Title

Sample-Specific Cancer Pathway Analysis Using PARADIGM

Permalink

https://escholarship.org/uc/item/9zr859ks

Author

Benz, Stephen Charles

Publication Date

2012

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
SANTA CRUZ

SAMPLE-SPECIFIC CANCER PATHWAY ANALYSIS USING PARADIGM

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

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Stephen C. Benz

June 2012

The Dissertation of Stephen C. Benz
is approved:

Professor David Haussler, Chair

Professor Joshua Stuart

Professor Nader Pourmand

Dean Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures **v**

List of Tables **xi**

Abstract **xii**

Dedication **xiv**

Acknowledgments **xv**

1 Introduction **1**

1.1 Identifying Genomic Alterations 2

1.2 Pathway Analysis 5

2 Methods to Integrate Cancer Genomics Data **10**

2.1 UCSC Cancer Genomics Browser 11

2.2 BioIntegrator 16

3 Pathway Analysis Using PARADIGM **20**

3.1 Method 21

3.2 Comparisons 26

3.2.1 Distinguishing True Networks From Decoys 27

3.2.2 Tumor versus Normal - Pathways associated with Ovarian Cancer 29

3.2.3 Differentially Regulated Pathways in ER+ve vs ER-ve breast cancers 36

3.2.4 Therapy response prediction using pathways (Platinum Free Interval in Ovarian Cancer) 38

3.3 Unsupervised Stratification of Cancer Patients by Pathway Activities . . 42

4 SuperPathway - A Global Pathway Model for Cancer **51**

4.1 SuperPathway in Ovarian Cancer 55

4.2 SuperPathway in Breast Cancer 61

4.2.1 Chin-Naderi Cohort 61
4.2.2 TCGA Breast Cancer 63
4.3 Cross-Cancer SuperPathway 67

5 Pathway Analysis of Drug Effects 74
5.1 SuperPathway on Breast Cell Lines 75
5.2 Culling Targets Using Machine Learning 83
5.3 Simulating Knockdowns in PARADIGM 83
5.4 Breast Cell Line Simulated Knockdowns 89

6 Conclusion 94
6.1 Conclusions 94
6.2 Future Expansions 95
6.3 Final Thoughts 100

List of Figures

1.1 TCGA GBM samples show overall alterations across GBM pathway, but changes to specific genes are limited. Part A) shows an OncoPrint profile of the 122 GBMs with complete genomic information. Part B) illustrates the underlying pathway connections that explain why alterations within any single node can deactivate TP53 function. OncoPrint generated from The cBio Cancer Genomics Portal (http://www.cbioportal.org/). 3

1.2 NCI Pathway interactions in TCGA GBM data. For all (n=462) pairs where A was found to be an upstream activator of gene B in NCI-Nature Pathway Database, the Pearson correlation (x-axis) computed from the TCGA GBM data was calculated in two different ways. The histogram plots the correlations between the A's copy number and B's expression (C2E, solid red) and between A's expression and B's expression (E2E, blue). A histogram of correlations between randomly paired genes is shown for C2E (dashed red) and E2E (dashed blue). Arrows point to the enrichment of positive correlations found for the C2E (red) and E2E (blue) correlation. 6

2.1 UCSC Cancer Genomics Browser proportions view of chromosome 9 across three copy number variation datasets. The yellow box represents the genomic region that is commonly deleted across all three datasets and contains the tumor suppressor CDKN2A/B. 13

2.2 Integrative visualization of TCGA glioblastoma multiforme (GBM) genomic data across four separate pathways. (a,b) Histogram of non-silent mutations (e.g. missense, insertion, deletions, etc.) per sample per base in GBM somatic and germ line tissues, respectively. (c) Copy number variation (CNV) in GBM tumor and normal samples. Red or blue color represents amplified or deleted genes. Note that the clinical data to the right of (c) are sorted by tissue type: tumor (black) vs. normal (yellow). The vast majority of copy number alterations seen in the heatmap occur in the tumor samples. (d) A Bonferroni-corrected t-test comparing the distribution of CNV in tumor versus normal samples. A green bar represents a significantly deleted gene ($p < 0.05$, after Bonferroni correction) and a red bar represents a significantly amplified gene in the tumor samples. 15

2.3 BioIntegrator web interface, designed to allow users to access both raw and analytical results in an exploratory fashion. Part A highlights the ability to flexibly search for genes and clinical features, as well as drag and drop reordering. Part B shows the dynamic nature of the sorting, allowing discrete clinical features to be placed at top with the data summarized under each label, allowing intuitive visualization of differences. 16

