficial bladder cancer has not grown into the main muscle layer of the bladder and invasive bladder cancer has grown into the main muscle layer of the bladder. Sample sizes for each group in each dataset are given. (GSE19915.GPL5186 originally had 43 superficial samples and 45 invasive samples, but samples with missing data for 25% or more of the genes were filtered out.) The data are publicly accessible and available as part of the curatedBladderData package [3].

<table>
<thead>
<tr>
<th>Gene</th>
<th>$-\log_{10}p^{GS}$ (GSE13507)</th>
<th>$-\log_{10}p^{GS}$ (GSE32894)</th>
<th>$-\log_{10}p^{GS}$ (GSE19915.GPL5186)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td>2.994</td>
<td>3.864</td>
<td>2.859</td>
</tr>
<tr>
<td>ITGAM</td>
<td>2.859</td>
<td>3.452</td>
<td>3.154</td>
</tr>
<tr>
<td>VIM</td>
<td>3.314</td>
<td>3.019</td>
<td>3.944</td>
</tr>
</tbody>
</table>

Table 3.6. **“Disruptive” disease genes in bladder cancer consistently found by GeneSurrounder**: At a threshold of $p = 0.05$ and with a diameter of $D = 34$, the Bonferroni corrected threshold is $-\log_{10}(p) \geq 2.83$. Listed are the genes that pass this threshold in all three studies.
We used the DOSE R package [87] to analyze the enrichment of these genes with Disease Ontology (DO) terms [88]. We found that the 379 genes that pass our Bonferroni corrected threshold in at least one bladder cancer were significantly enriched with the DO term “bladder cancer” (DOID:11054) \( (p = 1.05 \times 10^{-7}) \), supporting the biological relevance of genes identified by GeneSurrounder. Furthermore, our method found three genes, \( C2, ITGAM \) and \( VIM \), that pass our Bonferroni corrected threshold in all three studies (Table 3.6). \( C2 \) plays a role in inflammation and removing debris from cells and tissues. \( ITGAM \) plays a role in cell adhesion molecules and transcriptional misregulation in cancer. \( VIM \) plays a role in cell attachment, migration, and signaling and microRNAs in cancer. As we are comparing samples between superficial-vs-invasive bladder cancer (superficial bladder cancer has not grown into the main muscle layer of the bladder, whereas invasive bladder has grown into the main muscle layer), the finding of these three genes from studies of superficial-vs-invasive bladder is sensible and suggests that GeneSurrounder is able to accurately identify mechanistically relevant genes. A table of the full results is provided as an additional file in [80].

**GeneSurrounder results represent a true integration of pathway and expression data**

The method that we have developed combines gene expression data with an independent network model. To investigate whether our results are driven solely by either the network or the expression data or represent a true integration of biological knowledge (the pathway networks) and experimental data, we consider the association between our results, the centrality, and the differential expression for each gene. If the results were driven solely
by the network, the evidence a gene is a disruptive gene would correlate strongly with its centrality in the network. We therefore calculate the correlation between our results and two different measures of centrality. If the results were driven solely by the expression data, the evidence a gene is a disruptive gene would correlate strongly with its differential expression. We therefore calculate the correlation between our results and the differential expression for each of the studies. The results are given in Table 3.7. We find that for each of the studies, the correlations are small (at most +0.101), confirming that GeneSurrounder is not driven solely by network features or the expression data, but rather represents a true integration of biological knowledge (the pathway networks) with experimental data.

<table>
<thead>
<tr>
<th>Network/Gene Statistic</th>
<th>GSE13507</th>
<th>GSE31684</th>
<th>GSE19915.GPL5186</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree Cor.</td>
<td>+0.070</td>
<td>+0.089</td>
<td>+0.045</td>
</tr>
<tr>
<td>Betweenness Cor.</td>
<td>+0.038</td>
<td>+0.038</td>
<td>+0.036</td>
</tr>
<tr>
<td>$p_{DE}$ Cor.</td>
<td>+0.101</td>
<td>−0.033</td>
<td>+0.096</td>
</tr>
</tbody>
</table>

Table 3.7. Correlation between GeneSurrounder results and network/gene statistics: The three columns are the rank correlation between GeneSurrounder results ($p^{GS}$) and network/gene statistics (Degree, Betweenness, and $p^{DE}$) across all genes in each dataset. The Degree and Betweenness are two different network centrality measures. The Degree is the number of connections a node has and the Betweenness is the fraction of shortest paths that passes through the node. $p^{DE}$ is the $p$-value obtained from a standard differential expression $t$-test.
GeneSurrounder findings are more concordant than differential expression analysis

The intuition underlying evaluating cross-study concordance is that methods that detect true biological signals should find them across different data sets measuring the same conditions. To investigate the cross-study concordance of our analysis technique (i.e. its consistency across different data sets measuring the same conditions), we consider each pair of the three studies and calculate the correlation between our results. As a point of reference, we also calculate the correlation between the gene level statistics obtained using the customary \( t \)-test for differential expression. The results are given in Table 3.8.

