Linear Methods for Joint Analysis of Multivariate Genomics Data

M. Henry Linder
University of Connecticut - Storrs, mhlinder@gmail.com

Follow this and additional works at: https://opencommons.uconn.edu/dissertations

Recommended Citation
https://opencommons.uconn.edu/dissertations/2272
Linear Methods for Joint Analysis of Multivariate Genomics Data

Matthew Henry Linder, Ph.D.
University of Connecticut, 2019

ABSTRACT

The rise of Big Data has enabled sophisticated analysis of the human genome in unprecedented detail. Large datasets are now collected as a matter of routine, and their scope spans multiple data types and multiple functional units at the molecular level of the cell. The breadth and depth of these data offer the opportunity for complex experiments and extensive structural modeling. But, given the intricacies of these data and the nuanced challenges they pose, robust and rigorous methods are essential to ensure the value and validity of the resulting scientific research. In this dissertation, we consider statistical methods for networks, applied to signaling pathways in the human genome. We construct joint, integrative models that employ a variety of data types simultaneously. These pathway models provide a unified approach to analysis of genetic, epigenetic, transcriptomic, and other types of genomic data, and incorporate functionally meaningful biological relationships. In particular, we propose a new pathway model that integrates non-coding micro RNAs, proteins that play a regulatory role with respect to genes. We also propose methods to address obstacles that arise in the course of real-world research. We consider missing data, a fundamental reality of -omics Big Data due to variability in data quality and experimental design. We adapt a low-rank method for matrix completion to apply
to bioinformatic datasets with arbitrary patterns of missing data. We apply the imputation and pathway methods to a large-scale research study that profiles more than 30 cancer types. We also propose an algorithm to identify important subnetworks within large signaling pathways, in order to hone our understanding of the drivers of complex diseases. Through the use of interactive data visualization and analysis, we promote access to -omics analyses. Taken together, these methods provide a suite of tools that empower biological research using -omics data. Our methods span functional genomic models, address real-world problems in data analysis, and seek to make analysis of complex datasets more tractable, all while maintaining a statistically sound foundation.
Linear Methods for Joint Analysis of Multivariate Genomics Data

Matthew Henry Linder, B.A., M.S.

A Dissertation
Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy at the University of Connecticut

2019
Copyright by

Matthew Henry Linder

2019
Doctor of Philosophy Dissertation

Linear Methods for Joint Analysis of Multivariate Genomics Data

Presented by
Matthew Henry Linder, B.A., M.S.

Major Advisor
Yuping Zhang

Associate Advisor
Ming-Hui Chen

Associate Advisor
Zhiyi Chi

Associate Advisor
Lynn Kuo

University of Connecticut
2019
Dedication

To my mother Deborah, who taught me to love learning.

To my father Brett, who showed me how to think, inside and outside of the box.

To my brother Andrew, my role model for adventure and exploration.

To my partner Jaya, my endless companion in curiosity, fascination, and delight.
Acknowledgements

I thank the Department of Statistics for the financial and intellectual support necessary to earn this Ph.D. I would also like to thank Utility Operations & Energy Management at the University’s Facilities Operations for ample opportunities to apply the statistics I learned to real-world problems.

Thank you to Jaya Wen for her never-ending emotional, intellectual, and logistical support. Without her, I might not have begun a Ph.D., and certainly would not have finished it.

Thank you to Nitis Mukhopadhyay for the passion with which he taught me mathematical statistics; his humor; and the joy he brings to the classroom and those he teaches.

Thank you to Haim Bar for instruction in applied statistics, as well as consulting—in particular, for introducing me to R Shiny, and cheering me on along the way.

Thank you to Jun Yan for nurturing my interest in computing, and permitting me the freedom to learn on the job.

Thank you to Nalini Ravishanker for her generosity with her time and willingness to collaborate with me. In particular, I am grateful for her masterful intuition for organization and logistics, and her encouragement that I dream big.

Thank you to Ming-Hui Chen for mentorship and friendship. So much of my experience of graduate school was enriched and enlivened by his instruction, influence, and inspiration. I cannot imagine this journey without him as a guide, role model, and companion.

Finally, thank you to my advisor Yuping Zhang, a source of enthusiasm, ambition, and
support. Without her instruction and supervision, I could never have accomplished this thesis.
# Contents

**Dedication** iii

**Acknowledgements** iv

1 **Introduction** 1
   1.1 Overview .................................................. 1
   1.1.1 Literature ............................................... 4
   1.2 Overview of Datasets ...................................... 5
      1.2.1 The Cancer Genome Atlas ............................. 5
      1.2.2 Pathway Interaction Database ......................... 10
   1.3 Dissertation Outline ..................................... 10

2 **Network models for integrative pathway analysis** 13
   2.1 Pathway analysis for -omics observations of gene activity ........................................ 15
   2.2 miRNA-gene targeting ..................................... 19
      2.2.1 miRNA-gene target dataset ........................... 22
      2.2.2 Data processing ....................................... 23
      2.2.3 Method ................................................ 24
      2.2.4 Results ............................................... 28
   2.3 Discussion ................................................. 47
# 3 Low-rank imputation of -omics missing data

3.0.1 Structured matrix completion ......................................... 52

3.1 Integrated iterative imputation ........................................... 54

3.2 Results ............................................................................. 58

3.2.1 Simulations .................................................................... 58

3.2.2 EMC-NetGSA ................................................................. 64

3.2.3 miRNA-EMC-NetGSA analysis of breast tumor subtypes ......... 81

3.3 Discussion ....................................................................... 88

# 4 Interactive data visualization and analysis

4.1 Interactive visualization for pre-defined pathways ....................... 92

4.2 Interactive pathway analysis for subpathway identification .......... 97

4.2.1 Greedy algorithm ............................................................. 99

4.2.2 Algorithm performance ...................................................... 102

4.2.3 Analysis: results and visualization .................................... 105

4.2.4 Interactive analysis .......................................................... 111

4.3 Discussion ....................................................................... 113

# 5 Concluding Remarks and Future Research

5.1 Concluding Remarks ............................................................. 116

5.2 Future Research ................................................................ 118

Bibliography ........................................................................ 125
## List of Tables

1. Sample sizes for cancer and control populations in data for 32 cancers from The Cancer Genome Atlas (TCGA). The cancer “code” is an abbreviation for each cancer. All samples are from tissue in the afflicted, cancerous region. For the cancer population, tissue samples are from tumorous tissue. For the normal population, tissue is from healthy normal tissue. Cancers marked with a star (⋆) were included in the analysis of pathway disturbance, on the basis of each sample population including more than 10 samples. .......................... 8

2. In-degree by BRAF pathway gene, for edges into the gene from miRNAs. Values count the number of edges into a given gene from any miRNA. .................. 41

3. TCGA cancer types used for BRAF signaling pathway data analysis. Given are the TCGA cancer code abbreviation, cancer name, and sample population sizes for tumor and normal tissue samples. Only TCGA cancers with gene expression observations for all 10 BRAF pathway genes, as well as more than 10 samples in each population, were considered. ................................. 42

4. Breast tumor subtype definitions and prevalence in data from The Cancer Genome Atlas (TCGA). IHC status refers to immunohistochemical status for the three specified genes. Prevalence measures a proportion. ............... 67
Pairs of cancer and pathway for which the residual from a regression of the logit($p$)-value was negative, and larger in magnitude than a Bonferroni-adjusted critical value. In sarcoma, we observe 8 disturbed pathways; in thymoma, we observe 2 disturbed pathways; and we observe 1 pathway disturbance in each of kidney renal clear cell carcinoma, ovarian serous cystadenocarcinoma, and pancreatic adenocarcinoma.

Breast tumor molecular subtype sample sizes and prevalence in the TCGA dataset. Also given are the overall sample size of all tumors and healthy breast tissue.

Subpathways which were consistent between the expression-only and integrated analyses, and contained three or more genes. Cancers are labeled according to the codes in Table 1 (BLCA is bladder urothelial carcinoma; BRCA is breast invasive carcinoma; LUAD is lung adenocarcinoma; and THCA is thyroid carcinoma). $p$ gives the total number of genes in the pathway, while $q$ gives the number of genes in the subpathway.

Subpathways that recurred exactly in multiple pathways for the same cancer, under the EMC-NetGSA analysis. Cancer gives the cancer code from Table 1 (KIRP is kidney renal papillary cell carcinoma; LUAD is lung adenocarcinoma; THCA is thyroid carcinoma). $q$ is the number of genes in the subpathway, and $n$ is the number of pathways in which the subpathway recurred.
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sample sizes for TCGA populations with more than 10 samples.</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>A toy example of the integrated graph $G_{\text{miEMC}}$. White vertices (labeled “G”) are gene expression nodes, which specify a signaling pathway topology. The three miRNA nodes (red, labeled “mi”) target the genes, and methylation and copy number nodes integrate the multi-platform, gene-level data (black and grey, respectively).</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>A simulated five-level binary tree signaling pathway. Shown is the treatment network, which consists of: a full five-level binary tree of genes (large grey nodes); miRNA targeting 1, 2, and 3 genes (colored red, yellow, and green, respectively); and methylation and copy number nodes (white and black, respectively). Gene “G01” corresponds to the root node, which is the first level of the binary tree; “G02” and “G03” the second level; and so on. All integrated edges, that is, those from miRNA, methylation, or copy number nodes, are directed into the relevant gene node.</td>
<td>30</td>
</tr>
</tbody>
</table>
Boxplots of $-\log_{10} p$-values from simulation study. The left-hand panel shows significance tests for four gene sets of interest, under the first mean scenario in which no -omics features exhibit differential activation. The right-hand panel shows the second mean scenario, in which the top two-thirds of the simulated binary tree signaling pathway is differentially activated, but the final third is equal in the control and treatment populations. miEMC-NetGSA is shown in green, miE-NetGSA in yellow, EMC-NetGSA in blue, and NetGSA (“E”) on expression only in white.

Network diagrams for the BRAF signaling pathway. The left-hand network is the directed graph representing the BRAF genetic signaling pathway, consisting of 10 genes with 12 directed edges between. The right-hand network contains the same 10 gene vertices, but edges represent a shared miRNA parent. In other words, two genes that share an edge are both targets of a single miRNA, and possibly several. The graph contains 25 such edges. miRNA-gene targets are chosen as those entries in the mirDIP database with an “very high” confidence score. Neither the BRAF nor MAP2K2 genes are targeted by any miRNAs.
miRNA-gene interaction scores for all miRNA in TCGA dataset that target any of the 10 genes that comprise the BRAF pathway. Scores are grouped by gene, and colored by confidence class: very high, high, medium, and low confidence are colored blue, green, yellow, and red, respectively. For our data analysis, we included only miRNA-gene target interactions with “very high” confidence that the interaction exists, i.e., the blue scores.

Topology of subnetwork of the integrated BRAF signaling pathway corresponding to the unweighted adjacency matrix $A_{mi}^*$ in Equation 2.9, namely, the graph formed by miRNA-gene targets. All edges are directed from the miRNA vertex to the gene vertex. The miRNA node colors, edge widths, and edge colors correspond to the degree of the miRNA vertex, that is, the number of genes in the BRAF pathway targeted by a given miRNA. Grey graph nodes are genes, while colored graph nodes are miRNAs, targeting 1 gene (red), 2 genes (yellow), 3 genes (green), 4 genes (blue), and 5 genes (purple). The genes BRAF and MAP2K2 are not targeted by any genes. Darker edge colors correspond to higher out-degree of the associated miRNA.
Results of significance tests for pathway disturbance in the BRAF pathway, across 17 TCGA cancers. Barplots give values of $-\log_{10} p$-values, after Benjamini-Hochberg adjustment for multiple comparisons. Full gene expression miRNA expression, and gene-level methylation and copy number is shown in green ("miEMC"); integration of gene and miRNA expression is in yellow ("miE"); integration of gene expression and gene-level methylation and copy number is in blue ("EMC"); and the original, expression-only NetGSA results are in white ("E"). The horizontal dotted line gives the significance threshold at the $\alpha = 0.05$ level.

Test statistics for pathway disturbance in 17 TCGA cancers, for the BRAF pathway. Barplot heights give the test statistic value. miEMC-NetGSA is in green, miE-NetGSA in yellow, EMC-NetGSA in blue, and NetGSA ("E") in white.

Iterative integrated imputation with a toy example. $X$ is a data matrix of 5 observation vectors with missing values that do not form a rectangle. Iteratively for each column with missing elements $x_i$ ($i = 3, 4, 5$), we apply structured matrix completion to the matrix $\chi_i$, consisting of all complete data vectors; $x_i$; and any columns with missing values that are a subset of those missing from $x_i$. Imputation is performed on the entire rectangular submatrix of all features missing in $x_i$, with any observed elements not from $x_i$ treated as missing. So, imputing $x_3$ and $x_4$ uses $x_5$, whereas $x_5$ is imputed using only the complete data vectors.
Boxplots of $-\log_{10}(p)$-values for simulation study tests of pathway disturbance. Network structure is a binary tree. $X$ (shown in blue) represents EMC-NetGSA applied to the complete data matrix, with no missing values. $\tilde{X}$ is the matrix with all samples with any missing values dropped. $I3$ (in green) refers to our integrated, iteratively imputed matrix. SMC uses imputed values obtained by treating the entire minimal covering submatrix as missing, and KNN is $K = 10$-nearest neighbors imputation. Horizontal line gives $-\log_{10}(0.05)$. Scenario 1 (left): No features differential. Scenario 2 (right): Top 2/3 features differential.

Simulation results for missing data in a rectangular submatrix, using the second mean scenario in which the top 2/3 of binary tree nodes are differentially expressed. In this special case, our iterative imputation method ($I3$, in green) has identical power to that of the basic SMC method.

$-\log_{10}(p)$-values for integrative tests of network disturbance in genetic signaling pathways. Shown are pathways that changed significance after imputation in at least one sample population. Separate differential analyses were performed for each of four breast cancer subtypes, as well as a composite population of all tumors.

$-\log_{10}(p)$-values for significance tests of integrated genomics data for the entire Smad2 / Smad3 signaling pathway and selected genes, by subtype. Gene-level significance tests considered integrated triplets of pathway nodes, namely, expression, copy number, and methylation observations for each gene.
Coefficient estimates for expression in the Smad2 / Smad3 signaling pathway, by subtype. Selected genes shown only. Coefficients were estimated separately by population (cancerous tumor and healthy control). Within the tumor population, separate estimates were obtained using the raw data with missing data removed, and using the imputed data. ......................................................... 71

Expression coefficient estimates for a selected subset of Smad2 / Smad3 signaling pathway genes, by subtype. Coefficients were estimated separately by population (cancerous tumor and healthy control). Within the tumor population, separate estimates were obtained using the raw data with missing data removed, and using the imputed data. ......................................................... 73

Shown is the $- \log_{10}(p)$ transformation of the BH-adjusted $p$-values obtained for 22 TCGA cancer types among 173 NCI PID pathways. Darker colors indicate stronger significance, in terms of a smaller $p$-value. A white value indicates the pathway was not tested for that cancer. We observe strong patterns of increased significance across all pathways within individual cancers, such as elevated significance in sarcoma (SARC) and thymoma (THYM) and reduced significance in colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ). We also observe patterns in significance within pathways, across cancers, such as elevated differential activity in B-cell receptor (BCR) signaling, and a lack of disturbance in the glypican 2 and 3 networks. ......................................................... 76
\[ -\log_{10}(p) \]-values for the EPHA forward signaling pathway. Axis labels indicate the integration scheme: miEMC is miRNA-EMC-NetGSA; miE is miRNA and expression only; EMC is EMC-NetGSA; and E is NetGSA on gene expression only. 

Network-adjusted expression parameters for EPHA4 in healthy tissue and tumor subtypes. Integration of miRNA data only substantially alters the parameter estimates for all subtypes, in terms of both magnitude and sign. But, joint integration of miRNA, methylation, and copy number data clarifies and highlights the difference between the ER- HER2 over-expressed subtype and the ER+ luminal subtypes, as well as the basal subtype. 

\[ -\log_{10}(p) \]-values for the BCR pathway. Integration of miRNA data alone captures only marginal significance in the basal subtype. But, integration of miRNA in conjunction with gene methylation and copy number boosts the signal in basal subtypes, as well as the others, compared with integration of methylation and copy number without miRNAs. 

HIF-2α transcription factor network.
− log₁₀(\(p\))-values (top) and signed test statistics (bottom) for the signaling events mediated by focal adhesion kinase pathway, in sarcoma cancer tumors. These plots were generated procedurally using an interactive data visualization application, publicly accessible online. Shown are the results of hypothesis tests for the full pathway, as well as the -omics features for each individual gene, tested as separate subpathways. The pathway analysis was performed on a fully-integrated dataset consisting of expression, methylation, and copy number, using the EMC-NetGSA model; expression and methylation (EM-NetGSA); expression and copy number (EC-NetGSA); and expression alone (NetGSA).


Pathway for “signaling events mediated by PRL.” The subpathway genes are shown in yellow (seed node) and orange. This subpathway of three genes was obtained from bladder urothelial carcinoma (BLCA), and was identified consistently in the NetGSA and EMC-NetGSA analyses.
Results of greedy algorithm for subpathway discovery, using NetGSA with expression values only (top), and integrated EMC-NetGSA (bottom). The subpathway search was initialized using the single most significantly-disturbed gene in the pathway. The subpathway was expanded by including adjacent genes that increased the subpathway’s significance. Counts of gene occurrence in subpathways for any signaling pathway, organized by cancer. Only genes that recurred in some cancer across four or more pathways are shown. Genes identified in this manner as “drivers” were broadly consistent in the two analyses. This suggests these driver genes contribute to subpathway activity in a robust manner, and are not merely spuriously significant genes. Integration of methylation and copy number resulted in a larger number of driver genes within each cancer. Moreover, these significant driver genes do not recur in the same patterns across multiple cancers, which indicates differential functions for these genes across different cancers.
Chapter 1

Introduction

1.1 Overview

Systematic -omics research studies offer opportunities for comprehensive modeling of the human genome at the molecular level. These studies produce rich datasets measured across hundreds of patients, providing not only biological insight into the healthy interoperation and function of the cell, but also the opportunity to compare and contrast healthy and diseased cells. Equipped with an improved understanding of the mechanisms that give genesis to complex diseases, researchers will be more targeted in the development of treatments based on specific molecular characteristics of disease, and clinicians will be better able to apply individualized treatments.

Integrated datasets observed on -omics data types beyond gene expression allow for an unprecedented level of detail. The diversity of data available for study improves the detail with which to characterize functional processes that underlie the genome. Multi-view datasets are now routinely collected in multiple modalities across separate biological structures, and large-scale research studies are coordinated to increase the quality and quantity of data available to advance knowledge, treatment, and prevention.

Especially because of the complexity of data collection, processing, and multi-step analyses,
it is critical to apply methods that are scientifically valid and statistically rigorous. Different data types can provide new insight into unique facets of genomic systems, and deepen understanding of complicated biological functions. However, development of flexible statistical methods for these trends in research and next-generation data platforms has not kept pace with the speed of data collection. Moreover, the viability and validity of a statistical method does not equate with accessibility and interpretability. Expressive tools for functional analysis are critical to advance scientific understanding of the systems processes of the cell, and to develop and utilize novel treatments for complex diseases with distinctive genomic characteristics. Robust methods are essential to ensure scientific rigor and validity, and new experimental techniques require novel analytic tools grounded in solid statistical theory.

This thesis contributes to active research in integrative genomics and the study of complex disease by focusing on developing applied methods for biological analysis. Broadly, we consider pathway analysis of functional gene networks, and implement several analyses on multiple cancer types. In this introductory chapter, we provide an overview of literature related to integration of multi-platform data collected at the level of the gene, in particular integration of gene expression, methylation, and DNA copy number. We also review methods for pathway analysis. We then introduce the primary datasets used in our analyses, specifically a large-scale -omics dataset, and a source for known functional gene networks.

In Chapter 2, we discuss a model framework for pathway analysis of gene expression. We also discuss an integrative extension of that model that incorporates additional -omics data types collected at the level of genes. From these gene-level analyses, we progress to a novel pathway model by further integrating non-gene functional units, embedded within the same
pathway framework. Finally, we apply the fully-integrative analysis to an example signaling pathway across multiple cancer types.

In Chapter 3, we consider missing data that often afflicts even the most comprehensive -omics studies. A critical aspect of the analysis of genetic systems processes in a variety of different cancer types is the availability of complete data for analysis. To remedy missing data in -omics datasets, we consider an existing imputation procedure based on an assumption that large-scale -omics data matrices exhibit low rank. We highlight the shortcomings of the simple assumptions of that method. We propose a more comprehensive approach to accomplish imputation of missing data in arbitrary configurations. We then perform pathway analysis of breast tumor subtypes using the imputed dataset.

In Chapter 4, we discuss the use of interactive data visualization for exploration of the output of pathway analysis for -omics Big Data. We provide an overview of a basic web application that we use for exploratory data visualization of pathway outputs from the integrative models discussed in Chapter 2 and implemented using imputed data from Chapter 3. We discuss four implementations of the web application, for displaying pathway models with gene-level and miRNA-gene integration, applied to multi-cancer datasets as well as breast tumor subtypes. After reviewing the use of web applications for interactive data visualization, we discuss a computational approach to identify subpathways of interest within larger signaling pathway networks. We apply the approach to several datasets, and discuss the use of interactive tools for interactive analysis.

Finally, Chapter 5 provides commentary on and discussion of the preceding chapters. We also discuss avenues for further work in integrative pathway analysis.
1.1.1 Literature

To match the complexity and novelty of -omics datasets, statistical methods have been developed to model complex biological processes and to integrate multiple data types. The variety of approaches available for analysis of -omics data is rich, and may address different dimensions of genomic Big Data. Recent work has gone beyond simpler methods for statistical analysis, such as gene-set enrichment analysis (GSEA; Subramanian et al., 2005), to emphasize application of system-level models of biological processes. With respect to -omics datasets, basic approaches often focus on a single data type. Gene expression and clinical covariates have been applied to assess cancer survival outcomes across many different cancer types [Chandrashekar et al., 2017], which leveraged the large sample sizes that are be available.

Systematic statistical models provide a solid theoretical foundation for integrative analysis. Much work has been done to identify common signals across genomic features and -omics data types. The iCluster model of Shen et al. [2009] applies PCA-style dimension reduction for cluster analysis. PARADIGM [Vaske et al., 2010] uses network information to integrate -omics data via a latent factor model. DIG [Zhang et al., 2017b] is a statistical framework for estimating networks integrating multiple -omics data types and biological conditions. In their clinical study, Danielsen et al. [2015] integrated the results of multiple separate analyses by data type in an ad-hoc, manual fashion. This approach is typical of the literature.