2.4 Genes were scored independent of sample labels using both gene expression and copy number data from a melanoma cohort [47] in the BioIntegrator. Samples were split 80/20 into training and test sets and feature selection was performed on the training data thresholding on correlation coefficient [15]. Samples were run through SVM-light [38] using default settings, and the top 20 models (minimum 91% accuracy) were selected for downstream analysis. Part A shows the top 35 features selected by the SVMs, with both SVM weight and frequency plotted. When those top 35 features were run through NCI's Pathway enrichment analysis, both Caspase Cascade in Apoptosis ($p < 4e-6$, Part B) and Fas Signaling Pathway (CD95) ($p < 0.001$) were found to be significantly enriched. The color of nodes in Part B illustrate the average enrichment towards the mutated class (red) or non-mutated class (blue) as calculated by the BioIntegrator gene level perturbation score. Nodes seen in both Part A and Part B are denoted with an orange asterisk. 18

3.1 Overview of the PARADIGM method. PARADIGM uses a pathway schematic with functional genomic data to infer genetic activities that can be used for further downstream analysis. 20

3.2 Conversion of a genetic pathway diagram into a PARADIGM model. **A**. Data on a single patient is integrated for a single gene using a set of four different biological entities for the gene describing the DNA copy number, mRNA and protein levels, and activity of the protein. **B**. PARADIGM models various types of interactions across genes including transcription factors to transcript and protein targets (upper-left), protein subunits aggregating in a complex (upper-right), protein post-translational modification (lower-left), and sets of gene products in a family performing redundant functions (lower-right). **C**. Toy example of a small sub-pathway involving P53, its inhibitor MDM2, and the high level P53 determined cell process, apoptosis as represented in the model. 25

3.3 Distinguishing decoy from real pathways with PARADIGM and SPIA. Decoy pathways were created by assigning a new gene name to each gene in a pathway. PARADIGM and SPIA were then used to compute the perturbation of every pathway. Each line shows the receiver-operator characteristic (ROC) for distinguishing real from decoy pathways using the perturbation ranking. In breast cancer, the areas under the curve (AUCs) are 0.669 and 0.602 for PARADIGM and SPIA, respectively. In GBM the AUCs are 0.642 and 0.604 respectively. 28

3.4 Summary of FOXM1 integrated pathway activities (IPAs) across all samples. The arithmetic mean of IPAs across tumor samples for each entity in the FOXM1 transcription factor network is shown in red, with heavier red shading indicating two standard deviations. Gray line and shading indicates the mean and two standard deviations for IPAs derived from permuted data. 30

3.5 Summary of IPAs of FOXM1 and all tested transcription factors. Histograms of each sample's IPA in FOXM1 or in any other transcription factor from the PID. A. Histogram of IPAs, with IPAs of 0 removed. By a Kolmogorov-Smirnov test, FOXM1 shows a different distribution with a p-value $< 10^{-225}$. B. Histogram IPAs including 0 valued IPAs. By a KS test, FOXM1 has a different distribution of IPAs with a p-value $< 10^{-236}$. 30

3.6 Expression of FOXM1 Transcription Factor Network genes in high grade vs. low grade carcinoma. Using data from Etemadmoghadam et al (2009) we plotted expression of genes in the FOXM1 transcription network in either low grade (I) or high grade (II/III) ovarian carcinomas. A student's t-test was used to assess the significance of differential expression between high grade and low grade, with the p-value below each pair of boxplots. 31

3.7 The PathOlogist approach shows significant activation of the FOXM1 pathway in tumor samples. (©2012 IEEE) 34

3.8 PathOlogist reveals that patients with higher activity levels in the Downstream signaling in naive cd8+t cells pathway are associated with improved response to platinum therapy. (©2012 IEEE) 40