As mentioned earlier, methods that do not take into account systems-level information tend to have poor agreement between studies because the individual genes contributing to disease-associated mechanisms can vary from one study to the next. Indeed, we find that the cross-study concordance of differential expression results is remarkably low (Table 3.8). By contrast our method is more consistent than differential expression analysis. This cross–study concordance suggests that our method reliably detects biological effects reproducibly across studies.

<table>
<thead>
<tr>
<th>Bladder Cancer Study Pair</th>
<th>( p^{GS} ) Cor.</th>
<th>( p^{DE} ) Cor.</th>
<th>( p^{LEAN} ) Cor.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE13507 - GSE32894</td>
<td>+0.276</td>
<td>−0.045</td>
<td>+0.023</td>
</tr>
<tr>
<td>GSE13507 - GSE19915.GPL5186</td>
<td>+0.206</td>
<td>+0.058</td>
<td>+0.154</td>
</tr>
<tr>
<td>GSE32894 - GSE19915.GPL5186</td>
<td>+0.296</td>
<td>+0.016</td>
<td>−0.076</td>
</tr>
</tbody>
</table>

Table 3.8. Cross study concordance of GeneSurrounder results compared to differential expression analysis and LEAN: The columns \( p^{GS} \) Cor., \( p^{DE} \) Cor., and \( p^{LEAN} \) Cor. are the Spearman rank correlations respectively between the results obtained from GeneSurrounder, differential expression analysis, and LEAN for each study pair.
GeneSurrounder findings are more concordant than LEAN

We also compare GeneSurrounder to LEAN, a recent method that also attempts to integrate gene expression and network data to identify significant genes. In contrast to our method, LEAN considers only the immediate neighborhood (i.e. at a radius of one) and assesses the enrichment of significant genes. To compare the performance of our analysis technique to LEAN, we compare their respective cross-study concordances. To ensure comparability between our method and LEAN, we use the same network and expression data for inputs to LEAN that we used for GeneSurrounder. Again, we consider each pair of the three studies and calculate the correlation between our results and the correlation between results of LEAN [83] (which is available as an R package on CRAN). The results are given in Table 3.8. We found that GeneSurrounder is more consistent than LEAN. That is, the list of “disruptive” genes detected by GeneSurrounder are more reproducible across studies than both differentially expressed genes and the results from LEAN.
CHAPTER 4

Identifying transcriptional consequences of impaired GCSF signaling in hematopoiesis

4.1. Introduction

The development of blood cells from hematopoietic stem cells (HSCs) involves complex mechanisms such as external cues and intracellular signaling that are important to understand for the treatment of human diseases [92, 93]. Hematopoiesis is the process by which all blood cells (both red blood cells and the six types of white blood cells involved in immune response) develop. It is one of the best-studied examples of cellular development and differentiation. In the bone marrow, stem and progenitor cells receive signals that send them down paths to become particular red or white blood cells. The failure of one or more cell-types to develop can lead to a compromised immune system.

Granulocyte colony stimulating factor (GCSF) is a hematopoietic growth factor naturally produced in humans that stimulates the bone marrow to produce granulocytes, which are essential for suppressing infections [94, 95]. As a drug, recombinant human GCSF is administered therapeutically to patients with severe congenital neutropenia (SCN) to avoid life threatening infections. These patients, however, have often been observed to develop acute myeloid leukemia (AML) [96]. Since hematopoietic growth factors such as GCSF affect multiple phenotypes (including survival and differentiation) through complex intracellular signaling networks, we hypothesize that a mutant GCSF receptor (GCSFR)
triggers system-wide perturbations that lead to cancer. We study expression data for samples with a mutated GCSFR gene (leading to a truncated receptor) and for samples with a wild type GCSFR gene assayed at 0, 1, 2, 4 hours after treatment with GCSF. From this data, we seek to investigate the transcriptional differences in response to GCSF signaling in cells with truncated GCSF receptors.

In the case of static gene expression, the most common method for detecting gene-association is via differential expression analysis, in which each gene is independently tested for significant differences in mean expression between phenotypes \[16\]. Recent techniques to identify differentially expressed genes in time-series data include Extraction of Differential Gene Expression (EDGE) \[97, 98, 99\] and Bayesian Estimation of Temporal Regulation (BETR) \[100\]. EDGE identifies genes showing “temporal differential expression” between phenotypes by first fitting a curve to the combined groups and then each group separately and quantifying the improvement in the goodness of fit. BETR identifies genes showing “temporal differential expression” by first fitting a generative model to each gene assuming identical mean profiles between conditions and then fitting a model allowing for differential expression and then using Bayes’s rule to calculate the likelihood of the data for each model.