Integrative analyses offer unified approach for joint analysis of a large number of genomic features across multiple data platforms. Linear methods have been applied to gene-level measurements of expression, methylation, and copy number within individual cancers [Network et al., 2017], emphasizing the aggregate information obtained from distinct data platforms. In
addition to using integration to increase knowledge of molecular processes, integrative analysis has also been applied with a focus on clinical survival outcomes across multiple cancers [Liu et al., 2018].

Pathway analysis offers a more systematic, structural approach to modeling genomic processes [Yan et al., 2017]. Pathway analysis incorporates knowledge of biological network structure into statistical models of expression. Pathway models such as the NetGSA model [Shojaie and Michailidis, 2009] (discussed in detail below) exploit results from graph theory to account for co-expression due to pass-through effects from features with differential activity to those that are not.

1.2 Overview of Datasets

1.2.1 The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) [Tomczak et al., 2015] is a long-running international study, funded by the National Cancer Institute (NCI), to collect observations of cancerous tissue in more than 30 cancers and across multiple distinct -omics data platforms. We downloaded the TCGA data from the NCI Genomic Data Commons (GDC; Grossman et al., 2016), using the TCGA-Assembler software, v2.0.0 [Zhu et al., 2014, Wei et al., 2017].

We obtained level-3 TCGA data for 33 cancers. The data were collected from individual subjects, and we obtained subject-level covariates that included sex, survival outcomes, and other phenotypic information. Within a single subject, multiple tissue samples may be obtained. For a given cancer, the tissue was collected from the anatomical region that corresponds with the cancer type. For example, breast cancer data was collected from breast tissue.
Tissue samples were classified at the time of collection as belonging to one of more than one dozen tissue types. The population sample sizes are given in Table 1. The samples sizes for populations with at least 10 samples are shown in Figure 1. The data exhibit a systematic imbalance between the two populations, with the number of tumor sample consistently higher than the number of healthy samples.

<table>
<thead>
<tr>
<th>Cancer</th>
<th>Code</th>
<th>Cancer</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adrenocortical carcinoma</td>
<td>ACC</td>
<td>92</td>
<td>5</td>
</tr>
<tr>
<td>Bladder urothelial carcinoma*</td>
<td>BLCA</td>
<td>412</td>
<td>36</td>
</tr>
<tr>
<td>Breast invasive carcinoma*</td>
<td>BRCA</td>
<td>1096</td>
<td>161</td>
</tr>
<tr>
<td>Cervical squamous cell carcinoma and endocervical adenocarcinoma</td>
<td>CESC</td>
<td>308</td>
<td>8</td>
</tr>
<tr>
<td>Cholangiocarcinoma*</td>
<td>CHOL</td>
<td>36</td>
<td>15</td>
</tr>
<tr>
<td>Colon adenocarcinoma*</td>
<td>COAD</td>
<td>455</td>
<td>92</td>
</tr>
<tr>
<td>Lymphoid neoplasm diffuse large B-cell lymphoma</td>
<td>DLBC</td>
<td>48</td>
<td>0</td>
</tr>
<tr>
<td>Esophageal carcinoma*</td>
<td>ESCA</td>
<td>185</td>
<td>64</td>
</tr>
<tr>
<td>Glioblastoma multiform*</td>
<td>GBM</td>
<td>586</td>
<td>31</td>
</tr>
<tr>
<td>Head and neck squamous cell carcinoma*</td>
<td>HNSC</td>
<td>528</td>
<td>82</td>
</tr>
<tr>
<td>Kidney chromophobe renal cell carcinoma*</td>
<td>KICH</td>
<td>66</td>
<td>57</td>
</tr>
<tr>
<td>Kidney renal clear cell carcinoma*</td>
<td>KIRC</td>
<td>534</td>
<td>427</td>
</tr>
<tr>
<td>Kidney renal papillary cell carcinoma*</td>
<td>KIRP</td>
<td>291</td>
<td>87</td>
</tr>
<tr>
<td>Brain lower grade glioma</td>
<td>LGG</td>
<td>516</td>
<td>0</td>
</tr>
<tr>
<td>Liver hepatocellular carcinoma*</td>
<td>LIHC</td>
<td>377</td>
<td>87</td>
</tr>
<tr>
<td>Tumor Type</td>
<td>Code</td>
<td>Cases</td>
<td>Patients</td>
</tr>
<tr>
<td>---------------------------------------------------------</td>
<td>-------</td>
<td>-------</td>
<td>----------</td>
</tr>
<tr>
<td>Lung adenocarcinoma*</td>
<td>LUAD</td>
<td>519</td>
<td>179</td>
</tr>
<tr>
<td>Lung squamous cell carcinoma*</td>
<td>LUSC</td>
<td>504</td>
<td>240</td>
</tr>
<tr>
<td>Mesothelioma</td>
<td>MESO</td>
<td>87</td>
<td>1</td>
</tr>
<tr>
<td>Ovarian serous cystadenocarcinoma*</td>
<td>OV</td>
<td>586</td>
<td>130</td>
</tr>
<tr>
<td>Pancreatic adenocarcinoma*</td>
<td>PAAD</td>
<td>185</td>
<td>37</td>
</tr>
<tr>
<td>Pheochromocytoma and paraganglioma</td>
<td>PCPG</td>
<td>179</td>
<td>5</td>
</tr>
<tr>
<td>Prostate adenocarcinoma*</td>
<td>PRAD</td>
<td>500</td>
<td>117</td>
</tr>
<tr>
<td>Rectum adenocarcinoma*</td>
<td>READ</td>
<td>167</td>
<td>17</td>
</tr>
<tr>
<td>Sarcoma*</td>
<td>SARC</td>
<td>261</td>
<td>22</td>
</tr>
<tr>
<td>Skin cutaneous melanoma</td>
<td>SKCM</td>
<td>104</td>
<td>3</td>
</tr>
<tr>
<td>Stomach adenocarcinoma*</td>
<td>STAD</td>
<td>478</td>
<td>131</td>
</tr>
<tr>
<td>Testicular germ cell tumors</td>
<td>TGCT</td>
<td>150</td>
<td>0</td>
</tr>
<tr>
<td>Thyroid carcinoma*</td>
<td>THCA</td>
<td>507</td>
<td>99</td>
</tr>
<tr>
<td>Thymoma*</td>
<td>THYM</td>
<td>124</td>
<td>12</td>
</tr>
<tr>
<td>Uterine corpus endometrial carcinoma*</td>
<td>UCEC</td>
<td>545</td>
<td>51</td>
</tr>
<tr>
<td>Uterine carcinosarcoma</td>
<td>UCS</td>
<td>57</td>
<td>6</td>
</tr>
<tr>
<td>Uveal melanoma</td>
<td>UVM</td>
<td>80</td>
<td>0</td>
</tr>
</tbody>
</table>
Figure 1: Sample sizes for TCGA populations with more than 10 samples.

Table 1: Sample sizes for cancer and control populations in data for 32 cancers from The Cancer Genome Atlas (TCGA).

The cancer “code” is an abbreviation for each cancer. All samples are from tissue in the afflicted, cancerous region. For the cancer population, tissue samples are from tumorous tissue. For the normal population, tissue is from healthy normal tissue. Cancers marked with a star (⋆) were included in the analysis of pathway disturbance, on the basis of each sample population including more than 10 samples.

For our analysis, we downloaded and considered tissue collected from either “primary solid tumor” or “solid normal” tissue. We constructed sample populations from these two tissue types, so the populations permit comparison of tumorous and healthy tissue from the same location. Because of their situation within the corresponding tissue, we refer to the non-tumor...
samples as matched normal tissue.

For each tumor and matched normal tissue sample, we obtained multi-platform -omics data, collected on the following platforms.

1. Gene expression was measured on the Illumina HiSeq RNASeqV2 platform. Level-3 TCGA data was available as both raw read counts, and a normalized format. The latter is provided as fragments per kilobase of transcript per million mapped reads (FPKM), normalized as the upper quartile (UQ) of nonzero counts [Dillies et al., 2013, Grossman et al., 2016], or FPKM-UQ. We used these normalized values rather than the raw read counts.

2. DNA copy number variation (CNV) was collected using the Affymetrix Genome-Wide Human SNP Array 6.0 platform. We used data with germline copy number variants removed.

3. DNA methylation was measured using the HumanMethylation450 BeadChip. Observations are organized at the level of CpG sites.

We also obtained micro RNA (miRNA) observations that were collected on the Illumina HiSeq platform. We defer details of miRNA data processing until Section 2.2 below.

We performed additional data processing after downloading from TCGA, prior to analysis.

1. We applied a $\log_2$ transformation to the RNASeq data.

2. We averaged gene-level CNV by DNA region.

3. We averaged gene-level methylation beta values across all CpG sites for each gene.
1.2.2 Pathway Interaction Database

Much of the analysis that follows is oriented toward network analysis of genetic signaling pathways. To this end, we downloaded the 212 signaling pathways that comprise the NCI Pathway Interaction Database (PID, Schaefer et al., 2008). To obtain the PID, a collection of known gene networks specified as directed, functional relationships between genes, we used the graphite software [Sales et al., 2018] within R.

The PID pathways are defined over 2393 genes, which correspond to 2369 genes observed on 6973 -omics features in the TCGA dataset, across the three gene-level platforms for expression, methylation, and copy number. In particular, within each data type, we observed 2363 genes with expression measurements, 2279 genes with methylation, and 2331 genes with copy number. Of the 212 pathways that comprise the PID, 173 pathways were such that observations of expression were available for every gene in the pathway. The number of missing genes per pathway was one gene for 32 pathways, two genes for four pathways, three genes for one pathway, and four genes for two pathways.

1.3 Dissertation Outline

In this thesis, we considered several methods and procedures oriented towards integrative pathway analysis. These methods not only offer statistical models for signaling pathways that integrate multiple -omics data types, but also address real-world obstacles to applications, including missing data, effective communication and exploration of statistical output, and a computational method to identify important subpathways in known pathways.

In Chapter 1, we discussed the importance of -omics studies for understanding complex
diseases. We discussed some contemporary research in integrative genomics. We introduced a large-scale, ongoing research study, The Cancer Genome Atlas (TCGA), that provides access to multi-platform genomic data on cancer. We also discussed a dataset of signaling pathways, the NCI Pathway Interaction Database.

In Chapter 2, we gave an overview of approaches to pathway analysis of functional networks of genes. We gave an overview of a basic pathway model, NetGSA, as well as EMC-NetGSA, an integrative adaptation that jointly models gene expression, methylation, and copy number. We then introduced non-coding micro RNAs (miRNAs), and discussed their role in gene regulation. We introduced a dataset of functional miRNA-gene target relationships, and we proposed an extension of EMC-NetGSA that also includes miRNAs. This approach provides a secondary mechanism to account for correlation across genes beyond the graph topology of the signaling pathway. We demonstrated the method on an example pathway, using the TCGA dataset.

In Chapter 3, we presented a procedure to impute missing data in integrative genomics. We applied a matrix completion method in an iterative fashion, imputing non-rectangular missing data in a way that uses all available information in a genomic sample. Our approach permits application of a theoretically-appealing model for imputation to real-world data. We applied pathway analysis to the imputed data, and demonstrated through simulation the increased power of our strategy, with a smaller rise in false discoveries than the original SMC method. We demonstrated that the I3 method improves the precision of the basic SMC method, while still leveraging the linear dependencies that make SMC attractive in the first place.

Our data analysis demonstrated the power to stabilize coefficient estimates of subtypes, despite small sample sizes. Furthermore, it reinforced the importance of subtype analysis of
tumors, rather than across entire populations. Among the pathways in the NCI Pathway Interaction Database, we found many possible pathway disturbances. Some of these pathways play known roles in breast cancer, which supports evidence of disturbance in other pathways not previously known. The FOXA2 and FOXA3 transcription factor networks and the ErbB2/ErbB3 signaling events pathways, in particular, displayed pathway disturbances that have been relatively unexplored in breast cancer. Based on the results of the simulation study and the data analysis, iterative integrated imputation (I3) struck a balance between local accuracy and global structural constraints when imputing missing data. We demonstrated the bias introduced by the assumption that missing data occurs in blocks, and we propose a procedure based on SMC that includes the original block-missing SMC as a special case, while flexibly handling non-rectangular missing data. Our method imputed values that permit analysis of datasets with missing data, without providing imputed values that distort the output statistical analysis.

In Chapter 4, we discussed the challenges posed by large-scale data analysis of -omics Big Data. We reviewed some existing approaches to interactive data visualization, and we provided details of an interactive data visualization that gives access to our own pathway analysis of the previous chapters. We also discussed subpathway discovery, for the purposes of identifying important subnetworks among a broader pathway network. We proposed a greedy search algorithm, and implemented the procedure on both expression-only and integrated datasets. We also discussed the combinatoric and computational challenges posed by subgraph searches, and gave details of an interactive analysis tool that permits user-specified subpathway analysis.
Chapter 2

Network models for integrative pathway analysis

In this chapter, we discuss pathway analysis for -omics data that comprise known signaling pathways. These pathways are specified as known, directed graphs, with graph vertices that represent -omics features, and directed graph edges that represent known functional relationships.

Application of GSEA to genes a priori known to comprise a signaling pathway can give comparative insight into patterns of differential activity across multiple cancers [Ge et al., 2018]. But, more mature pathway methods have been introduced and applied. The random walk with restart has been applied to assess both network cohesion and compare the explanatory power of several different pathway databases [Huang et al., 2018]. The SAFE model [Baryshnikova, 2016] identifies local neighborhoods of high enrichment situated within larger networks. A factor model approach, PARADIGM, was applied to a single unified network of multiple signaling pathways [Vaske et al., 2010], and has also been used for comparative analysis of individual pathways across cancer types [Campbell et al., 2018]. Graphical methods have also been developed, for directed as well as undirected Gaussian networks [Ma et al., 2014b].
In Section 2.1, we provide an overview of the NetGSA framework for joint modeling of gene expression observations that correspond to the vertices in a graph for a genetic signaling pathway. The framework also offers hypothesis testing of differential expression between two populations, which we situate within the context of cancer and healthy populations. We also discuss an integrative refinement of the NetGSA model, EMC-NetGSA, which provides a joint model for several -omics data types observed at the gene level.

A natural fusion combines data integration and pathway analysis. This has been applied to comprehensive analysis of individual signaling pathway in a single cancer using gene-level covariates including expression, methylation, and copy number, and somatic mutations to analyze androgen receptor signaling [Robinson et al., 2015]. Also available for integration in pathway analysis is inclusion of non-gene features, like miRNA expression along with gene expression, methylation, and DNA copy number [Danielsen et al., 2015]. Pathway analysis has also been applied across multiple cancers and multiple pathways [Sanchez-Vega et al., 2018], and the Lemon-Tree model [Bonnet et al., 2015] used a module network model to integrate data and identify novel pathway components.

In Section 2.2, we propose a further extension of the EMC-NetGSA model to integrate -omics data collected for biomolecular units other than the gene, namely, non-coding microRNAs. We implement and discuss a data analysis of the BRAF pathway to demonstrate the new integrative method.
2.1 Pathway analysis for -omics observations of gene activity

The NetGSA model, introduced by Shojaie and Michailidis [2009], offers a graphical model for joint modeling of gene expression. The model accounts for known functional dependencies between genes comprising a signaling pathway. As originally proposed, the model applied only to directed acyclic graphs (DAGs). The same model was shown by Shojaie and Michailidis [2010] to apply to a more general class of directed graphs. Moreover, that same work proposed a procedure to apply the basic NetGSA model to arbitrary directed graphs. The work of Ma et al. [2014b] extended the model further, to include undirected Gaussian networks.

Each of these iterations on the basic NetGSA framework focused on signaling pathway graphs and models of gene expression. The model was further extended by Zhang et al. [2017a] to the -omics setting. Their integrative EMC-NetGSA was so named because it provides a joint model for expression, methylation, and copy number data, each observed at the level of individual genes.

We proceed constructively. We start with a genetic signaling pathway specifying known functional relationships between \( p \) genes. We represent the pathway as a directed graph \( G = \{V, E\} \), where \( V \) is a set of \( p \) graph vertices, and \( E \) is a set of directed edges between elements of \( V \). The graph vertices in \( V \) represent genes, while the edges in \( E \) represent the functional pathway interactions.

We represent the graph topology of \( G \) by a \( p \times p \) graph adjacency matrix, denoted \( A^*_E \). We use the subscript “E” to emphasize that the graph specifies relationships between vertices that represent measurements of gene expression. The element \( \alpha_{jk} \) of \( A^*_E \) takes the value 1 if there exists a directed edge from vertex \( k \) to vertex \( j \), and 0 otherwise, for all \( j, k = 1, \ldots, p \). \( \alpha_{jk} \)
may be interpreted as an indicator function for the conditional dependence of gene $j$ on gene $k$. Corresponding to each gene we observe gene expression, which we format as a vector with $p$ elements, $y_{i1}$, where $i = 1, \ldots, N$ indexes each of $N$ independent samples.

For each element in $y_{i1}$, we also observe copy number and methylation beta values for the corresponding genes. Denote the vectors of methylation and copy number by $y_{i2}$ and $y_{i3}$, respectively. The EMC-NetGSA integration procedure, proposed in Zhang et al. [2017a], incorporates the methylation and copy number measurements into the graph topology of $\mathcal{G}$. In particular, we augment $\mathcal{G}$ by adding $2p$ vertices to $\mathcal{V}$, one each for methylation and copy number for every gene. We also add $2p$ edges to $\mathcal{E}$, directed from the methylation or copy number vertex to the corresponding gene expression vertex. Therefore, the final EMC-NetGSA graph consists of two separate layers of network topology:

1. The primary, inter-gene signaling pathway network, giving edges between elements of $y_{i1}$;
2. A secondary integration network of edges between elements of $y_{i2}$ or $y_{i3}$, and the corresponding elements in $y_{i1}$.

Denote the full observation vector for sample $i$ by $y_{i123} = (y_{i1}', y_{i2}', y_{i3}')'$. For some genes, copy number or methylation values are unavailable, and in these cases, the number of copy number (methylation) observations in $y_{i2}$ ($y_{i3}$) will be less than $p$. We resolve this by removing the non-expression vertex from $\mathcal{V}$ and the corresponding integrated edge from $\mathcal{E}$. Without loss of generality, we assume $y_{i2}$ and $y_{i3}$ each contain $p$ elements, and adjust $y_{i123}$ accordingly.

The EMC-NetGSA integration is implemented in the adjacency matrix by concatenating
identity matrices along the right-hand margin and zero matrices along the bottom:

$$A^*_\text{EMC} = \begin{pmatrix} A^*_E & I_{p\times p} & I_{p\times p} \\ O_{2p\times p} & O_{2p\times p} & O_{2p\times p} \end{pmatrix}$$  (2.1)

where $O_{2p\times p}$ is a $2p \times p$ matrix of zeros. Here, the identity matrices reflect the directed edges from copy number and methylation vertices to expression, and the zero matrices reflect the lack of any network structure within or between the additional data types, or a directed relationship from genes to -omics features.

Each pair of vertices in $\{(j, k) | \alpha_{jk} \neq 0\}$, where $\alpha_{jk}$ is an element of $A^*_\text{EMC}$, specifies a directed edge from $y_{ik}$ to $y_{ij}$, $i = 1, \ldots, N$, $y_{ik}, y_{ij} \in y_{i123}$. This is a conditional dependence relation of vertex $j$ on vertex $k$, given the effects of the other $(3p - 2)$ network features. In the context of Gaussian graphical models, conditional dependence of random variables $X_j$ and $X_k$, conditional on a set $Z$ of additional random variables of interest, is formalized as the partial correlation $\rho_{jk}$ with respect to $Z$. That is, $\rho_{jk} = \text{corr}(X_j \setminus Z, X_k \setminus Z)$, where $X_j \setminus Z = X_j - P_Z X_j$ is the orthogonal complement of $X_j$ with respect to $Z$, and $P_Z$ is a projection onto $Z$ [Krämer et al., 2009]. Intuitively, $\rho_{jk}$ represents the association between -omics features $j$ and $k$, controlling for each of their associations with other features in the pathway. We estimate $\rho_{jk}$ by the sample partial correlation $r_{jk}$, which is obtained by regressing each of $X_j$ and $X_k$ separately on $Z$, and then calculating Pearson’s correlation coefficient between the vectors of residuals. Using the partial correlations and $A^*_\text{EMC}$, we construct the weighted adjacency matrix $A_{\text{EMC}}$ with elements $a_{jk} = r_{jk} \alpha_{jk}$, $j, k = 1, \ldots, 3p$.

To build a statistical model for the data $\{y_{i123}\}_{i=1}^{N}$, we consider a transformation of $A_{\text{EMC}}$...
introduced by Shojaie and Michailidis [2009]. The influence matrix $\Lambda_{\text{EMC}}$ captures the cumulative network effect of each gene on the expression of all others. In the case of directed acyclic graphs (DAGs), the authors derive the identity $\Lambda_{\text{EMC}} = (I_{3p} - A_{\text{EMC}})^{-1}$.

Shojaie and Michailidis [2010] extended this formula to apply to general, non-DAG graphs. In particular, they demonstrate that the relationship between the influence and adjacency matrices applies to any graph that is substochastic, which occurs when the adjacency matrix has eigenvalues all of which have magnitude smaller than 1. They apply a graph theoretic result, Gershgorin’s Disk Theorem, to obtain a limit approximation for the influence matrix when the adjacency matrix is not substochastic, in particular,

$$\alpha(\delta) = (\alpha_{jk}(\delta))_{j=1,...,3p; k=1,...,3p}, \quad \delta > 0 \quad (2.2)$$

$$\alpha_{jk}(\delta) = \frac{A_{\text{EMC}jk}}{(\sum_{\ell=1}^{3p} |A_{\text{EMC}j\ell}|) + \delta} \quad (2.3)$$

$$\Lambda_{\text{EMC}} = \lim_{\delta \to 0} (I_{3p} - \alpha(\delta))^{-1} \quad (2.4)$$

Then, an approximation to $\Lambda_{\text{EMC}}$ is constructed using a small value of $\delta > 0$.

The EMC-NetGSA model uses $\Lambda_{\text{EMC}}$ to structure the mean in a mixed-effects model with unknown regression coefficients $\beta_{\text{EMC}} \in \mathbb{R}^{3p}$, with $E_{y_{123}} = \Lambda_{\text{EMC}} \beta_{\text{EMC}}$. We may interpret $\beta_{\text{EMC}}$ as the network-adjusted expression coefficients for the 3p-omics features, and we note $\Lambda_{\text{EMC}}$ also structures the covariance of $y_{123}$.