3.9 Clustering analysis of ovarian samples using Paradigm shows that that patients with lower IPA's for genes in the E-cadherin adherens junction pathway are associated with better response to platinum therapy (p=0.003). (©2012 IEEE) 41
3.10 Clustering of IPAs for TCGA GBM. Each column corresponds to a single sample, and each row to a biomolecular entity. Color bars beneath the hierarchical clustering tree denote clusters used for Figure 3.12. 43
3.11 Clustering of IPAs for TCGA OV. Each column corresponds to a single sample, and each row to a biomolecular entity. Color bars above the heatmap denote clusters used for Figure 3.13. 44
3.12 Kaplan-Meier survival plots for the GBM clusters from Figure 3.10. . . 46
3.13 Kaplan-Meier survival plots for the OV purple cluster versus the rest from Figure 3.11. 47
3.14 Clustering glioblastoma multiform (GBM) datasets with HOPACH did not reveal any obvious patient sub-types when using either gene-level copy number or expression data. Patient samples (columns) were clustered according to gene-level copy number or expression (rows). Clustering may be dominated by the nearly patient-wide amplifications and deletions (and correlated high and low expression) for a subset of the genes in the genome (large blue and red swaths in the heatmaps). A. Copy number estimates from competitive genome hybridization for 17,508 probes across 267 tumor (left) and 170 normal (right). B. Microarray gene expression data for 11,240 probes across 243 tumor (left) and 10 normal (right) samples. 50

4.1 Topographical visualization of the SuperPathway. 53
4.2 Number of subnets found for a variety of IPA cutoffs. The red line represents actual samples, with the black line representing the permuted samples. The max difference of 156 subnets between these two populations was found at IPA cutoff of 3.3. 58
4.3 Paradigm reveals the regulatory network underlying estrogen receptor positivity in the Chin-Naderi Cohort. 62
4.4 Paradigm concept hubs separating Luminal and Basal in the TCGA Breast Cancer Cohort. In order for a link to appear in the subnetwork, the parent node needed at minimum 10 children nodes in the overall network. In addition, the child node needed at minimum 5 children nodes in the overall network. Concepts are colored orange if their activity is higher in Basal samples, and cyan if their activity is higher in Luminal samples. 64

4.5 Mutations in different protein domains show differential pathway activity. A: Mutational comparison of TCGA breast cancer samples versus mutations seen in the COSMIC mutation database. B: Pathway hubs that show differential activity between kinase and non-kinase mutations, colored and sized by significance. 66

4.6 PARADIGM activities computed across 1382 TCGA samples from 8 tissues of origin with matched copy-number and median-centered gene expression data. 68

4.7 Receiver operating characteristics of SVM-Light classifications done across tissue types. The Cyan ROC indicates a linear SVM that was trained on lung squamous versus adenocarcinoma and given breast samples labelled basal or luminal for classification (AUC=0.971751). The orange ROC indicates the performance of a linear SVM that was trained on breast basal versus luminal and tested on lung squamous versus adenocarcinoma (AUC=0.894994). 70

4.8 Paradigm concept hubs found to have high SVM weights when trained on Lung Adenocarcinoma versus Lung Squamous. In order for a link to appear in the subnetwork, the parent node needed at minimum 10 children nodes in the overall network. In addition, the child node needed at minimum 5 children nodes in the overall network. Concepts are colored orange if their SVM weight is positive, meaning they contribute to the resulting classification, and cyan if they have a negative SVM weight, meaning they offer no discriminative power. 72

5.1 Clustering of TCGA samples with breast cancer cell lines. 76

5.2 Uncentered Pearson correlation of pathway concepts show samples and cell lines from the same subtype are more correlated, regardless of the origin of the sample. 77

5.3 Cell-line subtypes have unique SuperPathway network features. In all panels, each node represents a pathway concept corresponding to a protein (circle), a multimeric complex (hexagon), or an abstract cellular process (square). Node sizes are drawn in proportion to the DA score; larger nodes correspond to concepts more correlated with a particular subtype than with all other subtypes. Color indicates whether the concept is correlated positively (red) or negatively (blue) with the subtype of interest. Lines represent interactions, including proteinprotein interactions (dashed lines) and transcriptional interactions (solid lines). Interactions are included if they connect concepts whose absolute level of DA is higher than the mean absolute level. Labels on some nodes are omitted for clarity. (A) An ERK1/2 subnet preferentially activated in basal breast cancer cell lines. (B) A MYC/ MAX network activated in claudin-low cell lines. (C) A FOXA1/FOXA2 net- work up-regulated in the luminal subtype. (D) A CTNNB1 subnet down-regulated in the ERBB2AMP subtype. . 79

5.4 Pathway diagrams can be used to predict response to therapies. (A) (Left) Basal breast cancer cell lines respond preferentially to the DNA-dam- aging agent cisplatin. Each boxplot represents the distribution of drug response data for basal (right) and non-basal (left) cell lines. (Right) Basal cell lines show enhanced pathway levels in a subnetwork associated with the DNA-damage response, providing a possible mechanism by which cisplatin acts in these cell lines. (B) (Left) ERBB2AMP cell lines are sensitive to the HSP90 inhibitor geldanamycin. (Right) The ERBB2HSP90 network is up-regulated in ERBBP2AMP cell lines. Conventions are as in Figure 5.3. 81