However, these approaches to prioritizing dysregulated genes do not take into account the role that a given gene has in signaling pathways. As a result, they can miss crucial multi-gene interactions that underlie complex phenotypes. To investigate the interactions between genes and gene products using expression data, two complementary approaches can be taken: infer a context specific wiring diagram from the data or analyze expression data in the context of a given wiring diagram.
Methods to infer context-specific interaction networks can be broadly organized as either *de novo* or semi-supervised. *De novo* techniques, such as ARACNE [11] and the lagged ordered LASSO method [12], infer gene regulatory networks solely from expression data. Semi-supervised techniques, such as iRafNet [13] and the postPLSR method [14], take advantage of known regulatory dependencies (contained in pathway databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [8]) when inferring networks to refine existing network models using expression data.

Pathway analysis techniques, such as GSEA [101], SPIA [23] and jActiveModules [102], use these interaction networks to identify groups of related genes. By incorporating systems-level information, these methods can capture multi-gene interactions and thus yield mechanistically interpretable results that are more reliable than single-gene analyses. Newer analysis techniques, such as GeneSurrounder [80] identify precise gene targets while incorporating systems-level information. GeneSurrounder [80] is an analysis technique that determines the degree to which a given gene is a source of dysregulation on a network. The method seeks to identify genes with two defining characteristics: (i) they appear to influence other genes nearby in the network, as evidenced by strongly correlated expression with nearby genes; and (ii) their dysregulation is associated with disease, as evidenced by a pattern of (progressively weaker) differential expression centered about that gene. Mechanistically relevant genes are more reliably identified by GeneSurrounder than competing approaches [83], and their influential role in the network makes them appealing targets for intervention.

Our goal is to identify the genes that are sources of dysregulation in a mutant GCSF receptor background. The key idea is to combine our GCSF treated expression data with
a hematopoiesis-specific interaction network and then find the sources of differential time-course profiles on the gene network. We first (1) seek to identify the genes that respond to a GCSF signal in healthy, wild type cells. We then (2) seek to determine how those genes are connected in GCSF responsive cells (i.e. cells that proliferate or differentiate when exposed to GCSF). Since impairments in this response lead to aberrant proliferation or differentiation, we are interested in this network. To infer a network that is independent from our data yet still representative of hematopoiesis, we apply iRafNet [13] to data from GCSF responsive cells available through Haemopedia [103]. Since iRafNet is a semi-supervised method and can integrate other sources of data, we use KEGG networks alongside the Haemopedia data to infer a network. We finally (3) seek to identify the genes on this network that are disruptive when the GCSF receptor is truncated. To identify disruptive genes, we apply GeneSurrounder [80] to our expression data and find the sources of differential time-course profiles on the network. By finding genes involved in system–wide perturbations that may be associated with malignant transformation in GCSFR mutant cells, this analysis could suggest promising new gene targets for follow-up studies.

4.2. Methods

Our analysis integrates three sources of data (our GCSF–treated expression data, Haemopedia expression data, and a KEGG network) and proceeds in three phases (Figure 4.1). Since our goal is to identify disruptive genes involved in GCSF signaling, we first identify GCSF responsive genes in healthy, wild type cells. We do so by quantifying each gene’s dependence on time and dynamic range in the GCSF–treated wild type
expression data. The genes with a significant dependence on time and dynamic range in our data are considered responsive to GCSF. To then infer a network amongst those genes, we apply iRafNet [13] to the Haemopedia [103] expression data (from GCSF responsive cells) and a KEGG network (of GCSF responsive genes). We want to know the gene regulatory network for cell types that are responsive to GCSF signaling (i.e. cells that differentiate or proliferate in response to GCSF, rather than “generic” cells represented in KEGG) so that we can determine how this network is disrupted when GCSFR is mutated. To learn the gene regulatory network for GCSF-responsive cell types, we use data from Haemopedia, a source of independent haematopoiesis gene expression data that is more comprehensive and representative of GCSF-responsive cell phenotypes than our GCSF–treated expression data. Haemopedia contains gene expression profiles from haematopoietic cells and we choose samples of GCSF responsive cells (Common Myeloid Progenitors and Myleoblasts [93]). We then construct a network for downstream analysis by updating the KEGG network with high scoring edges from iRafNet. To find the sources of dysregulation (i.e. differential time-course profiles between the wild type and truncated receptor conditions), we apply GeneSurrounder [80] to the inferred network and our GCSF treated data.