The NetGSA framework also provides for significance testing of two populations, control (healthy) and treatment (disease). Denote the population label for sample $i$ as $c_i \in \{C, T\}$. The control and treatment populations are modeled using separate adjacency matrices $A_{\text{EMC}}^{C}$,
\( \mathbf{A}^T_{\text{EMC}} \), corresponding to influence matrices \( \mathbf{A}^C_{\text{EMC}}, \mathbf{A}^T_{\text{EMC}} \), and parameterized with population-specific pathway-adjusted expression parameters \( \beta^C_{\text{EMC}}, \beta^T_{\text{EMC}} \), respectively.

The EMC-NetGSA statistical model is then given by

\[
\mathbf{y}_{i123} = \mathbf{A}^C_{\text{EMC}} \mathbf{\beta}^C_{\text{EMC}} + \mathbf{A}^C_{\text{EMC}} \mathbf{\gamma}_{i123} + \mathbf{\epsilon}_{i123}, \quad i = 1, \ldots, N \tag{2.5}
\]

\[
\mathbf{\gamma}_{i123} \sim \mathcal{N}_{3p}(\mathbf{0}_{3p}, \sigma^2_{\gamma} \mathbf{I}_{3p}) \tag{2.6}
\]

\[
\mathbf{\epsilon}_{i123} \sim \mathcal{N}_{3p}(\mathbf{0}_{3p}, \sigma^2_{\epsilon} \mathbf{I}_{3p}) \tag{2.7}
\]

To test differential activity in subsets of the pathway’s genomic features, represented by elements of \( \beta^C_{\text{EMC}} \) and \( \beta^T_{\text{EMC}} \), we specify features of interest through an indicator vector \( \mathbf{b} \).

We use the NetGSA network contrast \( \ell_{\text{EMC}} = \langle -\mathbf{b} \cdot \mathbf{b}^T_{\text{EMC}} \rangle \) to compute the test statistic \( T_{\text{EMC}} \propto \ell \mathbf{\beta}_{\text{EMC}} \), where \( \mathbf{\beta}_{\text{EMC}} = (\beta^C_{\text{EMC}}, \beta^T_{\text{EMC}})' \). \( T_{\text{EMC}} \) follows a Student’s \( t \) distribution with degrees of freedom estimated by the Satterthwaite approximation.

### 2.2 miRNA-gene targeting

Bioinformatic data is increasingly collected from functional genomic units in the cell other than genes. These new genomic entities present new opportunities for integrative analysis, and include non-coding micro RNAs (miRNAs). These small, noncoding RNAs are understood to play a regulatory role for genes, and the identification of miRNA-gene targets is critical to understanding the biological function of each [Hammond, 2015].

However, to date, development of advanced statistical methods has lagged behind the rapid growth in experimental methods and available data. We address the lack of sophisticated
technical methods by proposing an integrated model for simultaneous network analysis of
miRNA-gene target interactions and their association with genetic signaling pathways. We
extend existing methods for graphical analysis of gene-level -omics data, specifically the EMC-
NetGSA model, to include functional relationships between miRNAs and genes. In particular,
we represent miRNA-gene target relationships through the same graphical structure as we used
in the previous section to integrate methylation and copy number.

Individual miRNA are believed to target genes in a functional manner [Lewis et al., 2005],
and it is often the case that single miRNAs target multiple genes. To model correlated gene
activity due to a shared miRNA parent, miRNA-gene target interactions must be known and
available to researchers. Early miRNA-gene target research validated individual targets exper-
imentally, but the combinatoric problems introduced by large numbers of genes and miRNAs
motivated meta-analytic and computational approaches. One study to identify miRNA-gene
targets was miRTarBase [Hsu et al., 2010], which identified gene targets for fewer than 700
miRNAs by manual aggregation of experimental evidence. Despite the intensive cost of this
type of study, the authors anticipated the use of small, manually-integrated datasets as inputs
to computational approaches to identify interactions.

Computational resources such as the miRTar [Hsu et al., 2011] and miRDIP [Tokar et al.,
2017] aggregate individual experimental findings on miRNA-gene target relationships. These
centralized databases offer a starting point for more sophisticated functional analysis across
datasets with thousands of genomic features, reducing the combinatoric burden of a purely
computational association analysis between miRNAs and genes.
Early statistical analyses of miRNAs were characterized by straightforward statistical methods. For example, the CORNA method introduced by Wu and Watson [2009] applied methods including the hypergeometric and Fisher’s exact tests to assess differential activity in miRNA-gene interactions and networks. Subsequently, miRNA analysis has gained popularity in application to cancer datasets. Enerly et al. [2011] studied miRNA suppression in a novel miRNA and gene expression breast tumor dataset. They observed roles played by miRNAs in genetic activity, immunohistochemical characteristics, and system-level cellular processes. Yu et al. [2019] identified specific biomarkers with differential survival outcomes in lung cancer. Li et al. [2018] used differential correlation analysis between miRNAs and genes in cancer and normal populations, which they applied across multiple distinct cancer types.

Joint analysis of separate functional components may integrate marginally univariate data, such as miRNA expression paired with gene expression [Chu et al., 2015]. But, multi-level integrative analysis is also performed, such as an analysis of breast and gynecological cancers including somatic mutations and copy number, gene expression, methylation, and multiple types of non-coding RNA [Berger et al., 2018].

The nature of miRNA research is foundationally integrative in nature, given the fundamental nature of miRNA-gene interactions. Nevertheless, integrative analysis is not restricted to focus on the relationship between miRNAs and genes. Some researchers integrate epigenetic data, including Miao et al. [2017], who analyzed the relationship between miRNAs and DNA methylation in sheep. Du and Zhang [2015] integrated methylation in a small-sample analysis of lung cancer that also included expression in genes and miRNAs. Volinia and Croce [2013] applied a similar approach to breast cancer data. Transcriptomic data has also been
Modern informatic methods permit large-scale analyses to identify miRNA targets. Frameworks such as miRTarBase and DIANA-miRPath [Vlachos et al., 2015] utilize web interfaces to access and explore association analyses between miRNAs and genes. DIANA-miRPath adapts methods originally applied to gene expression. In addition to Fisher’s exact test, they also test for differential activity in miRNAs using the enrichment analysis method of Bleazard et al. [2015]. Computational approaches have been used to identify targets, too. Hsu et al. [2011] proposed miRTar, a successor to miRTarBase that used gene set enrichment analysis for significance testing of differential activity. These are often applied to specific phenomena or systems. Other databases include DIANA-TarBase [Karagkouni et al., 2017] and TargetScan [Agarwal et al., 2015]. Coll et al. [2015] used correlation analysis to find miRNA-gene targets related to cirrhosis of liver tissue. Godard and van Eyll [2015] performed pathway analysis of miRNA in the context of Alzheimer’s disease with an approach derived from enrichment analysis. But, methods for comprehensive, graph-theoretic pathway analysis remain undeveloped.

2.2.1 miRNA-gene target dataset

As described above in Section 1.2.1, it is straightforward to construct a direct mapping between the basic methylation and copy number features, on the one hand, and a single gene, on the other. In contrast, no such standard, direct mapping exists between genes and miRNAs. Instead, we use resources on functional miRNA-gene targets to construct an integrated statistical model. Substantial work has been done to identify miRNA-gene targets. One resource that quantifies the degree of experimental evidence in support of a given miRNA-gene
target interaction is mirDIP. Tokar et al. [2017] compiled the database as a meta-analysis to integrate predicted miRNA-gene targets from 30 separate sources of experimentally-validated interactions. It includes information on the degree to which the source databases overlap in their conclusion.

2.2.2 Data processing

As discussed above in Section 1.2.1, we aggregated expression, methylation, and copy number data at the level of individual genes. Similar to gene expression, miRNA expression data are available from TCGA in two formats: raw read counts, and normalized reads-per-million (RPM). Raw read counts were collected on the miRNASeq platform, and the TCGA processing pipeline outlined by Chu et al. [2015] is consistent with procedures in comparable projects, such as ENCODE [Consortium et al., 2012]. After alignment and read trimming, a library of approximately 22 base pairs of mature strands was used with an insert length of approximately 22.

Typical miRNASeq analyses use methods traditionally developed for RNASeq. For example, Stokowy et al. [2014] employed RPM normalization and cited its original definition from Mortazavi et al. [2008], in the context of gene expression. They cited other work that applies RPM normalization to miRNASeq data, including Chen et al. [2013]. Following RPM normalization but prior to the primary analysis, those authors applied a log<sub>2</sub> transformation. Han et al. [2018] integrated gene and miRNA expression. They first applied FPKM-UQ normalization RNASeq gene expression values, then calculated RPM for miRNASeq values and transformed logarithmically. TCGA provides RPM normalized transformation of the data, so
for our integrative analysis, we applied a log2 transformation to the RPM-normalized values. Empirically, we observed that this normalization was comparable to FPKM-UQ applied to the raw read counts.

2.2.3 Method

Integration of miRNA observations into the NetGSA framework proceeds from a known set of functional miRNA-gene target relationships. In addition to the signaling pathway topology represented in the adjacency matrix \( A_E \), we also consider \( g \) miRNAs with known functional targets among the \( p \) genes observed in \( y_{i1} \). For each sample, we observe a vector of \( g \) elements \( y_{i4}, i = 1, \ldots, N \), the values of which measure miRNA expression.

To represent the miRNA-gene target interactions, we construct a \( p \times g \) graph adjacency matrix \( A_{mi}^* \), with columns that correspond to miRNAs and rows that correspond to genes. Then, as before in the NetGSA and EMC-NetGSA models, the element \( \tau_{j\ell} \) of \( A_{mi}^* \) is an indicator value for miRNA \( \ell \) targeting gene \( j \), \( j = 1, \ldots, p \), \( \ell = 1, \ldots, g \). Here, we have represented each miRNA-gene target as a directed edge, and \( \tau_{j\ell} = 1 \) (Gene \( j \) is conditionally dependent on miRNA \( \ell \)).

We construct an integrated adjacency matrix \( A_{miE}^* \) that includes both the signaling pathway and the miRNA-gene targets:

\[
A_{miE}^* = \begin{pmatrix}
A_E^* & A_{mi}^* \\
O_{g\times p} & O_{g\times g}
\end{pmatrix}
\] (2.8)

which is a square \((p + g) \times (p + g)\) matrix with elements taking values of either 0 or 1. Correspondingly, we integrate the graph \( G \) by adding \( g \) vertices to \( V \), one for each miRNA.
For each miRNA-gene target, we add to $\mathcal{E}$ a directed edge from the miRNA vertex to the gene vertex.

Then, we use the EMC-NetGSA method to construct a fully-integrated adjacency matrix, by augmenting $\mathcal{V}$ to contain a total of $(3p + g)$ vertices, and adding the corresponding $2p$ additional edges from the methylation and copy number vertices to the corresponding gene expression vertex.

Mechanically, this integration scheme produces another, fully-integrated adjacency matrix, $A_{\text{miEMC}}^*$, which composes $A_{\text{miE}}^*$ with identity matrices:

$$A_{\text{miEMC}}^* = \begin{pmatrix} A_E^* & I_{p \times p} & I_{p \times p} & A_{\text{mi}}^* \\ O_{(g+2p) \times p} & O_{(g+2p) \times p} & O_{(g+2p) \times p} & O_{(g+2p) \times g} \end{pmatrix} \quad (2.9)$$

where $O_{m \times n}$ is a $m \times n$ matrix of zeros. The identity matrices reflect that the network topology consists of (1) directed edges between elements of $y_{i1}$, and (2) directed edges from elements in $y_{i2}, y_{i3}, y_{i4}$ to associated elements in $y_{i1}$. In particular, we do not model any directed network effect from genes to any of miRNAs, methylation, or copy number vertices. As before, without loss of generality, we consider the full $(g + 3p) \times (g + 3p)$ adjacency matrix $A_{\text{miEMC}}^*$, with the knowledge that its true dimension $q$ is such that $q \leq (g + 3p)$.

We may interpret the network adjacency matrix specified in Equation 2.9 as a composition of three distinct network layers: (1) a primary network specified between elements of $y_{i1}$; (2) a miRNA-gene integration layer giving directed network relationships, possibly many-to-one, from miRNA elements of $y_{i2}$ to gene elements of $y_{i1}$; and (3) a within-gene integration layer between copy number elements of $y_{i3}$ and methylation elements of $y_{i4}$ and their matching gene
element in \( y_{i1} \). Therefore, the graph represented by \( A_{\text{miEMC}}^{*} \) simultaneously provides for causal relationships between genes due to the signaling pathway; correlation between genes due to miRNA-gene targeting; and reduces the statistical noise associated with observations of gene expression by removing effects due to epigenetic and transcriptional factors. Moreover, each of these three components is motivated by scientific understanding of the complex functional processes that produce the observed genomic features.

For expositional clarity, we define \( m = (g + 3p) \) and denote the full observation vector of \( m \) elements by \( y_{i} = (y_{i1}', y_{i2}', y_{i3}', y_{i4}')' \). Writing \( A^{*} \equiv A_{\text{miEMC}}^{*} \), denote the elements of \( A^{*} \) by \( \delta_{jk}, j, k = 1, \ldots, m \). Depending on the location of \( \delta_{jk} \) in \( A^{*} \), \( j \) and \( k \) may index vertices for gene expression, miRNA expression, gene copy number, or gene methylation. Each pair of vertices indexed by \( \{(j, k) \mid \delta_{jk} \neq 0\} \) corresponds to a directed edge from \( y_{ik} \) to \( y_{ij} \), \( i = 1, \ldots, N \).

Then, the fully-integrated NetGSA model is

\[
y_{i} = \Lambda^{c_{i}}\beta^{c_{i}} + \Lambda^{c_{i}}\gamma_{i} + \epsilon_{i}, \quad i = 1, \ldots, N
\]

(2.10)

\[
\gamma_{i} \sim N_{m}(0_{m}, \sigma_{\gamma}^{2}I_{m})
\]

(2.11)

\[
\epsilon_{i} \sim N_{m}(0_{m}, \sigma_{\epsilon}^{2}I_{m})
\]

(2.12)

where, again, \( c_{i} \in \{C, T\} \) indexes the sample population as either control (healthy) or treatment (cancer). Here, \( \gamma_{i} \) is a sample-level random effect.

Inference proceeds as with the expression-only NetGSA model, and the integrated EMC-NetGSA model: We test for differential activity in subsets of the pathway features, corresponding to elements in \( \beta^{C}, \beta^{T} \), by indicating the features of interest in an indicator vector \( b \). The
NetGSA network contrast is \( \ell = (b \cdot b \Lambda_C, b \cdot b \Lambda_T) \), and this yields a test statistic \( T \propto \ell \beta \), \( \beta = (\beta_C', \beta_T')' \). \( T \) follows a Student’s \( t \) distribution with degrees of freedom estimated using Satterthwaite’s approximation.

We may also characterize the miRNA integration into the network topology through a set-theoretic lens, rather than the previous focus on the adjacency matrix. We again proceed from a signaling pathway defined on \( p \) vertices. Denoting the set of expression nodes by \( \mathcal{V}_E \) and the set of signaling pathway edges by \( \mathcal{E}_E \), we characterize the signaling pathway as the directed graph \( G_E = \{\mathcal{V}_E, \mathcal{E}_E\} \).

Next, we consider the set of \( g \) miRNAs known to target the \( p \) genes in \( \mathcal{V}_E \). Denote the set of miRNAs by \( \mathcal{V}_{\text{mi}} \), which represent observations miRNA expression, and the set of edges for the miRNA-gene targets by \( \mathcal{E}_{\text{mi}} \). The elements of \( \mathcal{E}_{\text{mi}} \) are directed edges defined from a miRNA vertex in \( \mathcal{V}_{\text{mi}} \) to a gene vertices in \( \mathcal{V}_E \). Then, we may consider the integrated miRNA-expression graph \( G_{\text{miE}} = \{\mathcal{V}_{\text{miE}}, \mathcal{E}_{\text{miE}}\} \), where \( \mathcal{V}_{\text{miE}} = \mathcal{V}_E \cup \mathcal{V}_{\text{mi}} \), and \( \mathcal{E}_{\text{miE}} = \mathcal{E}_E \cup \mathcal{E}_{\text{mi}} \).

For each gene, we define two additional vertices, one for methylation and the other for copy number. Denote the sets of methylation and copy number vertices by \( \mathcal{V}_M \) and \( \mathcal{V}_C \), respectively. We integrate these vertices with the vertices of \( \mathcal{V}_E \) by constructing edge sets \( \mathcal{E}_M, \mathcal{E}_C \), each of which contains \( p \) edges. Without loss of generality, we consider a single index \( j = 1, \ldots, p \), which relates the \( p \) gene-level observations in \( \mathcal{V}_{\text{EMC}} = \mathcal{V}_E \cup \mathcal{V}_M \cup \mathcal{V}_C \), such that \( \nu_{Ej} \in \mathcal{V}_E \) gives the expression vertex for gene \( j \), \( \nu_{Mj} \in \mathcal{V}_M \) gives the methylation vertex for gene \( j \), and \( \nu_{Cj} \in \mathcal{V}_C \) gives the copy number vertex for gene \( j \).

Denote the final, joint graph by \( G_{\text{miEMC}} = \{\mathcal{V}_{\text{miEMC}}, \mathcal{E}_{\text{miEMC}}\} \), where \( \mathcal{V}_{\text{miEMC}} = \mathcal{V}_E \cup \mathcal{V}_{\text{mi}} \cup \mathcal{V}_M \cup \mathcal{V}_C \) and \( \mathcal{E}_{\text{miEMC}} = \mathcal{E}_E \cup \mathcal{E}_{\text{mi}} \cup \mathcal{E}_M \cup \mathcal{E}_C \). To illustrate the integrative graph topology, we
show in Figure 2 a toy example. In this example, we set $p = 3$ and $g = 3$. The signaling pathway is a 2-layer binary tree that contains a root node with two child nodes. These gene expression nodes are labeled with “G”. Two of the miRNAs target one gene each, while the third miRNA targets all three genes. miRNAs are labeled “mi”. Finally, for each gene, the toy graph contains one methylation vertex and one copy number vertex.

Figure 2: A toy example of the integrated graph $G_{\text{miEMC}}$. White vertices (labeled “G”) are gene expression nodes, which specify a signaling pathway topology. The three miRNA nodes (red, labeled “mi”) target the genes, and methylation and copy number nodes integrate the multi-platform, gene-level data (black and grey, respectively).

2.2.4 Results

First, we use a simulation study to demonstrate the improved power of our method due to the integration of miRNA-gene target information. We then apply our method to the TCGA
dataset for pathway analysis of the BRAF signaling pathway.

Simulations

Our simulation study follows the template of the simulations in the EMC-NetGSA paper [Zhang et al., 2017a]. Broadly, we construct a pathway composed of a binary tree signaling pathway; we model miRNA-gene targets that drive correlated gene expression; and we integrate the methylation and copy number within genes. This supplies three separate layers of network information that may contribute to gene expression, and we examine the relationship between -omics integration and statistical power.

We modeled a signaling pathway consisting of a five-level binary tree containing 31 genes. We integrated miRNA with directed edges from miRNA to corresponding genes. For every gene, we added three distinct miRNA vertices, i.e., each had out-degree 1. We then partitioned the genes in $\mathcal{V}$ into disjoint sets of two, proceeding from the root node. We assigned to each pair of genes one shared miRNA, i.e., with out-degree 2. Finally, we repeated this procedure for sets of three genes. We assigned each triplet one miRNA with out-degree 3. The network topology for the control population signaling pathway used the same integrated binary tree structure, with all edges in the tree’s left branch removed. For EMC-NetGSA integration, we add to $\mathcal{G}$ directed edges into each gene from two vertices representing methylation and copy number.

The network topology of the simulated binary tree is shown in Figure 3. We set the correlation between expression vertices to 0.8 in the top third (two levels) of the tree; association is 0.5 in the middle third (third level); and association 0.2 in the final third (final level).
Figure 3: A simulated five-level binary tree signaling pathway. Shown is the treatment network, which consists of: a full five-level binary tree of genes (large grey nodes); miRNA targeting 1, 2, and 3 genes (colored red, yellow, and green, respectively); and methylation and copy number nodes (white and black, respectively). Gene “G01” corresponds to the root node, which is the first level of the binary tree; “G02” and “G03” the second level; and so on. All integrated edges, that is, those from miRNA, methylation, or copy number nodes, are directed into the relevant gene node.

We set the magnitude of the association strength between miRNA and their gene targets to 0.4. In the TCGA dataset, we found that the partial correlation coefficients between miRNA
and expression were generally symmetric. Therefore, we assigned alternating edges from miRNAs to genes to have positive and negative association, respectively. This may be understood as simulating cases where miRNAs with multiple gene targets have the same sign for association, as well as different associations with different genes for a single miRNA. We set the association 0.5 between copy number and gene expression, and -0.25 between methylation and expression.

We generated observation vectors \( y_i, i = 1, \ldots, N \), from the statistical model given in Equation 2.10, where \( N = N_C + N_T \). The number of control samples was \( N_C = 50 \), and the number of treatment samples was \( N_T = 150 \). This reflects the imbalanced sample sizes in the real cancer data sets. We set the variance parameters as \( \sigma^2_\gamma = 5 \) and \( \sigma^2_\epsilon = 0.5 \).

Denoting the mean vectors for gene expression, miRNA expression, gene copy number, and gene methylation by \( \beta^C_1, \beta^C_2, \beta^C_3, \beta^C_4 \), we simulated two scenarios for the network-adjusted mean parameter \( \beta \). Here, \( c \in \{C, T\} \) indexes the control and treatment populations. In the first scenario, we assigned \( \beta^C_j = 0 \), \( c \in \{C, T\}, j = 1, 2, 3, 4 \). In the second mean scenario, we held \( \beta^C_j = 0 \) for all \( j \). For the top two-thirds levels of the binary tree, we set \( (\beta^T_1, \beta^T_2, \beta^T_3, \beta^T_4) = (0.25, 0.5, 1.0, 0.5) \). In the bottom third of the binary tree, we maintained \( (\beta^T_1, \beta^T_2, \beta^T_3, \beta^T_4) = 0 \), as in the first scenario.