5.5 Schematic indicating how knockdown is performed within the factor graph central dogma. Part A represents the standard factor graph dogma model provided by PARADIGM. Part B represents the modified dogma used for the proteins that are being knocked out, with an example of how signal might propagate post KO. Of note is the disconnection of considering the mRNA evidence node and transcriptional regulation, with the resulting signal propagation to the protein node and any interactions that may occur with the knocked down node. 84

5.6 Visualization of IPLs for the region around GNAI3 for the best Knockdown effectiveness from Table 5.1: A) GFPc of GFP_scrCvsHR, B) GFPssi of GFP_scrCvsHR and C) siRNAt of GNAI3_06_1vsHR 87

5.7 Classification of response to QNZ (an NF-KB inhibitor) in the breast cell lines, before and after simulated knockdown experiments to E2F3. Tail of the arrow represents the predicted resistance score using the pathway activities, and the head represents the prediction after simulated knockdown was performed. Blue arrows indicate cell lines that are labelled sensitive and red arrows indicate resistant cell lines. 91

5.8 Classification of response to Oxaliplatin in the breast cell lines, before and after simulated knockdown of MAPK1. 91

6.1 An example expanded PARADIGM dogma model that accounts for promoter states and allows attachment of Methylation data. (©IEEE 2012) 95

6.2 Allele and Isoform Specific Factor Graph Representation. An example of how we might represent a gene ("GENE1") with allele-specific data and two isoforms. 98

List of Tables

3.1 The FOXM1 transcription factor network shows up as being upregulated in ovarian tumors compared to normal consistently across pathway methodologies. 33
3.2 FOXA1 transcription factor network is upregulated in ER+ve tumors whereas the Hif-1-alpha transcription network and Syndecan-1 mediated signaling events are down-regulated. 37
3.3 Pathways associated with platinum-free interval in ovarian carcinoma. . 39

4.1 Features and Interactions in SuperPathway (July 2011 version) 53
4.2 Top TCGA Ovarian Concepts from SuperPathway 56
4.3 Top pathway concepts consistently found in subnets across the OV samples using IPA cutoff 3.3. Only HSP90AA1 has more null samples than real, indicating a likely false positive. 59

5.1 Correlations between simulated knockdowns and true siRNA knockdowns for a series of genes in a colon cancer cell line. 86
5.2 Resitance shift and p-values for the top 20 simulated KO and drug combinations ranked by shift score. 92

Abstract

Sample-specific cancer pathway analysis using PARADIGM

by

Stephen C. Benz

Identifying key somatic alterations in cancer is a critical step in understanding the mechanisms that ultimately determine a patient's treatment outcome. High-throughput data are providing a comprehensive view of these molecular changes for individual samples, and new technologies allow for the simultaneous genome-wide assay of genome copy number variations, gene expression, DNA methylation, and epigenetics of patient tumor samples and established cancer cell lines. Analyses of current datasets find that genetic alterations between tumors can differ but often involve common pathways. It is therefore critical to identify relevant pathways involved in cancer progression and detect how they are altered in different patients. This work presents a novel method called PARADIGM for inferring tumor-specific genetic pathway activities incorporating curated gene-pathway interactions. A gene is modeled by a factor graph as a set of interconnected variables encoding the expression and known activity of itself and its upstream and downstream products, allowing the incorporation of many types of -omic data as evidence. The method predicts the degree to which a given pathway's activities (e.g. internal gene states, interactions, or high-level outputs) are altered in the tumor sample using probabilistic inference. Compared to a competing pathway activity inference approaches, PARADIGM identifies altered activities in cancer-related pathways with fewer false-positives, as shown in glioblastoma multiform (GBM), ovarian (OV) and breast cancer datasets. PARADIGM also identified consistent pathway-level activities for subsets of the GBM and ovarian serous cystadenocarcinoma patients that are overlooked when genes are considered in isolation. Furthermore, grouping GBM and OV patients based on their significant pathway perturbations divides them into clinically-relevant subgroups having significantly different survival outcomes. Further analysis was done using an integrated pathway termed the SuperPathway that gives a more consistent global view of biology, illustrated through the analysis of ovarian, breast, and cross-cancer analysis. Finally, PARADIGM was used to simulated knockdowns in a pathway model as an approach to understand drug effects and provide a rational approach towards combination therapies. These findings suggest that therapeutics might be personalized and chosen to hit target genes at critical points in the commonly perturbed cancer pathway(s) of specific patients using this model.