Data sources: Analysis integrates transcriptomic and network data

GCSFR time-series expression data. To investigate our hypothesis, our collaborators in the Corey lab assayed expression of 6920 genes in 32 samples at $t = 0, 1, 2, 4$ hours after treatment with GCSF in mouse cell lines. There are four d715 (truncated receptor) and four GCSFR.I (wild type) samples at each timepoint. We processed the expression
Workflow for identifying sources of dysregulation amongst GCSF responsive genes

**DATA SOURCES**

- GCSF treated expression data
- Haemopedia expression data
- KEGG network

**Phase 1:** Identify GCSF responsive genes from wild-type data by quantifying dependence on time and dynamic range.

**Phase 2:** Infer network of GCSF responsive genes by applying the semi-supervised network inference technique iRafNet.

**Phase 3:** Identify sources of dysregulation on the updated network by applying the network-based method GeneSurrounder.

---

**Figure 4.1.** Workflow for identifying sources of dysregulation amongst GCSF responsive genes

Data by filtering out genes with more than 25% missing data amongst all wild type and truncated receptor samples and imputing the remaining missing data using \((k = 10)\) nearest neighbor averaging, where \(k\) is the number of neighbors used.

**Haemopedia RNA-seq data.** Haemopedia [103] is a database of gene expression profiles from haematopoietic cells and includes data for both mice and humans. We used data from the Haemopedia-Mouse-RNAseq dataset, which contains counts for 47643 genes and 129 samples of sorted wild type mouse haematopoietic cells. We chose samples of GCSF responsive cells (Common Myeloid Progenitors and Myleoblasts [93]). We then filtered out genes with median expression values across samples less than five and normalized the
counts with DEseq2. The resulting dataset for downstream analysis consisted of normalized counts for 15030 genes in 19 samples of GCSF responsive cells.

KEGG network. The KEGG database organizes experimentally derived pathway information into individual networks of functionally related molecules. In the KEGG representation, the nodes (i.e. vertices) are genes or gene products, and the links (i.e. edges) are cellular interactions. We downloaded Mus musculus (mouse) KEGG pathway maps and merged the individual pathways by taking a graph union to create a global network that does not have artificial boundaries. This global network has \( N = 6069 \) nodes, \( L = 70400 \) edges, and a diameter \( D = 27 \).

Phase 1: Identify GCSF responsive genes

To find GCSF responsive genes, we quantify separately each gene’s dependence on time and each gene’s dynamic range after treatment with GCSF. If a gene has both a significant dependence on time and a large dynamic range, we consider it a GCSF responsive gene. To quantify the dynamic range, we compute the difference between the maximum and minimum expression values amongst the 16 samples. To quantify dependence on time, we fit a separate linear model for each gene using its (wild type) expression and test the significance of the time-dependent coefficients. We have data at four time points and seek to capture not only linear trends, but any pattern of observations across four times points. Any pattern of four observations in time can be represented as a combination of linear, quadratic, and cubic trends (Figure 4.2).

To model the expression of a gene as a combination of linear, quadratic and cubic trends, we use an orthogonal polynomial basis where \( hr.L, hr.Q, hr.C \) are vectors.
corresponding to linear, quadratic and cubic terms respectively. These vectors are constructed so that the linear vector corresponds to a straight line relationship in the data, the quadratic vector corresponds to evidence of a curvature in the data, the cubic vector corresponds to a point of inflection in the data and the vectors are orthogonal to each other. To illustrate, vectors that satisfy these properties are \( \text{hr.L} = \langle -3, -1, 1, 3 \rangle \), \( \text{hr.Q} = \langle 1, -1, -1, 1 \rangle \), and \( \text{hr.C} = \langle -1, 3, -3, 1 \rangle \).

More precisely, the linear model we use to quantify dependence on time is

\[
y = \beta_0 + \beta_1 (\text{hr.L}) + \beta_2 (\text{hr.Q}) + \beta_3 (\text{hr.C})
\]

where \( \text{hr.L}, \text{hr.Q}, \text{hr.C} \) are an orthonormal polynomial basis. When implemented in R, these terms have different numerical values than illustrated above, but satisfy the same properties \[105\]. If \( \beta_1 \) is significant, there is evidence of a linear trend in the data. If \( \beta_2 \) is significant there is evidence of a quadratic trend in the data (e.g. upregulated expression and then downregulated expression). If \( \beta_3 \) is significant, there is evidence of a cubic trend in the data. (e.g. upregulated expression, downregulated expression, and then upregulated
expression). By using this model, we can capture not only linear trends, but any pattern of observations across four time points.

The coefficients $\beta_1, \beta_2, \beta_3$ therefore correspond to expression depending on time. We test the hypothesis that these coefficients are all simultaneously zero. Under this null hypothesis, the $F$-statistic computed from the residual sum of squares of the model with and without these terms follows an $F$ distribution. We therefore quantify dependence on time by computing an $F$-statistic for $\beta_1, \beta_2, \beta_3$. This procedure yields a list of genes with a transcriptional responses to GCSF administration in wild–type pro-B cells.