For each simulated dataset, we tested four gene sets for differential activation: (1) the full binary tree; (2) the top third of the tree; (3) the top two thirds of the tree; and (4) the bottom third of the tree. We estimated the miEMC-NetGSA model for the entire simulated dataset, as well as the NetGSA variants with the adjacency matrices \( A_{\text{miE}}, A_{\text{EMC}}, \) and \( A_{E} \), as well as the corresponding hypothesis test for each network. We ran 1000 replicates of the simulation.
We calculated the power for each method by the proportion of hypothesis test $p$-values that were significant at the $\alpha = 0.05$ level, i.e., the proportion of replicates for which we reject the null hypothesis of no difference in pathway-adjusted mean parameters.

Figure 4 shows boxplots of the $-\log_{10} p$-values from the significance tests. The left-hand panel shows the results of the first mean scenario, in which no features are differentially expressed. The right-hand panel shows the second mean scenario, in which the top two-thirds of the binary tree signaling pathway are differentially expressed. We compare the performance of miEMC-NetGSA with the other integrated models described above.
Figure 4: Boxplots of $-\log_{10} p$-values from simulation study. The left-hand panel shows significance tests for four gene sets of interest, under the first mean scenario in which no -omics features exhibit differential activation. The right-hand panel shows the second mean scenario, in which the top two-thirds of the simulated binary tree signaling pathway is differentially activated, but the final third is equal in the control and treatment populations. miEMC-NetGSA is shown in green, miE-NetGSA in yellow, EMC-NetGSA in blue, and NetGSA (“E”) on expression only in white.

The first mean scenario permits assessment of the false positive rate under different -omics integration schemes. We observe that in all gene sets that we tested, all four NetGSA-based methods have low false positive rates. Most importantly, integration of miRNA with expression alone (“miE”) does not cause an elevated the false positive rate over the original NetGSA method. Likewise, although the false positive rate is somewhat elevated in miEMC-NetGSA, we observe that it is not elevated significantly over the existing EMC-NetGSA method. Therefore, we do not attribute to the miRNA integration a meaningful increase of the type I error rate.
Further, we note that the number of miRNAs is large relative to the number of genes. This causes the type I error rate to be lower for the methods that integrate miRNA-gene targets, shown in the test set of the top 1/3 genes. The reason is that the larger overall number of features provides increased accuracy to miEMC-NetGSA and miE integration, so more substantial information is available to the inference procedure than the methods with only gene-level network features.

The second mean scenario provides an assessment of the method’s power. We observe that in the three test sets that contain differentially activated features, the power of models that integrate miRNA-gene targets dominates the gene-only analyses. In fact, an integrated model of miRNA and expression has power comparable to EMC-NetGSA integration of methylation and copy number, although miRNA-gene expression remains under-powered relative to EMC-NetGSA.

Also prominent is the increased power of the new method, which integrates both the miRNA-gene targets and gene copy number and methylation. Integration of miRNA reduces the type II error rate of the EMC-NetGSA model; equivalently, it increases the analytic power. Finally, the miRNA methods continue to exhibit low rates of type I errors for the gene set that is not differentially activated.

Taken in tandem, the results of the two simulation scenarios confirm the value of pathway analysis that integrates miRNA-gene targets. We find miRNA contributes to pathway analyses improved statistical power, relative to analyses conducted solely at the level of individual genes. At the same time, we find that miRNA integration does not artificially elevate the false positive rate. Finally, although the addition of miRNA to an expression-only analysis improves power,
the increase is marginally less substantial than is provided by integrating copy number and methylation. But, our composition of these two integration schemes achieves the highest statistical power, and does not noticeably increase the type I error rate.

Pan-cancer analysis of the BRAF pathway

To demonstrate our procedure, we performed data analysis of the BRAF pathway, a genetic signaling pathway previously studied by Zhang et al. [2017a]. The left-hand panel of Figure 5 shows the network topology of the BRAF pathway. It consists of 10 genes—AKT1, BRAF, MAP2K1, MAP2K2, MAPK1, MTOR, NRAS, PIK3CA, PTEN, and RAF1—which are connected by 12 directed edges. The BRAF pathway is a DAG, and this property is preserved under integration with miRNA, copy number, and methylation.
Figure 5: Network diagrams for the BRAF signaling pathway. The left-hand network is the directed graph representing the BRAF genetic signaling pathway, consisting of 10 genes with 12 directed edges between. The right-hand network contains the same 10 gene vertices, but edges represent a shared miRNA parent. In other words, two genes that share an edge are both targets of a single miRNA, and possibly several. The graph contains 25 such edges. miRNA-gene targets are chosen as those entries in the mirDIP database with with an “very high” confidence score. Neither the BRAF nor MAP2K2 genes are targeted by any miRNAs.

The mirDIP database compiled by Tokar et al. [2017] aggregates predicted miRNA-gene target relationships from several experimentally-validated sources. The database assigns each miRNA-gene pair found across any of the 30 sources a composite integrative score. The score,
valued on the interval $[0, 1]$, quantifies the strength of experimental evidence that supports the existence of the interaction. The scores are stratified by so-called “confidence classes,” expressed in the labels “very high,” “high,” “medium,” and “low” confidence. These classes respectively represent the top 1% of scores (very high); the next 4% of scores (high); remainder of top 33% of scores (medium); and all other scores (low). The classes offer a discrete criterion for determining whether to include in $A_{\text{miEMC}}^{*}$ a specific miRNA-gene interaction. We downloaded their mirDIP unidirectional database, version 4.1, and we considered both 3’ and 5’ UTR miRNA entries for miRNAs that were also present in the TCGA dataset.

Figure 6 shows a scatterplot of the miRNA-gene target scores from mirDIP for all genes in the BRAF pathway. For a given gene, we identified all miRNAs in the database that target that gene. In the figure, the scores are grouped by gene, and colored by confidence class: blue represents very high confidence, green is high confidence, yellow is medium confidence, and red is low confidence.

In our analysis, we used the subset of miRNA-gene targets for which the scores in Figure 6 belong to the “very high” confidence class. The right-hand panel of Figure 5 shows the secondary graph formed between genes that are both targeted by a single miRNA parent. Compared with the genetic signaling pathway, this secondary network has a denser edge set: whereas the BRAF pathway contains 12 edges, the miRNA co-target graph contains 25 edges.

More generally, this larger edge set due to miRNA-gene targets indicates that miRNA integration substantially complexifies the network structure used as input for the pathway analysis. This contrasts with the underlying simplicity of the original graph: whereas the signaling pathway consists of 10 genes, miRNA integration introduces to the network 238
Figure 6: miRNA-gene interaction scores for all miRNA in TCGA dataset that target any of the 10 genes that comprise the BRAF pathway. Scores are grouped by gene, and colored by confidence class: very high, high, medium, and low confidence are colored blue, green, yellow, and red, respectively. For our data analysis, we included only miRNA-gene target interactions with “very high” confidence that the interaction exists, i.e., the blue scores.
vertices for miRNA observations. Nearly half of these miRNAs target multiple genes in the BRAF pathway.

The miRNA-gene target subnetwork, corresponding to the unweighted adjacency matrix $A_{mi}^*$ in Equation 2.9 is shown in Figure 7. This graph shows the subnetwork produced by the directed edges from miRNA vertices to genes, based on miRNA-gene targets. The number of miRNA targeting a given gene varies substantially, from as many as 143 miRNAs targeting PTEN to as few as to 0. These in-degrees from miRNA vertices are given in Table 2.
Figure 7: Topology of subnetwork of the integrated BRAF signaling pathway corresponding to the unweighted adjacency matrix $A_{mi}^*$ in Equation 2.9, namely, the graph formed by miRNA-gene targets. All edges are directed from the miRNA vertex to the gene vertex. The miRNA node colors, edge widths, and edge colors correspond to the degree of the miRNA vertex, that is, the number of genes in the BRAF pathway targeted by a given miRNA. Grey graph nodes are genes, while colored graph nodes are miRNAs, targeting 1 gene (red), 2 genes (yellow), 3 genes (green), 4 genes (blue), and 5 genes (purple). The genes BRAF and MAP2K2 are not targeted by any genes. Darker edge colors correspond to higher out-degree of the associated miRNA.
Table 2: In-degree by BRAF pathway gene, for edges into the gene from miRNAs. Values count the number of edges into a given gene from any miRNA.

Although most miRNAs in the TCGA dataset target a single gene in the BRAF pathway, 47% of the miRNAs target two or more genes. In the network diagram in Figure 7, miRNAs and their edges are colored according to the degree of the miRNA node, that is, the number of genes in the BRAF pathway targeted by the miRNA. Although there are 128 miRNAs that target only a single gene, the remaining 110 target multiple genes. This intricate structure is not balanced, in neither the in-degree of gene nodes nor the out-degree of miRNA nodes. The number of unique miRNAs that targets each gene varies from 58 (PTEN) to 4 (PIK3CA); and while one miRNA targets 5 genes, there are 19 that target 4, 31 that target 3, and 59 that target 2.

We obtained observations of gene expression, miRNA expression, methylation, and copy number from TCGA, as described in Section 1.2.1 and Section 2.2.2. We considered the subset of cancers for which all 10 BRAF pathway genes were observed, and for which more than 10 samples were available in both the tumor and normal tissue sample populations. This yielded 17 cancers for integrative pathway analysis of the BRAF pathway. The cancers and the sample sizes we used in the analysis are given in Table 3.
<table>
<thead>
<tr>
<th>Code</th>
<th>Cancer</th>
<th>Tumor</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLCA</td>
<td>Bladder urothelial carcinoma</td>
<td>412</td>
<td>36</td>
</tr>
<tr>
<td>BRCA</td>
<td>Breast invasive carcinoma</td>
<td>1096</td>
<td>161</td>
</tr>
<tr>
<td>CHOL</td>
<td>Cholangiocarcinoma</td>
<td>36</td>
<td>15</td>
</tr>
<tr>
<td>COAD</td>
<td>Colon adenocarcinoma</td>
<td>455</td>
<td>92</td>
</tr>
<tr>
<td>ESCA</td>
<td>Esophageal carcinoma</td>
<td>185</td>
<td>64</td>
</tr>
<tr>
<td>HNSC</td>
<td>Head and neck squamous cell carcinoma</td>
<td>528</td>
<td>82</td>
</tr>
<tr>
<td>KICH</td>
<td>Kidney chromophobe renal cell carcinoma</td>
<td>66</td>
<td>57</td>
</tr>
<tr>
<td>KIRC</td>
<td>Kidney renal clear cell carcinoma</td>
<td>534</td>
<td>427</td>
</tr>
<tr>
<td>KIRP</td>
<td>Kidney renal papillary cell carcinoma</td>
<td>291</td>
<td>87</td>
</tr>
<tr>
<td>LIHC</td>
<td>Liver hepatocellular carcinoma</td>
<td>377</td>
<td>87</td>
</tr>
<tr>
<td>LUAD</td>
<td>Lung adenocarcinoma</td>
<td>519</td>
<td>182</td>
</tr>
<tr>
<td>LUSC</td>
<td>Lung squamous cell carcinoma</td>
<td>504</td>
<td>242</td>
</tr>
<tr>
<td>PAAD</td>
<td>Pancreatic adenocarcinoma</td>
<td>185</td>
<td>37</td>
</tr>
<tr>
<td>READ</td>
<td>Rectum adenocarcinoma</td>
<td>167</td>
<td>17</td>
</tr>
<tr>
<td>STAD</td>
<td>Stomach adenocarcinoma</td>
<td>478</td>
<td>132</td>
</tr>
<tr>
<td>THCA</td>
<td>Thyroid carcinoma</td>
<td>507</td>
<td>99</td>
</tr>
<tr>
<td>UCEC</td>
<td>Uterine corpus endometrial carcinoma</td>
<td>545</td>
<td>51</td>
</tr>
</tbody>
</table>

Table 3: TCGA cancer types used for BRAF signaling pathway data analysis. Given are the TCGA cancer code abbreviation, cancer name, and sample population sizes for tumor and normal tissue samples. Only TCGA cancers with gene expression observations for all 10 BRAF pathway genes, as well as more than 10 samples in each population, were considered.
Within each cancer, we performed the same four NetGSA-based pathway analyses as in the simulation study: full integration of gene expression, miRNA expression, and gene-level methylation and copy number (miEMC-NetGSA); integration of gene and miRNA expression (“miE”); integration of gene expression, methylation, and copy number (EMC-NetGSA); and the original NetGSA for expression only (“E”). To correct for the multiple comparison problem, we adjusted $p$-values within each cancer using the method of Benjamini and Hochberg [1995] (BH).

The adjusted $p$-values are plotted in Figure 8. The pathway is significantly disturbed at the $\alpha = 0.05$ level across all pathway methods in all cancers, with the exceptions of (1) pancreatic adenocarcinoma (PAAD), which is not significantly disturbed under the miE analysis, and (2) rectum adenocarcinoma (READ), which is not found significantly disturbed using any method.
Figure 8: Results of significance tests for pathway disturbance in the BRAF pathway, across 17 TCGA cancers. Barplots give values of $-\log_{10} p$-values, after Benjamini-Hochberg adjustment for multiple comparisons. Full gene expression miRNA expression, and gene-level methylation and copy number is shown in green (“miEMC”); integration of gene and miRNA expression is in yellow (“miE”); integration of gene expression and gene-level methylation and copy number is in blue (“EMC”); and the original, expression-only NetGSA results are in white (“E”). The horizontal dotted line gives the significance threshold at the $\alpha = 0.05$ level.

Although each of the methods reaches the same conclusion, using a binary decision rule based on the $p$-value, in terms of the magnitude of the $p$-value the relative degree of significance varies greatly by cancer. We observe that in some cancers, integration of any features other than gene expression increased significance, and additional integrative features do not contribute further. These cancers include bladder urothelial carcinoma (BLCA); colon adenocarcinoma (COAD); esophageal carcinoma (ESCA); head and neck squamous cell carcinoma (HNSC); thyroid carcinoma (THCA); and uterine corpus endometrial carcinoma (UCEC). Figure 9 shows barplots of the test statistics corresponding to the $p$-values. We observe that in these 6 cancers with comparable significance, the test statistics retain the same sign and approximately
the same magnitude.

Figure 9: Test statistics for pathway disturbance in 17 TCGA cancers, for the BRAF pathway. Barplot heights give the test statistic value. miEMC-NetGSA is in green, miE-NetGSA in yellow, EMC-NetGSA in blue, and NetGSA (“E”) in white.

On the other hand, for some cancers, a small \( p \)-value in the analysis of expression only increases substantially with the integration of gene-level methylation and copy number, but we do not see a corresponding increase in significance with the integration of miRNA expression with gene expression, and in fact, we observe a decrease in significance relative to the baseline NetGSA. Furthermore, in these situations, the full integration of gene expression, miRNA expression, and gene methylation and copy number generally has comparable significance to the original NetGSA. The cancers that exhibit this pattern include cholangiocarcinoma (CHOL), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and pancreatic adenocarcinoma (PAAD). In lung adenocarcinoma, the miEMC-NetGSA method has the lowest significance of the four methods, whereas in the other three cancers types, the new method has significance comparable to NetGSA, if marginally higher. We observe significant downward pressure on the test statistics for these cancers after introduction of miRNA expression, even
causing a change in the test statistic’s sign in lung squamous cell carcinoma and pancreatic adenocarcinoma.

Finally, we observe in the remaining six cancers—breast invasive carcinoma (BRCA), the three kidney cancers (KICH, KIRC, KIRP), liver hepatocellular carcinoma (LIHC), and stomach adenocarcinoma (STAD)—a more variable pattern. For the breast cancer, we observe increased significance after integrating miRNAs, although the test statistics share the same sign and remain of comparable magnitude across the methods. In kidney chromophobe renal cell carcinoma (KICH), the test statistic changes sign after introduction of the miRNA features, but retains moderate to strong significance in the integrative methods. To different degrees, kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), liver cancer, and stomach cancer each display reduction in the test statistic’s magnitude after miRNA integration, although they retain the same sign. Nevertheless, we observe in three (KIRC, KIRP, LIHC) that the significance strengthens after the expansion of the network topology to include miRNAs.

These results indicate statistically meaningful contributions of the miRNA features to the pathway analysis, in a non-deterministic way. In some cancers, miRNA integration reinforces the conclusions of pathway analysis using existing methods. In others, the miRNAs identify a significant disturbance that is less apparent when the BRAF pathway is considered using only gene-level features. This may be due to the reduction in noise at the level of gene expression features using the augmented network, thereby accentuating the differential expression in the pathway. Finally, in the other cancers, the reduction of gene-level noise that is accomplished by the miRNA-gene target network effect clarifies the expression of pathway genes. Consequently,
although we observe apparent significance in the gene-level analyses of these cancers, it is not so much a true differential signal as it is noise attributable to common, non-genetic biological drivers.

2.3 Discussion

Based on the simple principle of the weighted adjacency matrix and its transformation into the influence matrix, we were able to not only model signaling pathways in terms of gene expression, but also to integrate multiple data types at several different levels. We found that graphical models provide flexibility in representing the relationships between various -omics data types. Graphical models provide a useful method to relate epigenetic and transcriptomic measurements to genetic observations, and this is readily accomplished by the EMC-NetGSA framework. Moreover, the directedness of graph edges provides a convenient mechanism to integrate data unidirectionally: there are known impacts of methylation and copy number variation on gene expression, but it is not nearly as conclusive that the reverse also holds true. Directed edges from supplementary data vertices to expression vertices permit this functional interpretation, without exceeding known biological functions.

There is structural similarity between the miRNA and the EMC-NetGSA integration: in both, we model an external influence directed into a signaling pathway, permitting more nuanced examination of the underlying network of interest. But, miRNA integration occurs at a different level, and operates at a fundamentally distinct biological level from genes. Thus, despite the surface-level similarities between the two integrations, they operate under different principles, and correspond to different functional interpretations. A significant strength of
miRNA integration is its endogenization of extra-signaling pathway correlation between genes. That is, because individual miRNAs typically target several genes within a single signaling pathway, miRNA integration provides a mathematical mechanism by which co-expression of genes may be disentangled. This is accomplished in a unique manner, outside the graph topology of the pathway itself. Integration of methylation and copy number occur not only at the level of genes, but also entirely within a single gene: in the EMC-NetGSA model, these integrated vertices do not “cross-pollinate” activity to other genes, but rather clarify the nature of expression within a gene, conditional on its relationship to other genes. In contrast, the miRNA integration unites components of a signaling pathway in such a way that it gives a fuller, more holistic picture of genetic activity.

miRNA integration supplies a fuller picture of genomic activity, but that picture is far from complete. A shortcoming of these integrated pathway analyses is their lack of dynamic interactions between gene expression and the integrative graph components. Indeed, Blenkiron et al. [2007] discuss the possibility that gene expression may in turn have implications for production of miRNAs, and a similar argument could be made for the activity observed in methylation and copy number. The present analyses stop short of modeling these feedback loops. But, this is not because of a lack of modeling flexibility, and in fact, the pathway models here could easily be extended to include directed edges from gene expression vertices into the secondary data platform vertex sets. Rather, this lack is due to limitations in current understanding of the biological mechanisms that give rise to copy number variation and methylation, specifically any possible genetic cause. Nonetheless, these integrative pathway analyses anticipate the future opportunity to construct even more detailed models of cellular and genomic activity.
Chapter 3

Low-rank imputation of -omics missing data

The pathway database introduced in Section 1.2.2 specify functional relationships between genes, and correspond to a large number of -omics features in the TCGA dataset. However, it is the case that many individual observation vectors among the $N$ total samples in a given cancer population exhibit missing values.

In general, statistical methods for -omics data are based on the assumption of complete data that exhibit no missingness. In practice, datasets collected across dozens of clinical locations on thousands of genomic features contain substantial missing data. This limits the viability of many statistical methods, a problem that will only worsen as integrative studies become more common. Some work has begun to look at imputation of integrative datasets. Fryett et al. [2018] reviewed imputation methods for transcriptome data, highlighting in particular the FUSION model [Gusev et al., 2016]. That method integrates genotype and expression, followed by downstream analysis to identify phenotypic drivers. Several imputation schemes were also considered, including a variant of nearest neighbors imputation, linear predictors, and the Bayesian linear mixed models introduced by Zhou et al. [2013]. Gamazon et al. [2015]
employed a penalized regression framework to obtain similar linear predictions.

Schulz et al. [2017] performed integrative analysis of methylation and expression data in the human brain. For expression imputation, they use the IMPUTE model. IMPUTE [Howie et al., 2009] uses hidden Markov models for simulation-based imputation of missing data. Chudasama et al. [2018] also applied IMPUTE for imputation, and performed an integrative analysis of expression and transcriptome observations in cancer.

Work on imputation for genomic pathway analysis is in the early stages. Some authors simply introduce methodological adjustments to work around missing data. Zhao et al. [2017] scored pathways and corrected for a deterministic impact of smaller sample sizes due to missing data on the pathway rank. Likewise, Köksal et al. [2018] appealed to parsimony by assuming that missing values have an insignificant effect on the overall analysis.


Dimension-reduction is also applied for imputation. Some methods based on the singular value decomposition expression imputation [Troyanskaya et al., 2001]. Mazumder et al. [2010] iteratively imputed missing data with a soft-thresholding algorithm. Tsuchiya et al. [2017] used reduced-rank methods to impute values in unequally-spaced gene expression time series.
As formulated in Chapter 2, the integrated NetGSA-style pathway analyses cannot accommodate any missing values. In the case of -omics features—that is, methylation and copy number observations—that are missing, we may omit features that are entirely missing. However, for -omics features that are not missing from every sample, we may encounter a decision between omitting samples to maximize the number of features in the analysis, or omitting features to maximize the number of samples. The first option ignores substantial information that may be present in the non-missing elements of samples that are missing some features. The second option ignores substantial information that may be present in the features we omit.

The choice of the best option is at best ad-hoc, and at worst may result in significant decreases in statistical power. To remedy persistent missing values in -omics datasets, we employ imputation methods that leverage the low-rank nature of the full data matrix. We first give an overview of an existing imputation method, and detail potential shortcomings of the method. We then propose an iterative procedure that adapts the existing method. The new procedure subsumes the existing method as a special case, and is applicable in a wider range of scenarios, including real-world data analyses. We then perform real-world data analysis for several integrative models, using the new imputation procedure.

Our approach enables analysis of a broader range of pathways than is possible without imputation. This allows analysis of biological processes using all data available, not just data that is well-formatted. In turn, this provides the opportunity to examine the organic systems that underlie complex diseases in finer detail and greater depth. We demonstrate both these perspectives with our data analysis: large-scale discovery among many pathways and granular analysis of specific pathways.
3.0.1 Structured matrix completion

Cai et al. [2016] proposed structured matrix completion (SMC) to address block-missing data. This type of missing data arises in the TCGA dataset when a specific -omics platform was collected at only a subset of clinical sites. The method takes advantage of the generally low rank of the -omics data matrix, to linearly impute reasonable values for missing observations.