To my parents,

Christopher and Constance Benz,

and my best friend and love of my life, Danielle.

Acknowledgments

I'd like to thank my advisor David Haussler for his support and encouragement of this work, and my committee chairman Josh Stuart for the advice and guidance he's given on PARADIGM. I thank Nader Pourmand for his guidance on my committee as we move towards sequencing data, and Scott Lokey for taking time out to be a part of my advancement committee. Finally I'd like to thank the other members of the UCSC Cancer Genomics team, without whom this work could not have been done: Zack Sanborn, Charlie Vaske, Jing Zhu, Chris Szeto, Larry Meyer, Fan Hsu, and Ting Wang. This work has been supported by a NIH Training Grant, the I-SPY INSTINCT project (QB3), TCGA Pilot and U24 funding sources, and the Stand Up To Cancer (SU2C) project.

Chapter 1

Introduction

There is no doubt about the importance of cancer - the World Health Organization has identified it as the leading cause of death worldwide (13% in 2004) [84]. Although all cancers are deadly, individual cancer frequencies and mortality rates vary widely. Lung cancer is currently the most deadly, responsible for an estimate 1.3 million deaths a year, with stomach, colorectal, liver and breast cancer close behind. It is projected that approximately 30% of life-threatening tumors are preventable and caused by avoidable risk factors [16], with early detection potentially capable of preventing another third. However, even after detection, standard therapies are only capable of successfully curing a minor proportion of all malignancies.

Cancer is fundamentally a disease of the genome, often defined by three main traits: uncontrolled growth, invasiveness, and potential for metastatic dissemination. Although genetically inherited germline alterations have been shown to be responsible for a subset of cancers, the majority of new cancers develop through somatic alterations within a tissue. These alterations can occur as a series of spontaneous mutations during cell division or, alternatively, from the effects of DNA carcinogens such as chemicals (exogenous or endogenous), tobacco smoke or viral infection. In addition, resulting genomic changes can effect two main classes of genes (oncogenes and tumor suppressors), with alterations often being required in both in order for a cell to undergo full tumorgenesis. It has become increasingly clear that in order to successfully treat tumors, identification of these genomic alterations is required so more directed therapies can be applied.

1.1 Identifying Genomic Alterations

While several high-throughput technologies have been available for identifying these alterations within each cancer, only a handful of successes have been achieved based on these advances. For example, 25% of breast cancer patients presenting with a particular amplification or overexpression of the ERBB2 growth factor receptor tyrosine kinase can now be treated with trastuzumab, a monoclonal antibody targeting the receptor [82]. However, even this success story is complicated by the fact that fewer than 50% of patients with ERBB2-positive breast cancers actually achieve any therapeutic benefit from trastuzumab, emphasizing our incomplete understanding of this well-studied oncogenic pathway and the many therapeutic-resistant mechanisms intrinsic to ERBB2-positive breast cancers [55]. This overall failure to translate modern advances in basic cancer biology is in part due to our inability to comprehensively