**Phase 2: Infer network of GCSF responsive genes**

To understand how GCSF-responsive *genes* contribute to gene regulation in GCSF-responsive *cells* (i.e. cells that show a phenotypic response of proliferation in response to GCSF signaling), we reconstruct the regulatory network of the genes found in Phase 1 using data from GCSF-responsive cell types in the Haemopedia data. To infer the regulatory network of GCSF responsive genes, we use the semi-supervised method iRafNet \[^{13}\]. This method determines the potential regulators of a gene $g_i$ by modeling the expression of $g_i$ with a random forest model and identifying the genes that are the strongest predictors of the expression of gene $g_i$.

Random forests is a collection of decision trees, each of which is a top down model in which predictors (here the potential regulators of gene $g_i$) are recursively chosen to predict the response (here the expression of gene $g_i$). At each step of a decision tree, a sample of $N$ predictors is uniformly drawn and the predictor which minimizes the variance of the response is chosen to be added to the model. iRafNet integrates other sources of data
(the KEGG network in our analysis) with the main expression data (the Haemopedia expression data in our analysis) by weighting this sampling scheme. As result, if there is evidence in the supporting data that a given predictor regulates gene $g_i$, it is assigned a higher weight than other predictors so that it is more likely to be drawn at each stage of the decision tree. In our analysis, the random forest parameters were set to $ntree = 1000$ and $mtry = \sqrt{p} - 1$, where $ntree$ is the number of trees, $mtry$ is the number of variables chosen at each node, and $p$ is the number of genes in the data.

In iRafNet, weights for network data are elements of the diffusion matrix on the network $F_{i,j}$, which represent accumulation of “flow” diffusing along network edges from node $i$ and amassing at node $j$. $F_{i,j}$ weights $g_j$ in the sampling scheme described above when modeling the expression of $g_i$. However, if two nodes are disconnected “flow” can not diffuse between them and the corresponding weight would be zero. Since edges are established in iRafNet by determining the strongest predictors of a given gene, an edge would never be assigned between these nodes. As a result, an edge could never be assigned between those nodes.

To address this, we equally weigh all pairs of nodes and then use the diffusion matrix to bias this equal weighting scheme so that an edge can still be established between disconnected nodes. Hence, we modify iRafNet to use sampling scores given by

$$W = 1 + e^{-L}$$

where $W$ is the matrix of sampling scores and $L$ is the Laplacian matrix of the network. In this equation, the first term (1) corresponds to equally weighing all pairs of nodes and the second term ($e^{-L}$) corresponds to a diffusion matrix that biases this equal weighing.
scheme by incorporating the network data and quantifying the accumulation of “flow” diffusing along network edges between genes. Without the first term (1), the weights between two disconnected nodes would be zero since no “flow” can diffuse between them. As a result, an edge would never be assigned between them by iRafNet. By including the first term (1), an edge can still be established between disconnected nodes, thus ensuring a connected graph.

To evaluate the significance of the importance scores between pairs of genes, we then used iRafNet to compute a null distribution (using 200 permutations) of importance scores by shuffling the target gene counts. We then construct a network for downstream analysis by updating the KEGG network with significant edges.

**Phase 3: Identify sources of dysregulation on network**

Having constructed the GCSF response network in healthy cells, we next investigate sources of transcriptional dysregulation on this network in GCSFR-mutant cells. To find sources of dysregulation on the GCSF responsive genes network, we use GeneSurrounder [80]. This method integrates expression data and network information to detect genes that drive the behavior of pathways associated with disease. The GeneSurrounder method consists of two tests that are run independently of each other (Figure 3.1) and then combined, for every neighborhood size on the network.

As described in Chapter 3, the first step, dubbed “Sphere of Influence,” tests if a candidate gene $i$ is more strongly correlated with its network neighbors than would be expected by chance. The second step, dubbed “Decay of Differential Expression,” tests whether the magnitude of dysregulation of other genes $j$ in the neighborhood is inversely related
to the distance $d_{ij}$ of gene $j$ from gene $i$. In our analysis, we assessed the significance of each step with permutation testing and used 1000 re-samplings.