Denote by $q$ the number of features in an -omics dataset of interest. For gene-level integrative pathway analysis using EMC-NetGSA, this is all expression, methylation, and copy number features, noted in Section 1.2.1 to be $q = 6973$ for the TCGA dataset. However, we may just as easily apply the method to only gene expression, or also include miRNA observations.

We format the -omics data as a $q \times N$ matrix $X = (x_1, \ldots, x_N)$, where $N$ is the total number of subjects. We denote by $q_E$ the number of genes with observed mRNA expression in the dataset; $q_C$ the number of genes with copy number observed; and $q_M$ the methylation features. In principle, a full dataset for the NCI PID would contain $q_E = q_C = q_M = 2,393$ features, for a total of 7,179 features. However, in practice, we have $q_E \neq q_C \neq q_M$.

The data exhibit several patterns of missingness, particularly entire -omics platforms missing in all samples collected at specific sites, and individual -omics features with missing values that vary according to subject, typically missing because of data quality problems.

Structured matrix completion (SMC) addresses the first of these scenarios. The method builds on results for recovery of missing-at-random elements of low-rank matrices, for example, Candès and Tao [2010]. The authors consider matrix $X$ that is approximately rank $r$, in the sense that its $r^{\text{th}}$-largest singular value is much larger than the next; contains continuous elements; and the matrix of deviations $(X - X_r)$ is well-conditioned. Here, $X_r$ is the rank-$r$
singular value decomposition (SVD) approximation to $X$ constructed using the first $r$ singular dimensions.

Further, suppose $X$ is a block matrix:

$$X = \begin{pmatrix} X_1 & X_2 \\ X_3 & X_4 \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}$$  \hspace{1cm} (3.1)$$

where $X_{22}$ is entirely missing. We approximate $X_{22}$ using the SVD. If $X$ were exactly rank $r$, one could exactly recover $X_{22}$ with the component $X_{21}X_{11}X_{12}$ of the Schur complement of $X$, where $X_{11}$ is the Moore-Penrose pseudo-inverse. For known $r$ but $X$ approximately rank $r$, we may recover $X_{22}$ using the rank-$r$ SVD approximation to $X_{11}, X_{12}, X_{21}$. Finally, we choose $\hat{r}$ as the largest value of $r$ for which the SVD approximation to $X_{11}$ is non-singular, and the approximation to $X_{22}$ is well-conditioned.

This block approach works when the missing data in $X$ has clean margins, i.e., the same features are either all missing or all observed in all samples. However, this is not typically true of real-world datasets, in which some subjects are naturally missing a small number of non-overlapping features. A direct solution is to form $X_{22}$ as be the minimal covering submatrix for all missing values. When the number of observed values in $X_{22}$ is small, we may justify discarding small portions of data prior to imputation, especially when the entire dataset at large is of interest, as opposed to individual features.

In real-world data matrices, however, the missingness in $X_{22}$ is sparse. In our TCGA dataset, for example, less than 5% of its elements are missing. The dimensions of $X_{22}$ must extend to cover all features with missing values, even those that are observed in most samples.
In the pathway analysis framework, \( X \) is a composite matrix of data generated from multiple signaling pathways. It is thus important to maximize the amount of information we use when imputing each given feature, so that we do not discard information of importance to a specific subset of pathway features of interest.

Also of critical importance is the assumption that \( X_{12} \) is non-degenerate, i.e., has a positive (nonzero) number of rows. This assumption requires the presence of at least some features observed in all samples, which is not generally guaranteed in -omics studies. We demonstrate a toy example, below, that illustrates this case.

### 3.1 Integrated iterative imputation

To address these limitations of SMC, we perform an iterative imputation independently across each sample. In the notation of (3.1), and without loss of generality, \( X_1 \) is \( q \times N_1 \), \( N_1 < N \), and contains no missing values; and \( X_2 \) is \( q \times N_2 \). Here, \( N_1 \) is the number of samples with complete observations for all \( q \) genomic features, while \( N_2 \) is the number of samples with any missing values, \( N_1 + N_2 = N \). Suppose further that \( q_1 \times N_2 \) matrix \( X_{12} \) contains no missing values, and \( X_{22} \) is the \( q_2 \times N_2 \) minimal covering submatrix for all missing values in \( X \), \( q_1 + q_2 = q \). \( q_1 \geq 0 \) is the number of features with no missing values, and may equal to zero, in which case every row of matrix \( X_2 \) contains at least one missing value, so that \( X_2 \equiv X_{22} \).

We may then consider individual columns \( x_i \) of \( X_2 \), \( i = N_1 + 1, \ldots, N \), each of which is a sample with complete observations on at least \( q_1 \) features. Of the elements of \( x_i \), we denote by \( q_i \leq q_2 \) the number of missing features.

If most of the values of \( X_{22} \) were missing, we might apply SMC directly, replacing the
imputed values with observed elements of $X_{22}$, where available. However, $X_{22}$ is often quite dense: in our data, more than 95% of the entries $X_2$ are observed, so imputing all elements of $X_{22}$ with SMC risks losing possibly considerable structural information in its many nonzero elements.

We apply SMC iteratively and independently to each of the samples in columns of $X_2$. Sample $x_i$ is a vector of $q$ elements, $i = N_1 + 1, \ldots, N$. By invariance of the singular values of $X$ under row and column permutations, we may suppose without loss of generality that the first $(q - q_i)$ elements of $x_i$ are entirely observed, so that the matrix $\begin{pmatrix} X_1 & x_i \end{pmatrix}$ is entirely observed except for a missing $q_i \times 1$ submatrix in its lower-right corner.

Turning our attention to the remaining columns of $X_2$, we identify all $0 \leq k_i \leq N_2 - 1$ other $x_{i'}$ with complete observed values for the $(q - q_i)$ features observed in $x_i$. Denoting the index set for these vectors by $\{\omega_{in}\}_{n=1}^{k_i}$, we form the matrix

$$\chi_i = \begin{pmatrix} X_1 & x_i & x_{\omega_{i1}} & \ldots & x_{\omega_{ik_i}} \end{pmatrix} \quad (3.2)$$

$\chi_i$ is $q \times (N_1 + k_i + 1)$ matrix, and we treat the entire lower-right $q_i \times (k_i + 1)$ submatrix as missing. The first column of that submatrix, which corresponds to the missing elements in $x_i$, is entirely missing. Subsequent columns come from $X_2$ and are missing a subset, possibly improper, of the same missing values in $x_i$. We impute this corner matrix using SMC and use the first column to form imputed vector $\tilde{x}_i$.

We repeat this procedure for all $i = 1, \ldots, N_2$, at the end of which, we have an imputed matrix $\tilde{X}_2 = \begin{pmatrix} \tilde{x}_1 & \ldots & \tilde{x}_N \end{pmatrix}$ which contains no missing values. In turn, this yields an imputed
matrix $\tilde{X} = \begin{pmatrix} X_1 & X_2 \end{pmatrix}$.

As a toy example to demonstrate the procedure, we consider a matrix with elements that equal to 1, or are missing (denoted “-”):

$$\begin{pmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & - & - \\ 1 & 1 & - & - & - \\ 1 & 1 & - & - & - \end{pmatrix}$$ (3.3)

We observe that, as originally proposed, SMC is unable to directly impute any of the missing values in $x_3$, $x_4$, $x_5$, because none of the rows of $X$ is completely observed.

Figure 10 displays the configuration of $X$, as well as the corresponding $\chi_i$, $i = 3, 4, 5$. $x_5$ contains a single missing value, which is a subset of the missing elements in $x_3$, whereas $x_4$ is missing that element. So, $k_3 = 1$, and we impute the missing values in $x_3$, using the element in $x_5$ corresponding to the observed value in $x_3$; but we ignore the observed element in $x_5$ that is missing in $x_3$ and $x_4$, so $k_5 = 0$. Implicitly, we impute this value of $x_5$, as well, but we discard this value, and keep only the imputed values for $x_3$. A symmetric argument applies to $x_4$.

On the other hand, because both $x_3$ and $x_4$ are missing elements that are observed in $x_5$, we cannot use either of these in imputing the missing value in $x_5$. 
Figure 10: Iterative integrated imputation with a toy example. $\mathbf{X}$ is a data matrix of 5 observation vectors with missing values that do not form a rectangle. Iteratively for each column with missing elements $x_i$ ($i = 3, 4, 5$), we apply structured matrix completion to the matrix $\chi_i$, consisting of all complete data vectors; $\chi_i$; and any columns with missing values that are a subset of those missing from $x_i$. Imputation is performed on the entire rectangular submatrix of all features missing in $x_i$, with any observed elements not from $x_i$ treated as missing. So, imputing $x_3$ and $x_4$ uses $x_5$, whereas $x_5$ is imputed using only the complete data vectors.

An appealing property of our iterative imputation procedure arises when the missing data has a block structure, as in the original SMC setting. In that case, iterative imputation produces identical imputed values to those from SMC. Moreover, as demonstrated in our toy example, we may apply our procedure to matrices for which $q_1 = 0$, that is, no features are fully observed across all subjects. In contrast, SMC cannot be used to impute that type of matrix.
3.2 Results

Applied to large-scale datasets, iterative imputation produces a complete matrix $\hat{X}$, using which we perform downstream analysis of pathway disturbance. We explore the power of our two-stage imputation and pathway analysis using a simulation. Then, we apply the method to breast cancer data to analyze pathway disturbances by tumor subtype.

3.2.1 Simulations

We follow a similar simulation design to that of Zhang et al. [2017a]. For the treatment group, we construct a five-level binary tree. The control group network derives from the treatment network, with all edges in the left branch (including the root) removed. Correlation between expression vertices is set to 0.8 in the tree’s top third (two levels); the middle third (third level) has association 0.5; and the lower third (last level) has association 0.2. We also add vertices for methylation and copy number, integrated via directed edges to the expression vertex. We set correlation of expression with copy number to 0.5, and correlation with methylation to -0.25.

We generate integrated data vectors $y_i$, $i = 1, \ldots, N$, from the EMC-NetGSA model in Equation 2.5, where $N = N_T + N_C$. $N_T$ ($N_C$) is the number of treatment (control) samples. We set $\sigma_\gamma^2 = 5$, $\sigma_\epsilon^2 = 0.5$, and $N_C = 50$ and $N_T = 150$. We consider four gene sets:

1. All genes in the network;
2. Top one-third levels of the tree;
3. First two-thirds levels of the tree;
4. The last level of the tree.
Denote by $\beta_{C1}, \beta_{C2}, \beta_{C3}$, the mean vectors of gene expression, copy number, and methylation, respectively, in the control population. The treatment population is defined analogously. We simulate two scenarios for the network-adjusted mean coefficient $\beta$:

1. $\beta_{T1} = \beta_{C1} = \beta_{T2} = \beta_{C2} = \beta_{T3} = \beta_{C3} = 0$;

2. $\beta_{T1} = 0.251, \beta_{T2} = 11, \beta_{T3} = 0.51$ for top two-thirds levels, otherwise same as first scenario.

The TCGA dataset displays missingness of two types. The first is block-wise missing data in samples collected at specific research sites for entire classes of measurements. This type of missing data motivated the original SMC method, and as noted previously, our iterative imputation is identical to SMC in the special case that the missingness in $X$ forms a block.

The second type of missingness varies the specific features that are missing within any single sample. In this case, the SMC procedure cannot simultaneously impute all missing values and use all observed data points. We note that the missingness does not alter the low-rank structure of the matrix. Therefore, we designed our simulation to mimic this variety of missingness, namely, where the features with missing values vary by sample, but we also include block-wise missingness, as well. This allows us to compare the novel aspect of our iterative imputation with the performance of the original formulation of SMC.

Within each simulation replicate, we generated a full data matrix, $X$. We randomly selected subsets of features and subjects to exhibit missing data, giving a data matrix with missing values. Additionally, we removed a full rectangular submatrix, which reflects the composition of missing data that is sometimes observed. We removed all samples with missing observations to form $\tilde{X}$. Finally, we iteratively imputed the missing values to obtain $\hat{X}$. For comparison, we
also imputed the matrix using SMC on the covering submatrix $X_{22}$, as well as the $K$-nearest neighbors (KNN) method taking median among $K = 10$ nearest neighbors, as implemented in \textit{bnstruct} [Franzin et al., 2017].

We performed 1000 simulation replicates for each mean scenario. We performed NetGSA on all five data matrices for both mean scenarios, and we calculated the power for each test by checking the significance of the Benjamini-Hochberg (BH) adjusted $p$-value [Benjamini and Hochberg, 1995], at the $\alpha = 0.05$ level. The power is then the proportion of simulations in which the null hypothesis of no difference is rejected.

Our combination of mean scenarios and gene sets provides simulations in which all genes in a gene set are differential; some (but not all) genes are differential; and none of the genes are differential. Moreover, our simulation scenario includes missing data that is characteristic of the real data, with a composition of data that is missing in block form, as well as data that is missing on a by-feature basis within individual samples. This simulation setting more accurately characterizes the real-world missing data, but this type of by-feature missingness was not considered by the authors of SMC.

The left panel of Figure 11 gives boxplots of $-\log_{10}(p)$ for the $p$-values in the first mean scenario, in which the expression coefficients are equal across the two populations. The power of the pathway analysis that uses our imputed matrix $\hat{X}$ is comparable to that of the true data. Our procedure corresponds to a minor increase in the type I error rate over the true data. In comparison, direct application of SMC to the minimal covering submatrix for the missing values results in a higher false positive rate.

The right panel of Figure 11 displays simulation significance results from the second mean
Figure 11: Boxplots of $-\log_{10}(p)$-values for simulation study tests of pathway disturbance. Network structure is a binary tree. $\mathbf{X}$ (shown in blue) represents EMC-NetGSA applied to the complete data matrix, with no missing values. $\tilde{\mathbf{X}}$ is the matrix with all samples with any missing values dropped. $\mathbf{I}_3$ (in green) refers to our integrated, iteratively imputed matrix. SMC uses imputed values obtained by treating the entire minimal covering submatrix as missing, and KNN is $K = 10$-nearest neighbors imputation. Horizontal line gives $-\log_{10}(0.05)$. Scenário 1 (left): No features differential. Scenário 2 (right): Top 2/3 features differential.
scenario, in which the top 2/3 levels of nodes in the binary tree are differentially expressed. Our iterative integrated imputation procedure exhibits power comparable to that of the true data, and consistently dominates use of the truncated data, $\tilde{X}$. Application of SMC to a rectangular submatrix is similar, but with a higher rate of false positives. In turn, this results in power that is higher than the true data, because of the inflated propensity to reject the null hypothesis.

KNN is comparable to our iterative imputation, but with weaker power. This relative performance of KNN to our method is robust to a range of $K$, both large and small. A benefit of our method is that the imputed values do not change the outcome of inference, as compared with the true data. The KNN imputation uses real observed values from other observations that are “close.” In contrast, the SVD-based methods construct a (linear) functional model for the data, and can thus directly predict the missing values.

The direct application of SMC to the minimal covering submatrix of the missing data in $X$ is similar to our iterative method, but with a higher false positive rate. The elevated false discovery rate reflects that SMC accentuates statistical noise, thereby reinforcing and strengthening spurious deviations due to statistical variation. SMC propagates dominant low-rank structure in $X$. But, discarding the information contained in samples with missing values results in an over-emphasis of features that are not reflective of the overall data matrix. This occurs even when imputing across a data matrix for which the mean structure is the same across all subjects, as in the first mean scenario.

Our imputation leverages the imputation using linear dependence due to the underlying
SMC method, while also maximizing the available information about each sample. This provides a more nuanced and complete picture of the linear structure of the network data. For pathway analysis, it is particularly important that any imputed values be locally accurate. The features in the TCGA data matrix \( X \), discussed below, encompass hundreds of different pathways, so the matrix-wide contamination we observed in the SMC simulations poses a serious obstacle to valid statistical inference.

Figure 12: Simulation results for missing data in a rectangular submatrix, using the second mean scenario in which the top 2/3 of binary tree nodes are differentially expressed. In this special case, our iterative imputation method (I3, in green) has identical power to that of the basic SMC method.

These results are robust to variations in the combination of block-wise and at-random missingness. The balance between these two types may vary by dataset: for instance, the original SMC method was designed for block-missing data from the TCGA study, but our data analysis indicates that at-random missing values are also widespread within individual subjects.
As the missingness progresses from at-random to block-wise, we find that the performance of SMC improves relative to the iterative imputation. Equivalent power is achieved when the missing data is fully rectangular, which reflects that our method gives the same result as SMC for this edge case. This power is shown in Figure 12. However, when the data composes both block-wise and at-random missingness, our method offers improved control of false discoveries, without a cost in terms of power.

Finally, we note that real-world data matrices, as in the TCGA dataset we analyze, the features in $X$ may be drawn from many different pathways, some of which may exhibit disturbance in tumor samples, but others of which may have equal means in the sample populations. This is reflected in the fourth test set in the second simulation scenario. The genes of interest in the test set are not differentially expressed, although other features in the data matrix do feature different means. Our iterative method identifies the differential structure in the regions of the data matrix, without a corresponding increase in false positives. If the minimal covering submatrix for the missing values contains all features that are differentially expressed, in the samples with the pathway disturbance, the original SMC method will fail to reveal any pathway disturbance. On the other hand, our method will use the partial information available in each sample, thereby retaining the differential structure in the imputed values it returns.

3.2.2 EMC-NetGSA

Breast cancer subtype analysis

To demonstrate the value of the imputed data in pathway analysis, compared with the truncated data, we performed an analysis of pathway disturbances in breast tumor subtypes.
Our TCGA dataset contained 838 tissue samples. After downloading the NCI Pathway Interaction Database (PID) using the graphite software [Sales et al., 2018], which spans 2,393 genes across 212 pathways, the matrix $X$ had $q = 6,973$ rows (genomic features), observed on $N = 838$ tissue samples.

It might be possible to impute the tumor and healthy sample populations as separate matrices. However, a major limitation of the TCGA datasets is imbalanced sample sizes, relative to the large number of features. This could result in sensitivity of the imputed values in the control population, and may be unnecessary. In principle, healthy and tumorous tissues should share many structural characteristics, namely, all non-malignant signaling activity. Therefore, we impute both populations jointly.

Of the 838 samples, $N_2 = 719$ samples contained missing values (681 tumor, 38 healthy), spread across $q_2 = 171$ features. 157 features measured copy number, and the other 14 measured methylation. The values $k_i$, giving the number of additional columns of $X_2$ used in forming $\chi_i$, were generally quite high: only 6 columns used fewer than 50 other samples for imputation, or less than 1% of the total number of columns we imputed. Our imputation procedure uses a substantial amount of the non-missing information contained in $X_{22}$ when forming $\chi_i$, despite our separate imputation of each column of $X_2$.

A key assumption of the SMC imputation method is that the data matrix is approximately low rank. Conceptually, the data satisfy the low-rank assumption: within each population, conditional on the subject-level random effects, the mean matrix is simply noisy observations of a rank 1 matrix. We verified empirically the low-rank structure of the submatrix $X_1$, i.e., all rows with no missing values.
Breast tumors are known to stratify into well-differentiated subtypes [Eroles et al., 2012]. In particular, immunohistochemical (IHC) characteristics of estrogen receptor (ER), progesterone receptor (PR), HER2, and Ki67 are used to define intrinsic molecular subtypes [Dai et al., 2015]. Classification of tumors by subtype is useful because of substantial variation in prognosis and behavior, with corresponding implications for survival outcomes and treatments of the various subtypes [Feng et al., 2018]. The subtypes are strongly associated with tumor grade, clinical outcome, and overall subtype prevalence. It is therefore critical to understand the distinctions between separate tumor types, which may operate under substantially different biological mechanisms.

To determine the breast tumor intrinsic subtype, we followed the classification given in Dai et al. [2015]. These subtypes are defined according to immunohistochemistry (IHC) status of several specific genes: ER, PR, HER2, and KI67. The TCGA dataset contains the IHC status for ER, PR, and HER2, but not KI67, so we employ four tumor subtypes: luminal A / HER2- luminal B; HER2+ luminal B; HER2 over-expression; and basal, or triple negative. The specific immunohistochemical criteria used is given in Table 4. We categorized tumors described as “normal-like” as luminal A / HER2- luminal B, since they cannot be distinguished on the basis of ER, PR, and HER2 IHC status. Also, we note that in the literature, it is also conventional to refer to basal-like tumors as triple negative, based on their molecular signatures with respect to ER, PR, and HER2.

Although we were unable to distinguish tumors on the basis of KI67, this reflected practical considerations that apply beyond the current study. Indeed, Cheang et al. [2009] note that “Ki67 is not included in routine clinical decision-making because of a lack of clarity regarding
Table 4: Breast tumor subtype definitions and prevalence in data from The Cancer Genome Atlas (TCGA). IHC status refers to immunohistochemical status for the three specified genes. Prevalence measures a proportion.

<table>
<thead>
<tr>
<th>Subtype</th>
<th>Luminal A / HER2- luminal B</th>
<th>HER2+ Luminal B</th>
<th>HER2 over-expression</th>
<th>Basal</th>
</tr>
</thead>
<tbody>
<tr>
<td>IHC status</td>
<td>ER+ or PR+, HER2-</td>
<td>ER+ or PR+, HER2+</td>
<td>ER-, PR-, HER2+</td>
<td>ER-, PR-, HER2-</td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.62</td>
<td>0.14</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>Tumor grade</td>
<td>2, 3</td>
<td>2, 3</td>
<td>3</td>
<td>Poor</td>
</tr>
<tr>
<td>Outcome</td>
<td>Good, Intermediate</td>
<td>Intermediate, Poor</td>
<td>Poor</td>
<td>Poor</td>
</tr>
<tr>
<td>Count</td>
<td>307</td>
<td>74</td>
<td>17</td>
<td>81</td>
</tr>
<tr>
<td>Prevalence</td>
<td>0.64</td>
<td>0.15</td>
<td>0.04</td>
<td>0.17</td>
</tr>
</tbody>
</table>

In the BRCA dataset, 479 samples had IHC status available for ER, PR, and HER. Frequency and prevalence of the subtypes is given in Table 4, and clearly matches that in Dai et al. [2015].