For the Decay of Differential Expression step, we must first compute a gene–level statistic $g_j$ that quantifies gene $j$’s dysregulation. The gene–level statistics that we use in this application quantify the degree to which the time courses of expression differ between the wild type and truncated receptor conditions. Therefore, GeneSurrounder will find sources of differential gene expression dynamics on the network. To compute these statistics, we fit a separate linear model for each gene using its expression. As in Equation 4.1, we use an orthogonal polynomial basis to capture any pattern of four observations in time. In the model below, we encode phenotype as a categorical variable.

$$y = \beta_0 + \beta_1 (hr.L) + \beta_2 (hr.Q) + \beta_3 (hr.C) + \beta_4 (I(\text{phenotype} = d715)) + \beta_5 (hr.L) I(\text{phenotype} = d715) + \beta_6 (hr.Q) I(\text{phenotype} = d715) + \beta_7 (hr.C) I(\text{phenotype} = d715)$$

(4.3)

where $hr.L$, $hr.Q$, $hr.C$ are coefficients derived from treating time as an ordered categorical variable ($0 < 1 < 2 < 4$) and $I(\text{phenotype} = d715)$ is an indicator variable that equals 1 when the sample corresponds to the d715 phenotype and 0 when it corresponds to the full length receptor phenotype. As in Equation 4.1, the coefficients are chosen so that the terms are an orthonormal polynomial basis and we can capture not only linear trends, but any pattern of observations across four time points.

The coefficients $\beta_5, \beta_6, \beta_7$ therefore correspond to expression depending on the interaction between time and phenotype. To quantify differences in the time course of expression, we therefore test for the significance of these interaction terms. In other words, we test the
hypothesis that these coefficients are all simultaneously zero. Under this null hypothesis, the $F$-statistic computed from the residual sum of squares of the model with and without the interaction terms follows an $F$ distribution. Therefore, the gene level statistics for GeneSurrounder are an $F$-statistic for $\beta_5, \beta_6, \beta_7$.

4.3. Results

GCSF responsive genes identified from GCSF treated data

To find GCSF responsive genes, we quantify dependence on time and dynamic range. We plotted these two quantities against each other for all 6920 genes in the data (Figure 4.3). We thresholded the adjusted $p$-values from the linear model (i.e. dependence on time) at 0.01 and the log$_2$ transformed difference between maximum and minimum expression (i.e. dynamic range) at 1. These thresholds filter out genes with expression that does not at least double sometime in our data (between $t = 0$ to $t = 4$ hours) and does not show significant dependence on time. Thus, we find 2178 GCSF responsive genes amongst the 6920 genes assayed in our data. The three genes with the highest combination of time-dependence and dynamic range (upper right hand corner of Figure 4.3) are $IL4$, $SERPINA3G$, and $SOCS3$. $IL4$ participates in B cell function, $SERPINA3G$ is a member of the serine protease inhibitor class, and $SOCS3$ inhibits JAK/STAT3.

Updated KEGG network of GCSF responsive genes

In our previous analysis, we found 2178 GCSF responsive genes. However, only 532 of those genes uniquely mapped to KEGG identifiers and are on the KEGG network. To infer a network specific to the context of hematopoiesis, we computed the evidence of an edge
Figure 4.3. **GCSF responsive genes identified from GCSF treated wild-type expression data** Displayed is the dependence on time vs dynamic range for 6920 genes computed from the 16 wild-type samples in the GCSF treated data. The horizontal and vertical lines correspond to a threshold of $p = 0.01$ and $\log_2(\text{Fold Change}) = 1$ respectively.

between every pair of these 532 GCSF responsive genes using iRafNet [13] (Figure 4.4). We find a natural cutoff between the top 51 most significant edges and the remaining edges. As illustrated in Figure 4.4, the $p$-values jump from $p = 0.105$ to $p = 0.725$. As a result, we threshold the $p$-values at this cutoff point and choose the top 51 edges with $p \leq 0.105$. To then construct a network for downstream analysis, we add these 51 edges and their corresponding genes to the KEGG network of GCSF responsive genes and find the largest connected component of this network. This graph has $N = 333$ nodes, $L = 671$ edges, and a diameter $D = 12$ (Figure 4.5).
Sources of dysregulation on the network

To find the sources of dysregulation on the network of GCSF responsive genes, we first computed gene-level statistics quantifying each gene’s dysregulation. We then applied GeneSurrounder to the 333 genes that are on the iRafNet updated KEGG network. We separately plotted the $p$-values from GeneSurrounder against each gene’s degree on the network and gene-level $F$-statistic (Figure 4.6). Note that the correlations are small (on the order of 0.01), confirming that the results are not driven solely by either the network features or the expression data. We generated a ranked list of genes based on the computed $p$-values (Table 4.1). Amongst those genes, we found those involved in regulating RNA transcription, chromatin, and metabolizing nucleotides. The genes on this list are between three and six hops away from the GCSFR gene (CS3FR) on the
Figure 4.5. **Largest Connected Component of Updated GCSF responsive genes KEGG network** Displayed is the network structure of the KEGG network of GCSF responsive genes after adding the 51 most significant edges returned by iRafNet and finding the largest connect component. (a) shows the degree distribution. (b) shows the distribution of shortest path lengths.

network. The most significant gene is *POLR3G* and in Figure 4.7 we display its results for each component of the GeneSurrounder algorithm. These genes as found by our analysis are possible sources of dysregulation associated with aberrant signaling due to the mutant GCSF receptor and candidates for follow-up studies.

### 4.4. Discussion

In this manuscript, we describe mouse cell line time-course data for samples with a wild type vs truncated GCSFR gene assayed at 0, 1, 2, 4 hours after treatment with GCSF. We also present an analysis of that data to identify the genes that are dysregulated in signaling networks due to a mutant granulocyte colony stimulating factor (GCSF) receptor. Recombinant human GCSF is administered therapeutically to patients with severe congenital neutropenia (SCN) to avoid life threatening infections, but they have
Figure 4.6. \textit{p-values from GeneSurrounder} Displayed are the \( p \)-values from GeneSurrounder. At a threshold of \( p = 0.05 \) and with a diameter of \( D = 12 \), the Bonferroni corrected threshold is \( \log_{10}(p) \geq 2.38 \).

<table>
<thead>
<tr>
<th>Gene</th>
<th>(-\log_{10}p^{GS})</th>
<th>Distance from ( Csf3r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( POLR3G )</td>
<td>3.99</td>
<td>4</td>
</tr>
<tr>
<td>( SHMT1 )</td>
<td>3.01</td>
<td>4</td>
</tr>
<tr>
<td>( POLR1A )</td>
<td>2.80</td>
<td>4</td>
</tr>
<tr>
<td>( SIRT1 )</td>
<td>2.73</td>
<td>3</td>
</tr>
<tr>
<td>( GART )</td>
<td>2.65</td>
<td>5</td>
</tr>
<tr>
<td>( MTHFD1 )</td>
<td>2.64</td>
<td>5</td>
</tr>
<tr>
<td>( MTAP )</td>
<td>2.53</td>
<td>6</td>
</tr>
<tr>
<td>( MYD88 )</td>
<td>2.41</td>
<td>2</td>
</tr>
<tr>
<td>( CAD )</td>
<td>2.40</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 4.1. \textit{Sources of dysregulation found by GeneSurrounder}. At a threshold of \( p = 0.05 \) and with a diameter of \( D = 12 \), the Bonferroni corrected threshold is \( \log_{10}(p) \geq 2.38 \). Listed are the genes that pass this threshold.

Often been observed to develop acute myeloid leukemia (AML) \cite{96}. We hypothesize that a mutant GCSF receptor (GCSFR) triggers system-wide perturbations that lead to cancer and aim to identify promising new gene targets from our mouse cell line time-course
Figure 4.7. **Displayed are the results for the gene Polr3g** (a) shows $-\log_{10}(p_{\text{Sphere}})$ vs the Neighborhood Radius. (b) shows $-\log_{10}(p_{\text{Decay}})$ vs the Neighborhood Radius. (c) shows $-\log_{10}(p_{\text{Combined}})$ vs the Neighborhood Radius. (d) shows the Number of Assayed Genes vs the Neighborhood Radius. In the top three plots, the dashed and dotted lines correspond to a significance level of 0.05 and 0.01 respectively. In the bottom plot, the solid line corresponds to the total number of genes assayed and on the network.

data. Though we may not be able to fix the GCSF receptor in patients directly with the mutation, it may be possible to treat the downstream consequences by identifying key regulators of the aberrant response.
We performed a three-phase analysis to combine our GCSF treated expression data with a hematopoiesis-specific interaction network and find the sources of differential time-course profiles on the gene network. In the first phase, we identify GCSF responsive genes. In the second phase, we infer a context specific network by applying iRafNet \[^13\] to data from GCSF responsive cells available through Haemopedia \[^103\]. We found 2178 GCSF responsive genes amongst the 6920 genes assayed in our data and then found 51 edges to add to the KEGG network of GCSF responsive genes. The largest connected component of this network had 333 genes.

In the third phase, we apply GeneSurrounder \[^80\] to the inferred network and our GCSF treated data to find the sources of differential time-course profiles between the wild type and truncated receptor conditions. After applying GeneSurrounder to the data, we found sources of dysregulation on the network that are involved in regulating RNA transcription, chromatin, and metabolizing nucleotides. We also confirmed that the results are not driven solely by either the network features or the expression data.