We ranked the pathways by the variance between the $-\log_{10}(p)$-values in the 4 subtypes. We considered the five pathways that exhibited the highest variability, which reflects major differences in the disturbance of these pathways between subtypes. These most-variable pathways were (1) EPHB forward signaling, (2) FOXA2 and FOXA3 transcription factor networks, (3) ErbB2/ErbB3 signaling events, (4) Validated nuclear estrogen receptor alpha network, and (5) E2F transcription factor network. The nuclear estrogen receptor network contains known oncogenes and plays an established role in breast cancer [Sommer and Fuqua, 2001], as does EPHB [Pasquale, 2010], [Kaenel et al., 2012]. But while the role of E2F is established in cancers such as retinal cancer [Nevins, 2001], its possible role in breast cancer has only been explored recently [Johnson et al., 2016]. This is also true of ErbB2/ErbB3 [Ma et al., 2014a], and while...
FOXA1 is implicated in breast cancer [Bochkis et al., 2012], this is not true of FOXA2 and FOXA3.

We identified the pathways for which the test for overall pathway disturbance changed significance between the raw and imputed datasets in at least one subtype. This resulted in ten pathways with different statistical conclusions between the raw and imputed datasets. Figure 13 shows the $-\log_{10}(p)$-values for the raw and imputed datasets in these pathways, across the different subtypes.

Changes in significance are in some cases minimal, and likely represent insubstantial statistical variation—for example, significance in any subtype is questionable in the HIV-1 and HDAC1-mediated signaling events. Likewise, $\beta_1$ integrin cell surface interactions appear significant only in luminal B subtypes, and $\beta_2$ integrin cell surface interactions is significant in HER2 over-expression subtypes. In these cases, the significance is strong in both the raw and
imputed datasets. Moreover, we verified the robustness of the pathways shown in Figure 13 with SMC on the entirety of $X_{11}$, as well as KNN imputation, and in both cases, the same pathways changed significance, with similar conclusions.

Of the pathways we tested, regulation of Smad2 and Smad3 signaling was identified as significantly disturbed across all subtypes, although only marginally so for luminal subtypes; and the pathway contains ESR1, an oncogene implicated in breast cancer. Smad2 and Smad3 play a well-established role in transcription growth factor $\beta$ (TGF-$\beta$; Brown et al., 2007), and we considered this regulatory pathway for further analysis.

Our data contained a total of $N = 479$ breast tumor samples with valid IHC observations. The Smad2 / Smad3 signaling pathway consists of 68 genes, comprising 200 -omics features after integration of copy number and methylation data. We performed NetGSA on the raw and imputed datasets, which contained 371 and 543 samples, respectively. Of the 172 samples with imputation, 119 were luminal A / HER2- luminal B subtype; 23 were luminal B; 6 were HER2 over-expression; and 24 were basal. We conducted significance tests for the full pathway, as well as each integrated triplet corresponding to one gene: three nodes, one for each integrated platform.

The network topology for the pathway consists of 423 directed edges between the expression nodes, as well as two directed edges for each expression node to integrate copy number and methylation, where available.

Figure 14 plots $p$-values for all breast cancer samples, as well as individual subtypes. Comparison is given between the raw and imputed datasets. Shown are the results of testing the
Figure 14: $-\log_{10}(p)$-values for significance tests of integrated genomics data for the entire Smad2 / Smad3 signaling pathway and selected genes, by subtype. Gene-level significance tests considered integrated triplets of pathway nodes, namely, expression, copy number, and methylation observations for each gene.

We note that, tested across all tumors, the pathway is only marginally significant. In contrast, analyzed at the subtype level, a clean partition is clear: Smad2 and Smad3 regulation is disturbed in HER2 over-expression and basal subtypes, with strong statistical significance, but not disturbed significantly in either luminal A or B subtypes. These results are consistent between the raw and imputed datasets, but exhibit stronger significance in the imputed dataset.

Considered more granularly, TG-interacting factor 1 (TGIF1) displays only marginal significance in HER2 over-expression subtypes, consistent with the low significance of TGIF2. TGIF2 is strongly significant in the other three subtypes, but in the raw data, TGIF1 is strongly significant only in basal subtypes. However, our imputed method identifies these as
Figure 15: Coefficient estimates for expression in the Smad2 / Smad3 signaling pathway, by subtype. Selected genes shown only. Coefficients were estimated separately by population (cancerous tumor and healthy control). Within the tumor population, separate estimates were obtained using the raw data with missing data removed, and using the imputed data.

significantly disturbed in luminal subtypes.

Our imputation method also reveals disturbance in Smad3 in luminal A / HER2- luminal B subtypes, and reinforces the lack of significance in basal tumors. The effect of Smad2 is reduced in luminal B tumors but increased in luminal A.

Across all subtypes, MED15 is not identified as disturbed, but within each subtype it is. This suggests a biological difference between the role this gene places in different subtypes. Figure 15 shows gene expression coefficient estimates $\hat{\beta}$ from the NetGSA analysis, and MED15 demonstrates down-regulation in HER2 over-expressed tumors, compared with up-regulation in basal tumors. Luminal B and basal tumors have comparable, mid-level expression in CEBPB, small in comparison with control; in HER2 over-expressed tumors, this value is under-expressed, whereas in luminal A tumors it is over-expressed.
HER2 over-expressed tumors exhibit expression in MYC in the opposite direction of the other subtypes, while the imputation actually causes a change in the sign of HDAC1 in luminal B tumors. Recent work found inconclusive results on the role of this gene in cancer tumors [Tang et al., 2015]. The sign switch is only in luminal B tumors, whereas in other tumors the difference is absent, or at least more muted.

Also of note is the HER2 over-expression subtype. The sample size for this subtype is already small, at 17, but with the raw data, only 11 samples are available for analysis of HER2 over-expression. We observed sign changes between the raw and imputed data among several genes in this subtype, including the gene CDKN1A, FOXH1, and MEF2C. The latter two genes are only marginally significant, especially compared with the very large p-values in other subtypes. But CDKN1A is strongly significant, and is known to relate to Smad signaling.

Smad3 is known to inhibit tumor cell growth [Zelivianski et al., 2010], but phosphorylation by CDK prevents this tumor-suppressive behavior [Liu, 2011]. Counteracting this effect, inhibiting CDK promotes transcription by Smad3, with the effect of suppressing cancer cell growth. This is speculated to reflect a non-canonical interaction between CDK and Smad3 that promotes tumor growth [Tarasewicz et al., 2014].

Addressing a lack of bio-marker targets for basal (triple-negative) breast tumors, recent research on this subtype has considered targeted suppression of CDK phosphorylation of Smad3. In particular, CDK has been observed to mediate a malignant interaction between Pin1 and Smad3 that increases proliferation of cancer cells in aggressive breast tumors [Thomas et al., 2017]. Targeted inhibition of CDK2 and CDK4 was prevented the Pin1-Smad3 interaction in basal subtypes, confirming the finding in [Tarasewicz et al., 2014] that suppression of CDK2/4
leads to increased Smad3 activity in basal subtypes.

Figure 16 shows expression coefficient estimates for the four tumor subtypes, for CDKN1A, CDKN2B, and Smad3. CDKN1A inhibits CDK2 phosphorylation of Smad3, while CDKN2B inhibits CDK4 [Driver et al., 2008]. In the healthy control tissue, CDKN1A expression is strong, whereas CDKN2B and Smad3 are not expressed at all.

We observe negative expression in Smad3 in the basal subtype, and the same in luminal A / HER2- luminal B tumors. This compares with slight positive expression of Smad3 in the other two subtypes. All four subtypes display positive expression for CDKN1A, which inhibits CDK2, however, the luminal and HER2 over-expression subtypes all display negative
expression for CDKN2B, with the strongest effect in luminal B tumors. This corresponds to small, positive expression in Smad3.

For comparison, we also analyzed the pathway using NetGSA on only expression data; with integration of only one of methylation and copy number at a time; and GSEA of expression [Wu et al., 2010]. The GSEA analysis identified only differential expression in HER2+ subtypes. The other three NetGSA methods were strongly significant in HER2+ and basal subtypes, as well as across all tumor types. However, while all four methods found marginal to no significance in luminal subtypes, the EC-NetGSA exhibited stronger significance. This supports the premise that Smad2 and Smad3 are central to transcription.

Taken as a whole, the results of our data analysis give strong support to the value of subtype analysis in complex diseases such as cancer. Among breast tumors, we identify Smad2 and Smad3 signaling pathway disturbance in basal and HER2 over-expression subtypes, with weaker disturbance in luminal subtypes. We identify separate roles of Smad3 in basal and HER2 over-expressed subtypes, suggesting the gene-level mechanism for the pathway disturbance may differ in the two subtypes.

**Pan-cancer analysis**

Following the EMC-NetGSA model of [Zhang et al., 2017a] discussed in greater detail in Section 2.1, we constructed for each signaling pathway in the PID a directed graph, with edges representing functional relationships connecting the corresponding vertices for gene expression. We also introduced graph vertices for gene-level methylation and copy number, with directed edges connecting each to the gene’s expression vertex. We estimated association weights for
the graph edges within each cancer, and applied the NetGSA hypothesis test [Shojaie and Michailidis, 2009] to test for differential activity. We tested the entire pathway for significance, as well as the subgraphs corresponding to the three vertices available for each gene.

We considered for pathway analysis the 22 cancers listed in Table 1 that are marked with a star (⋆). Of the 32 total cancers available from TCGA, these had more than 10 samples in both the tumor and normal populations. After removing the pathways that contained genes for which our data lacks observations of gene expression, we tested 173 of the PID pathways across the 22 paired populations. In order to correct for possible false discoveries due to the multiple testing problem, we used the procedure of [Benjamini and Hochberg, 1995] (BH) to adjust all p-values within each cancer.

Figure 17 shows the $-\log_{10}(p)$ transformation for the p-values in all pathways in all cancers. We observe that many p-values are strongly significant: even after BH-adjustment of the 3806 hypothesis tests we performed, only 88 pathways did not have a significant p-value at the $\alpha = 0.05$ level.
Figure 17: Shown is the $-\log_{10}(p)$ transformation of the BH-adjusted $p$-values obtained for 22 TCGA cancer types among 173 NCI PID pathways. Darker colors indicate stronger significance, in terms of a smaller $p$-value. A white value indicates the pathway was not tested for that cancer. We observe strong patterns of increased significance across all pathways within individual cancers, such as elevated significance in sarcoma (SARC) and thymoma (THYM) and reduced significance in colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ). We also observe patterns in significance within pathways, across cancers, such as elevated differential activity in B-cell receptor (BCR) signaling, and a lack of disturbance in the glypican 2 and 3 networks.

However, it is clear from the figure that there are associations among the $p$-values across all cancers in specific pathways. This includes consistent disturbance in the pathways B-cell receptor (BCR) signaling, ErbB1 downstream signaling, and transforming growth factor-beta (TGF-beta) receptor signaling, as well as consistent lack of disturbance, as in the glypican 2 and 3 networks, and PDGF receptor signaling. Likewise, we observe systematic elevation of $p$-values in some cancers, such as sarcoma (SARC) and thymoma (THYM), as well as depressed significance across other cancers, including colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ).
adenocarcinoma (READ).

The high significance in some pathways is in accordance with current scientific consensus. BCR signaling regulates B cell activity, and has additional downstream effects on other pathways that promote tumor growth [Fowler and Davis, 2013]. [Burger and Wiestner, 2018] discuss recent work to target malignancy in B cell signaling. ErbB1 plays a role related to epidermal growth factor (EGFR), and its disturbance is known to relate to a variety of cancers: [Roskoski Jr, 2014] found interactions in non-small cell lung cancer, colon cancer, and breast cancer. They also discussed ErbB1-related targets for clinical treatment of tumors. TGF-beta plays a central role in basic cellular activity [Jakowlew, 2006], and it can suppress or promote tumorigenesis, dependent on the cell type [Massagué, 2008]. [Fabregat et al., 2014] discussed treatments that target and inhibit malignant behavior in the TGF-beta pathway, in order to recover normal functioning.

Too, it is reasonable to observe co-disturbance of pathways within cancers. [Iengar, 2018] considered genes that consistently exhibit mutation across multiple cancer types, and identified a large number of pathways to which they relate. [Leiserson et al., 2013] directly considered the contrast between driver and passenger mutations in oncogenesis, and focus on finding multiple drivers that coexist within individual cancers.

Among the 3684 hypothesis tests shown in Figure 17, we considered a subset of 14 pairs of cancer and pathway that exhibited stronger significance than would appear typical for the combination of the two. The cancer-pathway pairs we considered are listed in Table 5. We considered 8 pathway disturbances in sarcoma, 3 in thymoma, and 1 pathway disturbance each
for kidney renal clear cell carcinoma, ovarian serous cystadenocarcinoma, and pancreatic ade-
nocarcinoma. In both sarcoma and thymoma, disturbance of T-cell receptor (TCR) signaling
in naïve CD4+ T cells was elevated, as was ErbB1 disturbance. The remaining 9 pathways
are found in only 1 cancer.
<table>
<thead>
<tr>
<th>Cancer</th>
<th>Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kidney renal clear cell carcinoma</td>
<td>Coregulation of androgen receptor activity</td>
</tr>
<tr>
<td>Ovarian serous cystadenocarcinoma</td>
<td>C-MYB transcription factor network</td>
</tr>
<tr>
<td>Pancreatic adenocarcinoma</td>
<td>Regulation of nuclear SMAD2/3 signaling</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>BCR signaling pathway</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>Beta1 integrin cell surface interactions</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>Ceramide signaling pathway</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>ErbB1 downstream signaling</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>Neurotrophic factor-mediated Trk receptor signaling</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>Signaling events mediated by focal adhesion kinase</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>TCR signaling in naïve CD4+ T cells</td>
</tr>
<tr>
<td>Sarcoma</td>
<td>TCR signaling in naïve CD8+ T cells</td>
</tr>
<tr>
<td>Thymoma</td>
<td>BCR signaling pathway</td>
</tr>
<tr>
<td>Thymoma</td>
<td>ErbB1 downstream signaling</td>
</tr>
<tr>
<td>Thymoma</td>
<td>TCR signaling in naïve CD4+ T cells</td>
</tr>
</tbody>
</table>

Table 5: Pairs of cancer and pathway for which the residual from a regression of the logit(p)-value was negative, and larger in magnitude than a Bonferroni-adjusted critical value. In sarcoma, we observe 8 disturbed pathways; in thymoma, we observe 2 disturbed pathways; and we observe 1 pathway disturbance in each of kidney renal clear cell carcinoma, ovarian serous cystadenocarcinoma, and pancreatic adenocarcinoma.
Some of the pathways in Table 5 reflect well-known drivers of the respective cancers. Differential expression in MYB in ovarian cancer has been suspected for a long time [Barletta et al., 1992], and some recent work has given attention to the role of C-MYB activation in other cancers [Jin et al., 2017]. The latter study found up-regulation in non-small cell lung cancer, a finding we corroborate—albeit to a lesser degree—in lung adenocarcinoma (LUAD), the corresponding TCGA cancer type. The prominent significance of ErbB1 is unsurprising in light of its known role in a wide range of cellular functions.

The significance of Trk signaling mediated by neurotrophic factor is perhaps surprising in sarcoma, a bone cancer. But while research on this type of neural signaling is relatively new, it does support the a tumorigenic role for Trk signaling. Increased expression in TrkB and TrkC has been found to correlate with tumor growth in brain tumors [Lawn et al., 2015]. The brain-derived neurotrophic factor (BDNF) was found to bind to the TrkB receptor, and elevated expression had strong downstream effects in other pathways [Meng et al., 2019]. A Phase I clinical trial was also performed for an inhibitor of Trk in several cancer types including sarcoma [Drilon et al., 2017]. Another study found that inhibition of Trk signaling corresponded to improved clinical outcomes in Ewing sarcoma [Heinen et al., 2016].

Increased adhesion of shed tumor cells in the presence of focal adhesion kinase (FAK) activation has been reported, and it is hypothesized that inhibition of FAK may yield superior patient outcomes in a variety of cancers, including sarcoma [Perry et al., 2010]. Others observed tumor suppression as a result of FAK inhibition [Crompton et al., 2013], and FAK inhibition-mediated signaling was found to complement and reinforce synergistically the therapeutic effects of Aurora kinase B inhibition in Ewing sarcoma [Wang et al., 2019b].
The role of androgen receptors has been considered in murine tissues, including kidney tissue, which exhibit a different biological effect from prostate and epididymis tissue [Pihlajamaa et al., 2014]. That effect suggests a novel, kidney-specific role for the androgen receptor pathway. Increased androgen receptor expression in kidney renal clear cell carcinoma was found to correspond to improved clinical outcomes [Foersch et al., 2017], which supports earlier findings that also found a possible tumor-suppressive role for androgen receptors via circadian regulation in the kidney [Zhao et al., 2016].

3.2.3 miRNA-EMC-NetGSA analysis of breast tumor subtypes

We also performed integrative analysis of the breast tumor subtypes from Section 3.2.2 using the miRNA pathway analysis framework presented in Section 2.2. Research on miRNA activity and disturbance in breast tumors by subtype has gained traction in recent research. Blenkiron et al. [2007] considered differential miRNA expression across subtypes, and identified clusters of miRNA with coordinated differential activity. Tsai et al. [2018] found differential miRNA activity among tumors, across patient ages. miRNAs can be used as a diagnostic tool to better understand the molecular processes in tumors: Søkilde et al. [2019] performed clustering on miRNA expression, which they found to emphasize and supplement more classical molecular subtypes. In particular, they found that miRNAs that corresponded to improved survival outcomes, a miRNA with high predictive power for HER2 expression, and improved discriminatory power among luminal subtypes. Kolesnikov et al. [2016] also found value to diagnosis of molecular intrinsic subtypes using miRNA expression as a proxy. And, given the regulatory and target relationships exhibited by miRNAs towards genes, it is also natural to consider
differential miRNA as it relates to gene activity. Oztemur Islakoglu et al. [2018] considered the relationship between miRNAs and genes constituting the PI3K-Akt and p53 pathways.

As a source of signaling pathway topologies, we used the PID pathways. For miRNA-gene target relationships, we used the mirDIP database. We used the breast invasive carcinoma dataset (BRCA), with iterative imputation performed on a matrix all-omics features that we consider in the analysis: gene expression, methylation, and copy number for all genes in the PID pathways, as well as miRNA expression features for all miRNAs targeting any gene in any pathway.

As mentioned above, the TCGA dataset contains IHC status for the three dominant biomarkers of subtypes, namely, ER, PR, and HER2. In total, the breast cancer dataset contain 1103 tumor samples, as well as 161 samples of healthy tissue. Of the tumor samples, 721 had valid IHC data. We assigned the label “other” to the remaining, unclassified samples. Table 6 gives the sample sizes and prevalence of the tumors. We note that, with the category for “other” removed, the sample sizes are comparable to those in the analysis of Section 3.2.2.
<table>
<thead>
<tr>
<th>Subtype</th>
<th>Sample size</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luminal A / HER2- luminal B</td>
<td>442</td>
<td>0.40</td>
</tr>
<tr>
<td>Luminal B</td>
<td>126</td>
<td>0.11</td>
</tr>
<tr>
<td>HER2 over-expression</td>
<td>37</td>
<td>0.03</td>
</tr>
<tr>
<td>Basal</td>
<td>116</td>
<td>0.11</td>
</tr>
<tr>
<td>Other</td>
<td>382</td>
<td>0.35</td>
</tr>
<tr>
<td>All tumors</td>
<td>1103</td>
<td></td>
</tr>
<tr>
<td>Healthy samples</td>
<td>161</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Breast tumor molecular subtype sample sizes and prevalence in the TCGA dataset. Also given are the overall sample size of all tumors and healthy breast tissue.

To assess the impact of miRNA integration in the pathway analysis, we considered the gene EPHA4, which is known to exhibit differential activity in several cancer types [Kou and Kandpal, 2018]. EPHA4 receptors are known to bind to ephrin B [Taylor et al., 2017]. We observed the ephrin B reverse signaling pathway to change significance in HER2 over-expressed tumors. Hachim et al. [2017] discussed the activation of the gene EPHA4 by different subtypes. They found upregulation of this gene in ER+ subtypes, that is, luminal A and B subtypes, and attribute to it transcription growth factor β-mediated cell migration. Based on this association between EPHA4 and ER positivity, we would not expect to observe significant disturbance in the EPHA forward signaling pathway for HER2 over-expressed subtypes. This pathway prominently features the EPHA4 gene.

Figure 18 gives $-\log_{10}(p)$-values for significance tests of differential activity in the EPHA
forward signaling pathway, across four subtypes and four integration schemes: miEMC is miRNA-EMC-NetGSA; miE is miRNA and gene expression only; EMC is EMC-NetGSA; and E is NetGSA on expression alone. Using the miRNA-EMC analysis, we observed strong significance in the luminal subtypes, and slightly less so in the basal (triple negative) subtype. As expected, HER2 over-expressed subtypes did not exhibit differential activity in the pathway. However, the HER2 over-expressed subtypes did exhibit differential activity in the analysis of gene-only features, and even in the miRNA-gene expression analysis.

Figure 18: $-\log_{10}(p)$-values for the EPHA forward signaling pathway. Axis labels indicate the integration scheme: miEMC is miRNA-EMC-NetGSA; miE is miRNA and expression only; EMC is EMC-NetGSA; and E is NetGSA on gene expression only.

Figure 19 displays the network-adjusted parameter estimates for the healthy and HER2 over-expression subtype tumors, under the four integration schemes. The healthy tissue parameters exhibit minor variability in magnitude across the analyses, but remain relatively stable. In contrast, the subtype parameter values change substantially with the miRNA integration, in both magnitude and sign. Integration of only miRNA with expression caused
substantial changes in all subtypes, although in the luminal and basal subtypes, the integration of EMC and miRNA clarifies the signs. On the other hand, the strong negative expression of EPHA4 in the HER2 over-expressed subtypes sustains the change in sign. We also note that the higher expression in the basal subtype is reflective of the findings of Hachim et al. [2017], who found EPHA4 most active in basal subtypes.

Figure 19: Network-adjusted expression parameters for EPHA4 in healthy tissue and tumor subtypes. Integration of miRNA data only substantially alters the parameter estimates for all subtypes, in terms of both magnitude and sign. But, joint integration of miRNA, methylation, and copy number data clarifies and highlights the difference between the ER- HER2 over-expressed subtype and the ER+ luminal subtypes, as well as the basal subtype.

As another example, we considered the B cell receptor (BCR) signaling pathway. Wang et al. [2019a] found that PTEN, a gene that features in this pathway, was targeted by miR-136, which correlated strongly to poor clinical outcomes in basal (triple negative) tumor subtypes. Qi et al. [2019] found the BCR signaling pathway to contribute strongly to basal subtype growth. Figure 20 shows the results of significance tests for the BCR pathway under the various
integration schemes. We observe a boosting effect due to miRNA integration in conjunction with methylation and copy number, which applies across all subtypes, and strengthens the evidence of pathway disturbance compared with only gene-level integration. In this setting, miRNA integration serves less to identify a novel disturbance, and more to clarify and reinforce the results of methylation and copy number integration. Moreover, this effect of miRNA integration in basal tumors is supported by the literature: Hua et al. [2013] found differential activity in the miRNAs miR-18a and miR-135b in basal subtypes, a finding which Kurozumi et al. [2017] confirmed, as well as differential activity in miR-93 and miR-155. All four of these miRNAs target genes in the BCR signaling pathway, which suggests a biological explanation for the strengthened significance.

Figure 20: $-\log_{10}(p)$-values for the BCR pathway. Integration of miRNA data alone captures only marginal significance in the basal subtype. But, integration of miRNA in conjunction with gene methylation and copy number boosts the signal in basal subtypes, as well as the others, compared with integration of methylation and copy number without miRNAs.

Finally, we also observe a super-additive effect of the two types of integration, of miRNAs
as well as methylation and copy number. The hypoxia-inducible factor-2α (HIF-2α) transcription factor network is understood to correlate to poor prognostic outcomes in breast tumors [Moreno Roig et al., 2018]. Although basal subtypes are associated with this low prognosis, Helczynska et al. [2008] found no specific correlation between the IHC status of ER or PR. Jarman et al. [2019] found that HER2 regulates HIF-2α.

In Figure 21, we plot $-\log_{10}(p)$-values for the HIF-2α transcription factor network. In the fully-integrated analysis, we observe strong significance across all subtypes. Of particular note is the prominence of the basal subtype: this is one of the most significantly disturbed pathways for the subtype in our study. Furthermore, the magnitude of the miRNA-EMC-NetGSA analysis far exceeds that of miRNA or gene-level analyses. This is also true of the luminal subtypes, but for HER2 over-expressed tumors, it is much less so. The miRNA and gene-level integrations complement each other and provide strong support for this pathway disturbance, despite a complete lack of significance in any subtype for the expression-only analysis.

![Figure 21: HIF-2α transcription factor network](image)

Figure 21: HIF-2α transcription factor network
3.3 Discussion

Although most analytic techniques for genomic Big Data depend on the assumption of complete data, i.e., data without missing values, it is rarely the case in practice that a large-scale -omics dataset will exhibit high-quality observations across every feature in every sample. Far from a shortcoming or failure for individual studies, this is a characteristic of real-world datasets that requires sustained attention in order to enable analyses that do not simply discard data. Low-rank methods are appealing in the context of our pathway analysis because the mean structure of the NetGSA model is low-rank by definition. Moreover, these methods employ structural representations of the overall data matrix, a procedure that again aligns with the linear nature of the pathway analysis.

Despite the conceptual appeal of SMC, we found that even this method for imputation of missing data rested upon an idealized conception of the character of missing values. Much as statistical methods require complete data, SMC makes the unrealistic assumption that the data matrix can be rearranged into a rectangular matrix. We found that in the TCGA dataset, the pattern of missing values did not permit any such reordering, and use of the minimal covering submatrix for all missing values was a dissatisfying solution because of the density of observed values within that submatrix.

Our solution to this lack of rectangular missing data was less a new method, and more a prescription for careful application of SMC. As noted above, our iterative application of SMC recovers an identical imputed submatrix in the case of a missing submatrix, while permitting application to a far broader range of datasets. More important than that edge case is the more plausible setting in which every -omics features contains at least one missing value in some
sample, despite global density of observations in the data matrix. In these settings, even naïve application of SMC to the minimal covering submatrix is impossible, because this submatrix would have full missing columns. In contrast, our applied procedure permits imputation in these situations.

A shortcoming of our method is best illustrated by the edge case where we reproduce the exact SMC solution: although the values are the same, the computational time to obtain those values will be substantially longer, because we blindly perform a new imputation for each sample that exhibits missing values. In general, we do not address composite situations with some rectangular missing data in addition to more at-random patterns. In these cases, computational time can be reduced substantially, by removing redundant iterations.

An innovative characteristic of our simulation study was to consider the downstream effect of our imputation on pathway analysis that employs the imputed data. Although this style of “full-stack” assessment is not strictly a joint analysis, in the sense that the imputation and pathway analysis represent a single analytic step, we are able to assess the practical value of our method. In particular, we found that our procedure imputed values that were interior to some degree, because they did not alter the statistical decision, and particularly did not increase the rate of false positives. More direct power analyses typically assess the imputation’s accuracy in terms of a loss function based on the magnitude of deviations. In some instances, this may even be the same loss against which the imputation is performed as a minimization problem. Rather, we consider a power analysis that is conceptually orthogonal to the mathematical characteristics of the procedure: by examining the effect of low-rank imputation on a downstream pathway analysis, we are able to better understand the practical
implications of imputation, rather than considering metrics divorced from actual applications.

This is not to say that our procedure offers an ultimate solution to the problem of missing data. Indeed, as we discuss below in Chapter 5, a far superior approach would internalize the missing values themselves into the analysis as latent variables, an approach that does not pose any particular technical obstacles, but would unify the pathway analysis with the imputation of missing values. This would even permit analysis of signaling pathways that contain genes not observed in any sample, which we remain unable to solve with iterative imputation. Still, despite the appeal of such an approach, our procedure provides an off-the-shelf way to move past the choices that would otherwise accompany missing data. And, the overall setting of our analysis represents a more holistic approach to considering the effects of imputation on real-world analyses with scientific aims. We are able to emphasize the importance of statistical practice, to achieve analytic goals beyond pure research, but which instead advance our understanding of genomic activity.
Chapter 4

Interactive data visualization and analysis

The proliferation of -omics Big Data has produced methods that can be applied at scale. In the preceding sections, we discussed methods for pathway analysis across a multiple sample populations and several hundred pathways. The output of any individual pathway analysis is large, including pathway topology, parameter estimates, and inference results. In general, the volume of these results is difficult to manage, at best, and prohibitive of in-depth analysis, at worst.

Moreover, statistical methods are valuable not only because of the output of a specific analysis, but because of the wide range of analyses that are enabled. Especially for non-technical researchers, who may lack familiarity with extensive implementation of advanced statistical methods, it is critical not only to develop new methods and show sample outputs, but to provide for easy, public use implementations of the methods.

In this chapter, we discuss interactive data visualization tools we developed for exploratory analysis of the many outputs of the pathway methods we discussed. We describe a general template for designing interactive tools using the R programming language, and illustrate
the implementation with the output of specific analyses. We then introduce a computational algorithm to identify subpathways that may drive differential activity in the pathways we consider, implemented as a greedy, seed-based search. Our subpathway search may be applied to expression data, but we also extend it to the EMC-NetGSA settings. We discuss findings of the subpathway analysis using the TCGA dataset, and we introduce interactive software for pathway analysis of user-supplied data. The tool also accommodates user-specified seeds, thereby enabling a wide range of analyses.

4.1 Interactive visualization for pre-defined pathways

One use of interactive tools is to provide an interface for easy exploration of a large number of analytic results. In particular, many model outputs can be represented visually, and the accessibility of data visualizations is greatly improved by dynamic, interactive tools.

Some work has been done on interactive data visualization. Phandango [Hadfield et al., 2017] provides an interactive web application for visualization of phylogenetic datasets and analyses. General tools for interactive visualization and analysis have introduced software packages tooled for downstream implementation of of -omics data in individual analyses [Wickham, 2016, Yin et al., 2012]. The software SeqPlots [Stempor and Ahringer, 2016] implemented novel statistical graphics for visualizing cluster analysis applied them to gene expression data.

Given the multivariate nature of our data, as well as the cross-sectional comparisons performed in our inference, even straightforward analyses generate large volumes of statistical output that requires detailed attention.

Our discussion in Section 3.2.2 of Smad2 and Smad3 signaling demonstrates this point.
However, considered across five sub-populations compared with normal tissue, separate raw and imputed datasets, 176 pathways, and many hypothesis tests within each pathway, it is impractical in the extreme to generate these results “on-demand” for individual pathways, or to manually explore thousands of procedurally-generated plots.

To address this, we implemented an interactive data visualization using the Shiny software package in R [Chang et al., 2018]. This framework generates modern, responsive web content from within R, with a focus on data-driven and statistical applications. Shiny does not require direct knowledge of HTML, and provides straightforward bindings to common web elements. Particularly useful are forms, including drop-down menus, checkboxes, and radio inputs. These can be interacted with, updated, and modified by procedurally-generated Javascript, and supplemented with text content.

Shiny operates under a client-server paradigm. The user interacts with a website, which serves as the front-end client. A Shiny server then receives interactions and processes them on the back-end. In implementation, the user interface and server are structured similarly, and specify the visual interface, on the one hand, and the R-based software functionality, on the other.

Our general template for Shiny-backed pathway visualizations contains three components: (1) interactive graph topology of the signaling pathway; (2) output of statistical inference; (3) model estimates. For the interactive graph topology, we use the igraph [Csardi and Nepusz, 2006] and visnetwork [B.V. et al., 2018] packages in R. igraph is software for representing and interacting with graphs, and visnetwork provides an interactive front-end to the graphs. For our pathway analysis, we summarize the pathway in terms of the number of features it contains,
and offer a dynamic and interactive applet to move the graph components. Many of the signaling pathways in the NCI PID contain a large number of edges, and static representations of the topology obscures important aspects of the underlying network. We use color to convey the association strength of directed edges in $E$, specifically, scaled according to the partial correlation coefficient used in constructing that adjacency matrix.

The inference outputs are presented using bar plots, such as is shown in Figure 22. The pathway analysis is from the EMC-NetGSA analysis of the breast cancer data after imputation in Chapter 3, and indicates that integration of copy number into the pathway analysis does not obtain increase significance, relative to an expression-only analysis, despite a large increase in the magnitude of the test statistic. On the other hand, integration of methylation using EM-NetGSA yields a smaller test statistic, in magnitude, but a much higher degree of statistical significance. And, full integration using the final EMC-NetGSA model sustains the significance of the pathway disturbance. This provides a fuller picture of the biological components that comprise the larger biological process. Moreover, our data visualization was easy to construct, using the R language, and easily deployed for open access, demonstrating the potential for widespread use of similar tools to analyze and assess a large volume of analytic outputs.
Figure 22: $-\log_{10}(p)$-values (top) and signed test statistics (bottom) for the signaling events mediated by focal adhesion kinase pathway, in sarcoma cancer tumors. These plots were generated procedurally using an interactive data visualization application, publicly accessible online. Shown are the results of hypothesis tests for the full pathway, as well as the -omics features for each individual gene, tested as separate subpathways. The pathway analysis was performed on a fully-integrated dataset consisting of expression, methylation, and copy number, using the EMC-NetGSA model; expression and methylation (EM-NetGSA); expression and copy number (EC-NetGSA); and expression alone (NetGSA).

These plots permit comparison of significance between different methods: for example, the figure shows $-\log_{10}(p)$-values and test statistics for a signaling pathway. Color is used to enable quick visual discrimination between methods or populations. In addition to a hypothesis test for differential activity across all genes in the network, we also perform hypothesis tests for individual genes. In EMC-NetGSA analyses, we consider the integrated triplets of vertices for each gene, representing expression, methylation, and copy number. Plots are procedurally
generated based on the characteristics of the specific pathway. Finally, we use similar plots for the estimated mean parameters $\beta$. The three interface components are shown in Figure 23, based on four different analyses. The four analyses are publicly accessible online:

1. EMC-NetGSA breast cancer subtypes analysis

   https://zhang-lab.shinyapps.io/pathway-analysis-missing-data/

2. EMC-NetGSA pan-cancer analysis

   https://zhang-lab.shinyapps.io/pathway-analysis-tcga-cancers/

3. miRNA-EMC-NetGSA breast cancer subtypes analysis

   https://zhang-lab.shinyapps.io/breast-subtypes-pathways-mirna/

4. miRNA-EMC-NetGSA pan-cancer analysis

   https://zhang-lab.shinyapps.io/pathway-analysis-mirna/
4.2 Interactive pathway analysis for subpathway identification

The preceding methods focus on pathway analysis for known signaling pathways. But, it is often the case that subpathways within a larger signaling network of genes may drive differential expression of the full pathway [Martini et al., 2012]. Indeed, individual signaling pathways have been modeled as subpaths within a larger, trans-genome “superpathway” [Vaske et al., 2010]. It is therefore of interest to identify specific subpaths among fuller signaling pathways to identify...
the most significant components of a larger network.

Many methods exist for subpathway discovery, from blunt and straightforward methods to more sophisticated algorithmic approaches. He et al. [2017] gave an overview of several methods for identifying subnetworks of interest. They found that some methods were able to recover important subpathways in simulation studies, but failed to find any one method that dominated others.

Exhaustive elaboration of every subpathway is so computationally difficult as to be impossible. Some research nonetheless considers all possible pathways: Koumakis et al. [2016], for example, consider all possible subpaths of regulatory networks. Others reduce the computational complexity: TEAK [Judeh et al., 2012] considers linear subpathways between roots (in-degree 0) and leafs (out-degree zero), as well as a critical path method to connect cliques of three genes. Similarly, Nam et al. [2014] noted the combinatoric obstacle, and introduced PATHOME to identify gene expression subpathways that partition two populations. They started from leaf nodes, i.e. pathway vertices with in-degree equal to 0, and expanding subpaths on the basis of hypothesis tests for correlation coefficients.

Other approaches consider pre- or user-defined subpathways of interest, rather than emergent identification of subpaths based computational search. Li et al. [2009] considered metabolic pathways, identified based on the significance of a hypergeometric test of subpathways of interest, against the full set of all pathway nodes. Subpathway-GM [Li et al., 2013] integrated gene expression and metabolite data, and considered shortest paths connecting nodes of interest to identify subpathways.

Seed-based methods offer a more adaptive approach to subpath discovery. Lee et al. [2017]
proposed MIDAS, a greedy approach that proceeded from an initial seed and expanded a graph based on the most active edges and a stopping rule based on exponential decay. Likewise, Liu et al. [2019] performed integrated subpath discovery with a greedy algorithm. They considered observations of expression, methylation, and copy number, and used marginal $t$-tests, integration via Fisher’s test, and greedy agglomeration to increase the pathway score. Like MIDAS, they used a stopping rule based on the subpathway size, as well as a threshold for the “value added” by expanding the subpath.

To address the lack of consensus on how to identify subpathway drivers, we propose a seed-based, greedy algorithm that is firmly grounded in graph-theoretic pathway analysis, via the NetGSA model. We given an overview of the basic method, applied to gene expression. Then, we extend the method to the -omics setting, following the EMC-NetGSA integration scheme. We demonstrate interactive visualization of the subpathway discovery. Finally, we introduce a web application for interactive analysis, both for pathway inference as well as on-demand, seed-based subpathway discovery.

4.2.1 Greedy algorithm

As in previous sections, we considered a signaling pathway specified in terms of a directed graph $G = \{V, E\}$, defined across $p$ genes. We estimated the NetGSA model given in Equation 2.10. We performed $p$ hypothesis tests, one for each gene in the pathway. Although the individual hypothesis tests consider differential expression of a single gene, we emphasize that the full NetGSA model controls for pass-through effects in the individual genes of interest from other genes in $G$. 
We chose a seed gene from which to begin the analysis, based on the gene with the strongest significance, that is, the smallest \( p \)-value from the NetGSA hypothesis test. This seed represented the initial subpathway, which contained a single gene.

We proceeded in a greedy fashion: for a given subpathway, we found all child genes of any gene that was already part of the subpathway. For each of these child genes that are not already in the subpathway, we formed a candidate subpathway, based on the union of the current subpathway and the child node. We performed NetGSA hypothesis tests for each of the candidate subpathways.

We then compared the significance of each candidate subpathway with the current subpathway. If any candidate subpathway exhibited stronger significance than the current subpathway, in terms of a \( p \)-value of smaller magnitude, that candidate subpathway became the current subpathway. If no candidate subpathway increased the significance over the current subpathway, we stopped, and identified the current subpathway as the driver of the overall pathway.
This procedure is summarized in Algorithm 1.

**Algorithm 1:** Greedy NetGSA subpathway algorithm

**Data:** A graph $\mathcal{G} = \{V, E\}$; an initial subpathway of one vertex, the seed; the seed’s NetGSA test $p$-value

**while** Vertices in current subpathway have children not in subpathway **do**

- Find all $k$ child vertices of current subpathway vertices;
- Form $k$ candidate subpathways as union of 1 child and current subpathway;
- Calculate $k$ NetGSA test $p$-values, one for each candidate subpathway;
- **if** Some candidate subpathway has $p$-value less than current subpathway **then**
  - Set current subpathway to candidate subpathway with smallest $p$-value;
- **else**
  - break;

**end**

**Result:** A subpathway, as a subset of $V$

Extension of the method to the EMC-NetGSA -omics setting is trivial: in identifying the seed node, we consider the test set for an “individual” gene to consist of the triplet of expression, methylation, and copy number nodes for that gene. Similarly, in forming candidate subpathways, we include all three nodes for each child gene. Child genes are chosen only from the main signaling pathway, although the testing is performed across all integrated nodes in the graph.
4.2.2 Algorithm performance

Care is needed in the implementation of combinatorically intensive algorithms such as this greedy algorithm. A naïve application of the search method would estimate the pathway model at each iteration of the search, as well as for each candidate subpathway. However, substantial improvements can be made to run time by the relationship between the pathway model estimation, on the one hand, and the hypothesis test, on the other. Existing implementations of the NetGSA framework provide an API with a single access point to estimation and hypothesis testing. However, these represent two fundamentally different components and steps of the overall analysis. All hypothesis tests can be performed using the same estimated model, if the tests consider only subsets of the features in the estimated model. Therefore, we produced a modified version of the NetGSA implementation that allowed us to estimate the model once, at the beginning of the analysis, followed by greedy application of the hypothesis test only.

To assess the performance of our greedy algorithm, we considered simulations and real-world data analysis. For a simulation, we used a similar structure to that employed in Section 2.2 for the miRNA-EMC-NetGSA. We used a five level binary tree for both the control and treatment populations, with integration of methylation and copy number. We did not consider any miRNA integration for the simulation. We again set the association between copy number and expression to be 0.5, and association between methylation and expression to be −0.25. The top two levels of the tree had association 0.8, the next two had association 0.5, and the final level had association 0.2. We considered one mean scenario: the control group had mean 0, and the left branch of the treatment group had mean values 0.25 for expression, 1.0 for copy
number, and 0.5 for methylation. This a significant driver subpathway in the left branch of the tree.

We ran 1000 replicates, and in each replicate, we used as the seed node the binary tree root, which is a member of the subpathway. This was intended to allow assessment of the degree to which the greedy subpath recovers the entire true driver subpathway. Because of the directed binary tree structure of the simulated network, starting with a subpathway node other than the root of the tree would deterministically prevent recovery of the full subpathway, despite the equal means, which can be interpreted as equal pathway disturbance. Under this simulation scheme, 80% of the time, we recovered at least half of the subpathway, and 30% of the time, we identified 13 of the 16 subpathway nodes. In more than 5% of the simulations, we discovered the full pathway. For the binary tree, which contained 31 genes and 93 total features, the average run time was 0.04 seconds.

We also ran the simulation using as the seed the single most significant gene. We emphasize that the caveats above hold: given the pathway disturbance has equal magnitude in all features in the tree’s left branch, the gene with highest significance should follow a uniform distribution, whereas seeding with any non-root node will prevent recovery of the node. Of the 1000 replicates, only one single replicate contained any genes not in the true subpathway with disturbance. On the other hand, the median of the proportion of true subpathway genes contained in search algorithm’s output was low, at 6.25%. However, this reflects that, as the seed node’s location in the tree proceeds from the top level to the bottom, it becomes deterministically less likely that the full pathway will be included. In fact, with each level, half of the remaining nodes become unreachable. Indeed, for the subset of subpaths that used as seed
a gene from the top two levels of the tree, the median proportion of true subpathway genes in the found subpathway was 0.5. Run time was identical to that in the simulations where the subpathway was always seeded with the root node.

Together, these simulation results indicate that the subpathway method is accurate at finding most subpathway genes, when the seed’s location is located favorably to reaching all other subpathway genes. But, a shortcoming of the method is its strong dependence on starting in the “right” place. As we will show below, it is possible to sidestep this issue by considering multiple seeds, rather than only the most significant single gene.

For real-world application of our subpathway procedure, we considered the bladder urothelial carcinoma dataset from TCGA, which contained 448 samples total. This was a representative sample size for most cancers in our dataset, and was somewhat larger than the mean. Therefore, we believe the behavior of the algorithm is well-represented by application to this dataset. To illustrate the data analysis, discussed in further detail in the next section, we considered a full analysis of the 212 PID pathways discussed above, using EMC-NetGSA integration. After completion, we had analyzed 173 pathways, and identified subpaths within each. The full analysis took 191 seconds, or 3 minutes and 10 seconds.

For more detailed analysis of a specific signaling pathway, we considered the pathway for TCR signaling in naive CD4- T cells, again in the bladder cancer data. We chose this pathway because it contained a moderate number of genes features, 60, and had the largest subpathway of any pathway in the bladder cancer data analysis, \( q = 17 \). We note that \( q \) corresponds directly to the number of iterations performed before termination of the algorithm, as described in Algorithm 1, so this served as a sort of upper bound for computation time of any network.
This pathway was comparable in size to the largest pathway (96 genes), so we believe it is an appropriate measure of the algorithm’s performance.

We ran the greedy algorithm 100 times for the pathway. We measured only the main loop of the algorithm, so the run time did not include the initial estimation of the pathway model, or estimation of the association weights. We found a 95% confidence interval for the run time was (3.56, 3.68) seconds, with a mean of 3.62.

4.2.3 Analysis: results and visualization

We performed the iterative procedure four times, using expression-only NetGSA and EMC-NetGSA for the breast cancer subtype dataset, as well as each of the TCGA cancer populations. In each analysis of each pathway, we used as the seed the gene with the strongest significance.

In the pan-cancer data, across 11 cancers, we searched for subpathways for 1899 pathways. Of these, 970 subpathways—more than half—were degenerate, in the sense that they contained only the seed gene. Among these pathways, we identified 37 pathways for which the subpathway was consistent between the two NetGSA analyses, on expression data and on expression, methylation, and copy number. Of these, five subpathways consisted of three or more genes. These subpathways are listed in Table 7.
Table 7: Subpathways which were consistent between the expression-only and integrated analyses, and contained three or more genes. Cancers are labeled according to the codes in Table 1 (BLCA is bladder urothelial carcinoma; BRCA is breast invasive carcinoma; LUAD is lung adenocarcinoma; and THCA is thyroid carcinoma). \( p \) gives the total number of genes in the pathway, while \( q \) gives the number of genes in the subpathway.

<table>
<thead>
<tr>
<th>Cancer</th>
<th>Pathway</th>
<th>( p )</th>
<th>( q )</th>
<th>Subpathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLCA</td>
<td>Signaling events mediated by PRL</td>
<td>22</td>
<td>3</td>
<td>CDKN1A, PTP4A1, RHOA</td>
</tr>
<tr>
<td>BLCA</td>
<td>Urokinase type plasminogen activator</td>
<td>38</td>
<td>4</td>
<td>ITGB3, ITGB5, SERPINE1, VLDLR</td>
</tr>
<tr>
<td></td>
<td>uPA and uPAR mediated signaling</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRCA</td>
<td>Reelin signaling pathway</td>
<td>28</td>
<td>6</td>
<td>ITGB1, MAP2K7, MAPK8, NCK2, PIK3R1, RELN</td>
</tr>
<tr>
<td>LUAD</td>
<td>EPHB forward signaling</td>
<td>33</td>
<td>3</td>
<td>EFNB1, MAP4K4, RAC1</td>
</tr>
<tr>
<td>THCA</td>
<td>Presenilin action in Notch and Wnt signaling</td>
<td>45</td>
<td>3</td>
<td>AES, CTBP1, TLE1</td>
</tr>
</tbody>
</table>

As one example among these pathways, for “signaling events mediated by PRL” (PRL is the hormone prolactin) in bladder urothelial carcinoma (BLCA), a subpathway of three genes was found that was consistent across the two analyses. This pathway is shown in Figure 24, with the seed (PTP4A1) shown in yellow, and the other subpathway genes in orange. Research literature on prolactin has identified association with cancers: Sethi et al. [2012] discussed evidence to support a relationship between the PRL and tumorigenesis, particularly in breast, prostate, and gynecological cancers. However, they did not discuss its relation to bladder cancer. RHOA overexpression and overabundance have long been known to correlate with poor clinical outcomes in bladder cancer [Kamai et al., 2003]. Liu and Kwiatkowski [2015] found widespread CDKN1A mutations in bladder cancer, and that knockdown suppression of...
its expression increased the receptiveness of the cancer to treatment.

Figure 24: Pathway for “signaling events mediated by PRL.” The subpathway genes are shown in yellow (seed node) and orange. This subpathway of three genes was obtained from bladder urothelial carcinoma (BLCA), and was identified consistently in the NetGSA and EMC-NetGSA analyses.

We also considered subpathways that recurred across multiple pathways. As a heuristic, we considered those subpathway that were identified exactly in multiple pathways. Table 8 gives the recurrent subpathways that were found by the EMC-NetGSA integrated subpathway search. We highlight that the laminin subunit α gene LAMA2 was found by both Schuetz et al. [2005] and Yang et al. [2017] to be associated with renal cell carcinoma, as well as other
genes from the LAMA and LAMC families.

<table>
<thead>
<tr>
<th>Cancer</th>
<th>q</th>
<th>Subpathway</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>KIRP</td>
<td>3</td>
<td>LAMA2, LAMA5, LAMC1</td>
<td>2</td>
</tr>
<tr>
<td>LUAD</td>
<td>2</td>
<td>MAP2K1, MAPK1</td>
<td>2</td>
</tr>
<tr>
<td>THCA</td>
<td>2</td>
<td>SUMO1, XPO1</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 8: Subpathways that recurred exactly in multiple pathways for the same cancer, under the EMC-NetGSA analysis. Cancer gives the cancer code from Table 1 (KIRP is kidney renal papillary cell carcinoma; LUAD is lung adenocarcinoma; THCA is thyroid carcinoma). q is the number of genes in the subpathway, and n is the number of pathways in which the subpathway recurred.

Finally, we also considered the individual genes that were consistently included in the subpathways of multiple pathways, within each cancer. Figure 25 shows these genes, for both the NetGSA and EMC-NetGSA analyses, for any gene that recurred in at least four pathways in at least one cancer. This offers another perspective on the drivers of cancers, particularly the individual genes that may strongly associate with the cancer, rather than the subpathways that may drive overall pathway disturbance.

The prominence of specific genes in specific cancers is quite consistent between the two analyses. For example, a mutation that downregulates PIK3R1 was found by Lin et al. [2015] to correspond to increased tumor growth, as well as elevated metastasis, in kidney renal clear cell carcinoma. Cizkova et al. [2013] identified PIK3R1 as a prognostic biomarker for breast cancer, and Chen et al. [2018] found mutations of PIK3R1 coincided with increased tumorigenesis in
breast cancer. FOS was found by Bakiri et al. [2017] to correlate with accelerated tumor growth in liver cancer.
Figure 25: Results of greedy algorithm for subpathway discovery, using NetGSA with expression values only (top), and integrated EMC-NetGSA (bottom). The subpathway search was initialized using the single most significantly-disturbed gene in the pathway. The subpathway was expanded by including adjacent genes that increased the subpathway’s significance. Counts of gene occurrence in subpathways for any signaling pathway, organized by cancer. Only genes that recurred in some cancer across four or more pathways are shown. Genes identified in this manner as “drivers” were broadly consistent in the two analyses. This suggests these driver genes contribute to subpathway activity in a robust manner, and are not merely spuriously significant genes. Integration of methylation and copy number resulted in a larger number of driver genes within each cancer. Moreover, these significant driver genes do not recur in the same patterns across multiple cancers, which indicates differential functions for these genes across different cancers.
We published two interactive data visualizations of the subpathway analysis online, one a pan-cancer study, the other for breast tumor subtypes. For each, we summarize results for both EMC-NetGSA and expression-only NetGSA. The websites are located online at:

- Cancer drivers: subpathway identification among signaling pathways
  https://zhang-lab-network-analysis.shinyapps.io/subpathway-analysis-pan-cancer/

- Breast cancer drivers: tumor subtype subpathway identification among signaling pathways
  https://zhang-lab-network-analysis.shinyapps.io/subpathway-analysis-breast-tumors/

4.2.4 Interactive analysis

Individually, these findings suggest our subpathway analysis finds real drivers that underlie differential activity observed in signaling pathways. As before, the subpathway analysis produces a large volume of results, and for this we built an interactive visualization of the subpathways. In particular, we used dynamic graphs to display the full pathway and the subpathway we identified, an exploratory visualization that matches Figure 24.

However, the analysis relies upon the strong assumption that the most significant individual gene will also be part of a driver subpathway, but there is no theoretical reason this should be the case. This choice of seed was motivated by the intuition that strong significance of individual genes might correspond to significantly disturbed subpathways, but is also reflective of the computational complexity that accompanies any subpathway search.

In order to achieve more versatile subpathway identification, we implemented a tool for interactive pathway analysis and subpathway search. Like the visualization applications, we
used shiny to build a reactive website, including interactive graphs. To permit a wide range of datasets not restricted to the NCI PID and TCGA samples, we provided web forms for users to input custom pathway specifications as well as arbitrary data files.

Inputs took the form of white space-delimited text files containing an edge list, for the pathway specification, and genomic data with identifying labels for the rows, containing genomic features, and population labels for the columns, corresponding to samples. Upon uploading the pathway specification, we displayed interactive network topology, and after uploading data, we displayed population information.

We then perform the NetGSA hypothesis test for differential activity in the pathway between the two populations, and plotted the output of the statistical estimation and inference: significance test $p$-values for the pathway and individual genes; test statistics; and expression parameter estimates. This provides on-demand pathway analysis for user-provided data.

Additionally, we provided an interface to the subpathway search method. The user specifies a seed gene of interest, and the algorithm proceeds to identify subpaths from this seed, as described in Section 4.2.1. The analytic benefit of this interface is large: rather than restricting the analysis to a single seed gene based on a heuristic, we provided access to subpathway analysis using any gene in the pathway as a seed. This permits exploratory analysis of the roles played by various components of a signaling pathway. This is particularly valuable for pathway analysis of large pathways, which may contain multiple subpathways of interest or with differential activity. Figure 4.2.4 shows the inputs summary page and the results of the subpathway algorithm for the interactive analysis website.
We published the interactive subtype analysis online at the following URL: https://zhang-lab-network-analysis.shinyapps.io/interactive-subpathway-analysis/.

4.3 Discussion

Despite the large volume of analytic output that accompanies analysis of large-scale -omics datasets, we found that interactive data visualization was an effective technique to make results easy to review and explore, and also provided a convenient method to distribute the results of statistical analysis. And, although it was previously necessary to obtain specialist skills in web development to create interactive websites, modern tools like the R package shiny are a
valuable resource for development of responsive and professional visualizations. In particular, these interactive tools can be implemented without specialist knowledge of web technologies, and can be implemented with ease by statisticians familiar with R.

We also found that the large-scale nature of many signaling pathways made it difficult to identify specific components of those pathways that contribute to pathway activity or disturbance. In order to identify these “drivers” of genomic activity, we implemented a greedy search algorithm. Our algorithm was designed for and embedded within the pathway analysis framework, which was a valuable characteristic for ensuring a unified analysis. We used a greedy procedure to identify the subpathway seed, which was an effective and reasonable choice for initialization of the method. Because of the large number of vertices that are contained in some pathway graphs, identification of subpathways is a combinatorically hard problem, but the internal consistency of using NetGSA pathway analysis to identify subpathways of interest was justified as a starting point. We found our subpathway identification algorithm had good performance when implemented efficiently.

Still, because of the large number of possible graphs, it was not possible to perform an exhaustive search, and we recognized that the most significant individual gene, which we used as the subpathway seed, might not deterministically lead to the most significant subpathway. From a theoretical perspective, using many seeds to identify many subpathways could achieve more consistent recovery of the true subpathway with significant disturbance. This would require integration of incompatible subpathways, which could be achieved using a consensus-type rule that identifies subpathways that are stable across multiple seeds. From a procedural perspective, it is important to be able to assess multiple seeds, and we applied the lessons
learned from interactive data visualization to build a web application for interactive analysis. Specifically, by allowing the user to select a seed for the subpathway algorithm, we obtained a good balance between comprehensive analysis of the subpathways, and constraints on our ability to search through every possible subgraph.
Chapter 5

Concluding Remarks and Future Research

5.1 Concluding Remarks

Real-world analysis of bioinformatic data requires dynamic statistical practice in multiple stages. It can be necessary to variously remedy shortcomings of a dataset, identify biologically meaningful models of the relationships between multiple data types, and situate these analyses in statistically valid modeling frameworks. In the previous chapters, we presented several models for integrative pathway analysis, using data types collected for genes as well as for non-gene functional units. These pathway methods contribute a model-based, statistically sound procedure for integrating these -omics data types into a single pathway analysis. Although integrative analysis has become much more prominent in recent years, far too often, practitioners perform integration manually, by running separate analyses for each data type. We hope our statistical methods will contribute to more sophisticated analysis of the functional interactions found in the human genome.

We also addressed practical problems in performing pathway analysis on large-scale datasets. We implemented a method for imputation of missing data, and demonstrated that it provides
sensible values for downstream analysis. It could be possible to directly and jointly model missing values and the pathway structure, as we discuss below. But, such an approach would require tailor-made customization for each subsequent analysis of a dataset. Rather, our method permits any downstream analysis that requires complete values, thereby empowering a wide range of analyses based on a single imputed dataset.

We also addressed the difficulty encountered by large-scale analysis of genomic Big Data, particularly the practical difficulty of exploratory data visualization and analysis. We implemented several interactive web applications to facilitate and promote analysis of -omics datasets. Moreover, we situated an exploratory method for subpathway analysis in the interactive framework, as well, permitting on-demand, interactive analysis. Subpathway discovery is a computationally intensive application, especially because of the combinatoric complexity of even simple graphs. By presenting an analytic interface, we permit more flexible analysis of a wide range of scenarios.

In aggregate, these methods provide a pathway model as a starting point for analysis, an imputation procedure to maximize the applicability of the pathway analysis, and several interactive applications that allow exploratory visualization and analysis of the results. We believe this presents a coherent, unified analytic pipeline that focuses on development of statistical methods, but also addresses practical problems in data analysis. Especially for bioinformatics datasets, the role of the statistician is not simply to develop new methods, but also to empower researchers to expand human knowledge in their respective fields, and we hope to have accomplished that, at least in part, through this dissertation.
5.2 Future Research

Research is never “complete,” one merely reaches a stopping point. This thesis is just such a stopping point: avenues for further research sprout like the heads of a hydra in several directions.

One direction for future work is more comprehensive integration of miRNAs. In particular, the TCGA dataset contains not only miRNA expression, but also copy number variation. Although biological understanding of miRNA activity is still relatively shallow, it is easy to imagine an integration scheme similar to that of EMC-NetGSA, with directed edges from copy number vertices to miRNA expression vertices. On the other hand, this further expansion of the graph would result in substantial expansion of the adjacency matrix: the largest pathway in the PID, ErbB1 downstream signaling, contains 653 features in total, of which 368 are miRNA expression vertices. Already, the number of miRNAs is larger than the number of gene-level features for every pathway in the PID, and integration of an additional vertex for each miRNA would simply exacerbate this problem. Quickly, the dimension of the observation vectors $y_i$—equal to the number of graph vertices—may well become larger than the number of samples, and special care must be taken in these situations.

Another opportunity for extension is to make the covariance structure of the integrative NetGSA variants richer. In particular, the miRNA-EMC-NetGSA and EMC-NetGSA models are parameterized with a single error variance and a single random effect variance. There is no reason a priori to suppose that the variances are equal across data platforms, and empirically, we observed substantial variation in the variance estimates for the same pathway under different integration schemes. A straightforward starting point would be to stratify the variances by
data platform, with separate variance magnitudes for each data type.

We also observed that the error variances also differ within sample populations, and even between different features of the same data type. Therefore, it is desirable to specify a model with a more flexible covariance structure than the isotropic error that treats the magnitude of the variances as equal. This might be accomplished through a hierarchical linear model, or separate error variances by individual genes. It might also be viable to employ cluster analysis to identify coherent groups of features or samples that exhibit similar variances, with heterogeneity between groups. More preferable would be a latent class model that internalizes the presence of separate groups for the variances of -omics features or subjects. This model is justified by our finding that, although there is some heterogeneity in error variance, it empirically does not appear that the variance is different between every feature.

A third direction for further research is network regression. Despite the apparent coherence of the NetGSA model discussed in Shojaie and Michailidis [2009, 2010], Ma et al. [2014b], we found that the model is not identifiable. A brief simulation study illustrates the problem:

We considered a network with three nodes arranged linearly, that is, with an edge from gene 1 to gene 2, and another from gene 2 to gene 3. Define the values $\sigma_1^2 = 1$, $\sigma_2^2 = 5$, $\beta_1 = 1_3$, and $\beta_2 = -1_3$.

We calculated $A_1, A_2$, and generated data from two models:

1. Random effect variance $\sigma_\gamma^2 = \sigma_1^2$, error variance $\sigma_\epsilon^2 = \sigma_2^2$;

2. Random effect variance $\sigma_\gamma^2 = \sigma_2^2$, error variance $\sigma_\epsilon^2 = \sigma_1^2$.

From each population, we generated 150 observations. NetGSA produced the following estimates of the variance components:
In each, the larger magnitude is clearly allocated more to $\sigma_\gamma^2$ than $\sigma_\epsilon^2$, regardless of the true magnitude. This identifiability may be resolved in the mixed-models framework by using some constraints on the relative magnitudes of $\sigma_\gamma^2$ and $\sigma_\epsilon^2$; further discussion on the identifiability of linear mixed-models is given in Wang et al. [2013].

However, another approach would be to consider a fixed-effects linear model without any random effects. Such a model would not model correlation in the errors across genes at the level of individual samples, so we may modify this to produce a network regression:

$$\mathbf{y}_i = \Lambda \beta + \Lambda \mathbf{e}_i$$

Rearrangement of terms, specifically under the assumption that $\Lambda$ is invertible, gives the linear means model model:

$$\Lambda^{-1} \mathbf{y}_i = (\mathbf{I}_p - \Lambda) \mathbf{y}_i = \beta + \epsilon_i$$

We note that the arguments of Shojaie and Michailidis [2009, 2010] depend upon the invertibility of $(\mathbf{I}_p - \Lambda)$ to define $\Lambda$ in the first place, and employ a limit approximation to address situations in which this is not the case. The network regression formulation would sidestep this requirement, and still accommodate correlated errors.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>$\hat{\sigma}_\gamma^2$</th>
<th>$\hat{\sigma}_\gamma^2$</th>
<th>$\hat{\sigma}_\epsilon^2$</th>
<th>$\sigma_\epsilon^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.55</td>
<td>1.0</td>
<td>1.12</td>
<td>5.0</td>
</tr>
<tr>
<td>2</td>
<td>4.59</td>
<td>5.0</td>
<td>0.90</td>
<td>1.0</td>
</tr>
</tbody>
</table>
In this framework, to perform a significance test of differential activity across all genes in the signaling pathway, we have a network contrast $\ell$ and an $F$ statistic

$$\ell = (-1 \cdot 1\Lambda_1, 1 \cdot 1\Lambda_2)$$

$$F = \frac{(\ell\hat{\beta})^2 / (\ell(X'X)^{-1}\ell'\hat{\sigma}^2)}{\hat{\sigma}^2}$$

where, under the null hypothesis of no difference, $F \sim F_{1, np-2p}$. This is straightforward linear model theory, and does not require substantial conceptual developments.

The network regression model formulated in Equation 5.2 is appealing. One attractive characteristic is that the network regression model directly exposes the elements of the weighted adjacency matrix $A$, without requiring intermediate matrix inversion prior to parameter estimation. The NetGSA framework supposes the adjacency matrix association weights between adjacent vertices are fixed and known, presumably obtained exogenously from a supplementary source. In practice, it is not generally viable to obtain partial correlations to use as weights except from the dataset of analysis. We typically employ a two-stage approach, estimating first the partial correlations, then estimating the regression mixed model conditionally. In principle, this will introduce bias into the estimation, and it would be preferable to estimate a joint model that simultaneously estimates the adjacency weights and the activity parameter $\beta$.

Such a joint model of the adjacency weights and the activity parameter introduces some mathematical obstacles. First, although the influence matrix is derived from the adjacency matrix, the interpretation of the term $(I_p - A)$ is not entirely clear, especially in the setting
where $A$ is obtained using the approximation of Shojaie and Michailidis [2010]. Second, many complex signaling pathways correspond to singular unweighted adjacencies $A^*$, in which situations the model may not be estimable. This joint model warrants further research, especially because non-singular signaling pathway $A^*$ are available, even if not the majority.

Another direction for further research is joint modeling of the missing values directly with the pathway analysis parameters. The linear method from Chapter 3 has a mathematical simplicity, but the two-stage procedure may introduce bias into the estimation, and moreover, the imputed values do not have a biological interpretation. To remedy this issue, we would like to employ a latent variable model, in which missing gene features are imputed as part of the model. Our pathway analyses can be understood to be structural models of biological phenomena, and latent variables for features with missing values would embed the imputation process fully within the functional pathway model.

Another shortcoming of our iterative imputation method is its dependence on the presence of at least some observed value for each feature of interest. However, we consistently found the TCGA dataset was missing observations of features of interest in any subject. As discussed in Chapter 2, when the missing features are not of primary interest, namely, gene methylation or copy number, or miRNA expression, we may simply omit them from the graph, and proceed with our analysis of the signaling pathway, which is primarily defined over gene expression vertices. But, this omission of missing features does not address the scenario in which gene expression features are missing. In principle, one may omit vertices that are leaves or roots, with in- and out-degree of one, respectively. But omitting the gene feature from the signaling network may fundamentally alter the graph topology in a way that destroys the connectedness
between other graph vertices. A latent variable model, on the other hand, might provide a mechanism by which to save the analysis.

To implement such a model, in the first place, would require either an endogenous model of the adjacency weights, as discussed above, or a “stand-in” default value for association weights. Then, simulation-based methods could be employed to sample the unobserved values, in order to permit pathway pass-through effects even in networks with missing gene vertices. Although it is unlikely that the unobserved gene could be accurately or precisely analyzed inferentially, this would at least enable pathway analysis in situations that are currently impossible because of missing features, despite most features being observed.

Finally, much work remains for subpathway identification. The assumption that the single most significant gene in a pathway will also be in the most significant subpath is a strong one, and it is entirely possible that no individual gene as seed will yield the maximally-significant subpathway. Nonetheless, an initial remedy would be to use each gene in the pathway as a seed, to identify its corresponding greedy pathway. This is possible in principle through our interactive data analysis, which lets the user select a gene to use as a seed for the greedy subpathway algorithm. To procedurally identify a single significant subpathway from many generated from multiple seeds, we require on the one hand an integration or consensus method to combine subpathways from different seeds, and on the other hand, a compatibility score to identify when there is no significant subpathway. This would be a function of the compatibility or overlap of subpathways generated from different seeds.

Computational cost of subpathway identification remains high, but it is not apparent where savings might be found: the initial estimation of the pathway model is a fixed cost, and the
greedy algorithm must proceed iteratively, expanding the subpathway by a single vertex at each step. However, the computational cost for an individual pathway is minimal, so some computational time can be reduced by parallelization when considering a large number of pathways. At this point, we have only implemented the subpathway algorithm for NetGSA on expression only, and EMC-NetGSA. A natural next step is to apply the same method to the miRNA-EMC-NetGSA analysis, to identify subpathways with the miRNA-gene target effects accounted for.
Bibliography


Youcheng Lin, Zhao Yang, Abai Xu, Pei Dong, Yi Huang, Huan Liu, Feida Li, Haifeng Wang, Qian Xu, Yongqiang Wang, et al. Pik3r1 negatively regulates the epithelial-mesenchymal transition and stem-like phenotype of renal cancer cells through the akt/gsk3β/ctnnb1 signaling pathway. *Scientific reports*, 5:8997, 2015.

Fang Liu. Inhibition of smad3 activity by cyclin d-ckd4 and cyclin e-ckd2 in breast cancer cells, 2011.


Sarah Wang, Elizabeth E Hwang, Rajarshi Guha, Allison F O’Neill, Nicole Melong, Chansey J Veinotte, Amy Saur Conway, Kellsey Wuerthele, Min Shen, Crystal McKnight, et al. High-throughput chemical screening identifies focal adhesion kinase and aurora