One limitation of our analysis that bears consideration is the use of KEGG pathways and iRafNet to construct a context specific network. The results of our analysis depend on the pathway maps contained in KEGG and iRafNet to identify significant edges. This analysis could be potentially modified to use other pathway databases or network-inference techniques. As presented, the GCSFR time-course expression data and three-phase analysis to identify dysregulated genes provides promising new gene targets for follow-up studies and could help to identify prognostic markers and therapeutic targets for patients with SCN at risk of developing AML.
CHAPTER 5

Conclusions

In this thesis, we studied the network-based analysis of transcriptomic data. The analytical techniques we considered and developed are “network-based” in the sense that they integrate the network data of interactions between genes and gene products with transcriptomic data to provide insights about the functions and interactions of cellular processes. We first outlined and applied an evaluation framework for pathway analysis methods, then presented a novel procedure to detect genes that appear to influence genes with disease–associated dysregulation, and finally studied the development of acute myeloid leukemia (AML) from severe congenital neutropenia (SCN) by analyzing time–series expression data (Figure 1.1).

In Chapter 2, we reviewed eight network-based pathway analysis techniques by comparing their methodological approaches, ease of use, and evaluating the consistency of their results across a set of comparable studies. We found that two major methodological choices distinguish the algorithms: gene $p$-value thresholding and the type of hypothesis (“self-contained” or “competitive”) being tested by the method. Since a gene $p$-value threshold introduces an arbitrary parameter in the analysis technique, we generally prefer threshold-free techniques and since the “self–contained” null is better justified biologically, we prefer techniques that test the “self–contained” null. When comparing the consistency of the methods, we, as expected, found greater concordance for the eight pathway analysis analysis techniques than gene-level analyses and the and then non-network GSEA [21].
method. Amongst the eight network-based algorithms, we found that CePa–GSA [25, 24] exhibits the best cross-study concordance and has the benefit of testing the preferred “self-contained” null. The pathway analysis methods we reviewed facilitate the understanding of the complex networks that coordinate cellular processes.

In Chapter 3, we presented a new analysis technique, GeneSurrounder, that integrates expression data and network information in a novel procedure to detect genes that are sources of dysregulation on the network. The key idea of our method is to score genes based on the evidence that they influence the dysregulation of their neighbors on the network in a manner that impacts cell function. The algorithm consists of two tests that are run independently of each other and then combined. The first test, Sphere of Influence, calculates the evidence that a putative disease gene is correlated with its neighbors, and the second test, Decay of Differential Expression, calculates the evidence that the neighbors of a putative disease gene are differentially expressed. Applying GeneSurrounder to real ovarian cancer and bladder cancer expression data, we show that our method is able to identify biologically relevant genes, integrate pathway and expression data, and yield more reproducible results across multiple studies of the same phenotype than competing methods. Together these findings suggest that GeneSurrounder provides a new avenue for identifying the genes that are sources of dysregulation for cellular processes.

In Chapter 4, we analyzed time–series expression data to study the development of acute myeloid leukemia (AML) from severe congenital neutropenia (SCN). We hypothesized that a mutant GCSF receptor (GCSFR) triggers system wide perturbations that lead to the clonal evolution to cancer and sought to identify the genes that are dysregulated in the signaling networks. We presented original mouse cell line expression data of
samples with both a truncated and wild type receptor at four different time points. We used this data to find genes with significant differences in the time course of expression. We combined our data with an interaction network specific to the context of hematopoiesis. To infer a network that is independent from our data, we applied the semi-supervised method iRafNet to data from GCSF responsive cells available through Haemopedia. We then found the genes that are the sources of differential time-course profiles on the network using GeneSurrounder. The genes we found from the original mouse cell line expression data provide insight about the cellular processes that could lead to the development of AML from SCN.

The next generation of analysis techniques can build on our methods by addressing some of the challenges of analyzing biological networks and transcriptomic data. First, the central dogma of biology is a simplified model of cellular biology and mRNA molecules can in fact be spliced after transcription, resulting in protein “isoforms” with possibly different functions. Future work could be done to apply and evaluate our methods to available isoform-level expression data. Furthermore, biological networks contain different types of interactions, such as gene regulation and signaling. Another future direction of our work could be to add information about edge type and directionality to our biological networks and use that in our analysis. Finally, there does not exist a gold standard against which we can benchmark our methods. We presented one approach for evaluating analysis techniques in this thesis, but future work could also be done on new validation methods, such as generating simulated data. These directions would not only extend our work, but advance the field.
With the growing volume of rapid, highly parallel measurements of biological molecules, statistical techniques have become essential for teasing out knowledge and ultimately developing therapeutics for human diseases. The insights provided by our study of the network-based analysis of transcriptomic data can be applied to generate new hypotheses and design experiments. In summary, this work advances the field by establishing a rigorous means to evaluate new algorithms for systems biology, presents a novel approach to integrate experimental data and network models to identify driver genes, and applies that method in combination with a new extension of iRafNet to elucidate GCSFR signaling (and as a proof of concept) (Figure 1.1).
References


[8] Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda,


[29] Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem