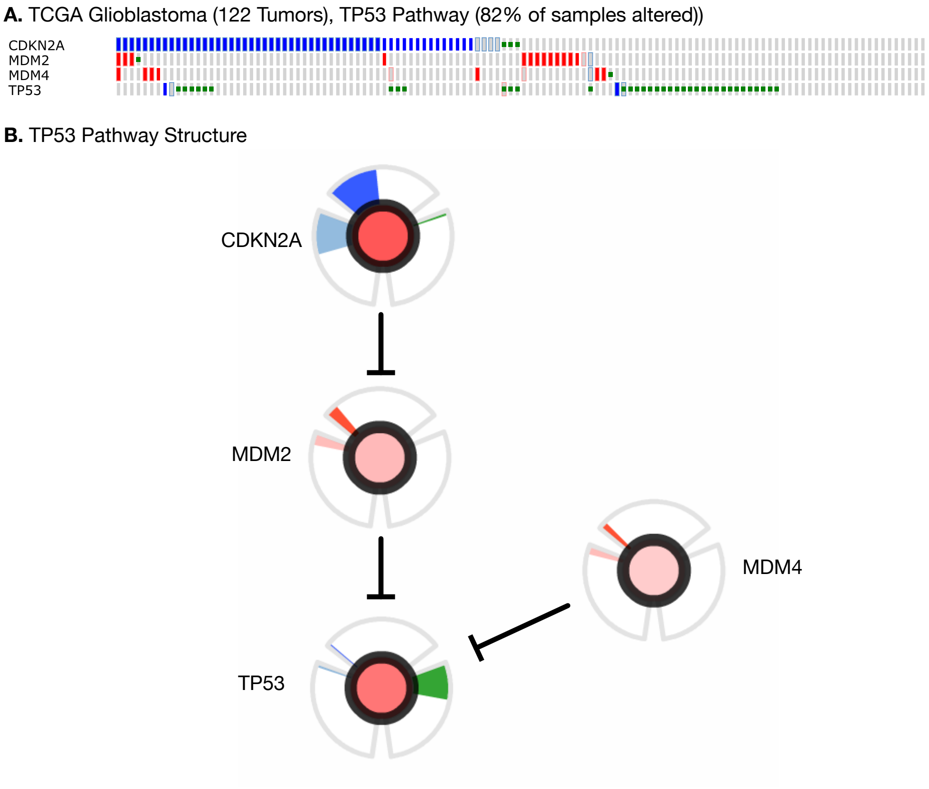


Figure 1.1: TCGA GBM samples show overall alterations across GBM pathway, but changes to specific genes are limited. Part A) shows an OncoPrint profile of the 122 GBMs with complete genomic information. Part B) illustrates the underlying pathway connections that explain why alterations within any single node can deactivate TP53 function. OncoPrint generated from The cBio Cancer Genomics Portal (http://www.cbioportal.org/).

organize and integrate all of the omic features now technically acquirable on virtually any cancer sample. Despite overwhelming evidence that histologically similar cancers are in reality a composite of many molecular subtypes, each with significantly different clinical behavior, this knowledge is rarely applied in practice due to the lack of robust molecular signatures that correlate well with prognosis and treatment options.

High-throughput functional genomics have made tremendous progress in the past decade towards understanding the alterations that lead to disregulation of cellular function [31, 1, 79]. However, the challenges of integrating multiple data sources to identify reproducible and interpretable molecular signatures of tumorigenesis and tumor behavior remain elusive. Recent pilot studies by TCGA and others [56, 74] make it clear that a pathway-level understanding of genomic perturbations is needed to understand the functional changes observed in cancer cells. These findings demonstrate that even when patients harbor genomic alterations or aberrant expression in different genes, these genes often participate in a common pathway (Figure 1.1). In addition, and even more striking, is that the alterations observed (e.g. deletions versus amplifications) often alter the pathway output in the same direction, either all increasing or all decreasing the pathway activation.

Approaches for interpreting genome-wide cancer data have focused on identifying gene expression profiles that are highly correlated with a particular phenotype or disease state, and have led to promising results [78, 20, 2]. Methods using analysis of variance [41], false-discovery [69] and non-parametric methods [77] have been proposed. However, these methods often result in sets of genes that are difficult to generalize between studies and have weak associations with known cellular processes.

Several pathway-level approaches use statistical tests based on overrepresentation of genesets to detect whether a pathway is perturbed in a disease condition. In these approaches, genes are ranked based on their degree of differential activity, for example as detected by either differential expression or copy number alteration. A probability score is then assigned reflecting the degree to which a pathway's genes rank near the extreme ends of the sorted list, such as is used in Gene Set Enrichment Analysis (GSEA) [70]. Other approaches include using a hypergeometric test-based method to identify Gene Ontology [4] or MIPS Mammalian Protein-Protein Interaction [54] categories enriched in differentially expressed genes [71].

Overrepresentation analyses are limited in their efficacy because they do not incorporate known interdependencies among genes in a pathway that can increase the detection signal for pathway relevance. In addition, they treat all gene alterations as equal, which is not expected to be valid for many biological systems. Because of these factors, overrepresentation analyses often miss functionally-relevant pathways whose genes have borderline differential activity. They can also produce many false positives when only a single gene is highly altered in a small pathway.

1.2 Pathway Analysis

Our collective knowledge about the detailed interactions between genes and their phenotypic consequences is growing rapidly. While the knowledge was traditionally scattered throughout the literature and hard to access systematically, new efforts are cataloging pathway knowledge into publicly available databases. Some of the databases that include pathway topology are Reactome [39], KEGG [53], and the NCI Pathway Interaction Database [62]. Updates to these databases are expected to improve our understanding of biological systems by explicitly encoding how genes regulate and communicate with one another. A key hypothesis is that the interaction topology of these

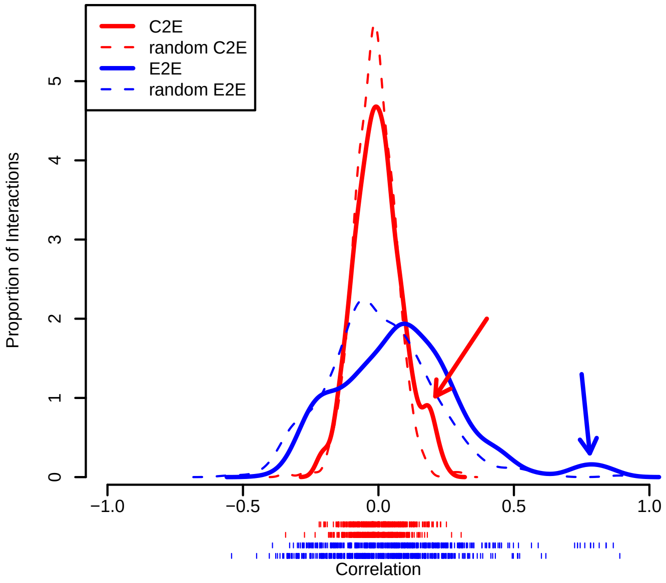


Figure 1.2: NCI Pathway interactions in TCGA GBM data. For all (n=462) pairs where A was found to be an upstream activator of gene B in NCI-Nature Pathway Database, the Pearson correlation (x-axis) computed from the TCGA GBM data was calculated in two different ways. The histogram plots the correlations between the A's copy number and B's expression (C2E, solid red) and between A's expression and B's expression (E2E, blue). A histogram of correlations between randomly paired genes is shown for C2E (dashed red) and E2E (dashed blue). Arrows point to the enrichment of positive correlations found for the C2E (red) and E2E (blue) correlation.

pathways can be exploited for the purpose of interpreting high-throughput datasets.

The hypothesis of pathway-based approaches is that the genetic interactions found in pathway databases carry information for interpreting correlations between gene expression changes detected in cancer. For example, if a cancer-related pathway includes a link from a transcriptional activator A to a target gene T, we expect the expression of A to be positively correlated with the expression of T (E2E correlation). Likewise, we also expect a positive correlation between A's copy number and T's expression (C2E correlation). Further, we expect C2E correlation to be weaker than E2E correlation because amplification in A does not necessarily imply A is expressed at higher levels, which in turn is necessary to upregulate B. In this way, each link in a pathway provides an expectation about the data; pathways with many consistent links may be relevant for further consideration. We tested these assumptions and found that the NCI pathways contain many interactions predictive of the recent TCGA GBM data [52] (Figure 1.2). As expected, C2E correlations were moderate, but had a striking enrichment for positive correlations among activating interactions than expected by chance. E2E correlations were even stronger and similarly enriched. Thus, even in this example of a cancer that has eluded characterization, a significant subset of pathway interactions connect genomic alterations to modulations in gene expression, supporting the idea that a pathway-level approach is worth pursuing.

Until recently, few computational approaches were available for incorporating pathway knowledge to interpret high-throughput datasets. However, several newer approaches have been proposed that incorporate pathway topology [21] . One approach, called Signaling Pathway Impact Analysis (SPIA) [73], uses a method analogous to Google's PageRank to determine the influence of a gene in a pathway. In SPIA, more influence is placed on genes that link out to many other genes. SPIA was successfully applied to different cancer datasets (lung adenocarcinoma and breast cancer) and shown to outperform overrepresentation analysis and Gene Set Enrichment Analysis for identifying pathways known to be involved in these cancers. While SPIA represents a major step forward in interpreting cancer datasets using pathway topology, it is limited to using only a single type of genome-wide data. Newer computational approaches are needed to connect multiple genomic alterations such as copy number variation, DNA methylation, somatic mutations, mRNA and microRNA expression. Integrated pathway analysis is expected to increase the precision and sensitivity of causal interpretations for large sets of observations since no single data source is likely to provide a complete picture on its own.

In the past several years, approaches in probabilistic graphical models (PGMs) have been developed for learning causal networks compatible with multiple levels of observations. Efficient algorithms are available to learn pathways automatically from data [25, 51] and are well adapted to problems in genetic network inference [24]. As an example, graphical models have been used to identify sets of genes that form "modules" in cancer biology [63]. They have also been applied to elucidate the relationship between tumor genotype and expression phenotypes [45], and infer protein signaling networks [60] and combinatorial gene regulatory codes [8].

More recently, a generalization of many existing graphical models (bayesian networks, markov random fields, etc.) called "factor graphs" has become popular in the field [43]. In particular, factor graphs have been used to model expression data [27, 28, 26]. A factor graph is a bipartite graph representing a set of factors, or functions, whose domain is a set of variables and range is the real numbers. By representing pathways as factor graphs, we avoid many issues often seen with bayesian networks (the problem of cycles in particular), and although we have chosen to use probability distributions in this work, we are able to use any function to represent the factors. Very efficient exact and approximate inference methods have been developed that allow graphs of thousands of nodes to be run quickly, making this solution attractive for analyzing thousands of samples in parallel. Armed with this approach, it is possible to determine the critical pathways altered in individual tumor samples and across cancer patient cohorts. In the following chapters, I will describe initial efforts to interpret integrated genomics data and a novel pathway approach that exploits the multiple measurements to provide a powerful tool for understanding misregulated networks in cancer.

Chapter 2

Methods to Integrate Cancer Genomics Data

Prior to 2008, very few tools were available to visualize and analyze cancer genomics data, and the few tools that did exist (Oncomap, Ingenuity) required expensive subscriptions and only handled individual types of data. Given the rich history of UCSC and the success of the genome browser for viewing annotations across genomes, we wanted to establish if a similar experience could be provided for cancer genomics data that would give users access to compare and contrast multiple high throughput genomic datasets. Without the appropriate tools, researchers are typically required to pass data back and forth via Excel or other text-based formats, which makes interpretation difficult. State of the art visualization techniques at the time included clustering and heat map analysis, so we wanted to replicate views that would be familiar to users, and add on top the ability to group by clinical information and compute genome-wide statistics at the push of a button. This chapter describes the creation of the UCSC Cancer Genomics Browser and subsequent analysis pipeline coined the BioIntegrator for interpretation of multiple genomic data points across a set of samples.

2.1 UCSC Cancer Genomics Browser

The UCSC Cancer Genomics Browser is a tool designed to allow hypothesis generating exploration in an intuitive and real-time fashion. The browser was originally designed as an extension to the widely popular UCSC Genome Browser and currently benefits from much of the underlying technology developed for the genome browser. The Cancer Genomics Browser was conceived in collaboration with I-SPY lead investigator Laura Esserman, a breast cancer surgeon at UCSF. The browser was designed to display the multidimensional I-SPY data and allow other researchers in the consortium to manipulate and explore the data through a web browser. In the summer of 2008, the browser was completely overhauled and the interface was re-written as a "Web 2.0" application utilizing asynchronous javascript (AJAX) to dynamically load only the parts of the browser that changed. This overhaul was written as a programming collaboration between myself and Zack Sanborn, with later assistance from Chris Szeto, Jing Zhu, and Larry Meyer. Zack primarily designed and wrote the C-based backend and drawing code, while I designed the interface and wrote the client-side Javascript. The AJAX-powered browser allows for a much more fluid user experience while providing scalability capable of handling hundreds of datasets viewed simultaneously. The browser was published in April of 2008 in Nature Methods [86], but I will discuss many of the features in more detail in the follow sections.

The default view of the UCSC Cancer Genomics browser is the “chromosome view,” which enables researchers to view high-throughput genomic data in the form of a heatmap across a set of patients for the entire genome. The X-axis represents the chromosomal position with the Y-axis representing the set of patients. Thus, a single patient’s set of tumor data across the genome is a single pixel slice of the heatmap as you move from the left to the right at a single Y value. Alongside the genomic heatmap is a clinical heatmap, visually representing key clinical data across the entire patient cohort. These data are generally colored according to the relative values, with bright yellow representing the highest value and black representing the lowest. The sample order is conserved between the genomic and clinical heatmaps in order to facilitate comparison between the two. The browser supports click-to-zoom, as well as click-to-sort functionality allowing researchers to reorder and focus on regions of interest in the genomic heatmap and visually identify clinical attributes that may correlate with the genomic observations. In addition, data for an entire cohort can be visualized in a summary view allowing the user to find regions of conserved copy number or expression data.

The true power in the cancer browser is the ability to simultaneously visualize multiple datasets allowing users to find global patterns that apply between cancers. When we look across three different cancer types (Figure 2.1) at chromosome 9 in the summary view, we see a large blue peak in the middle of the p arm. This peak

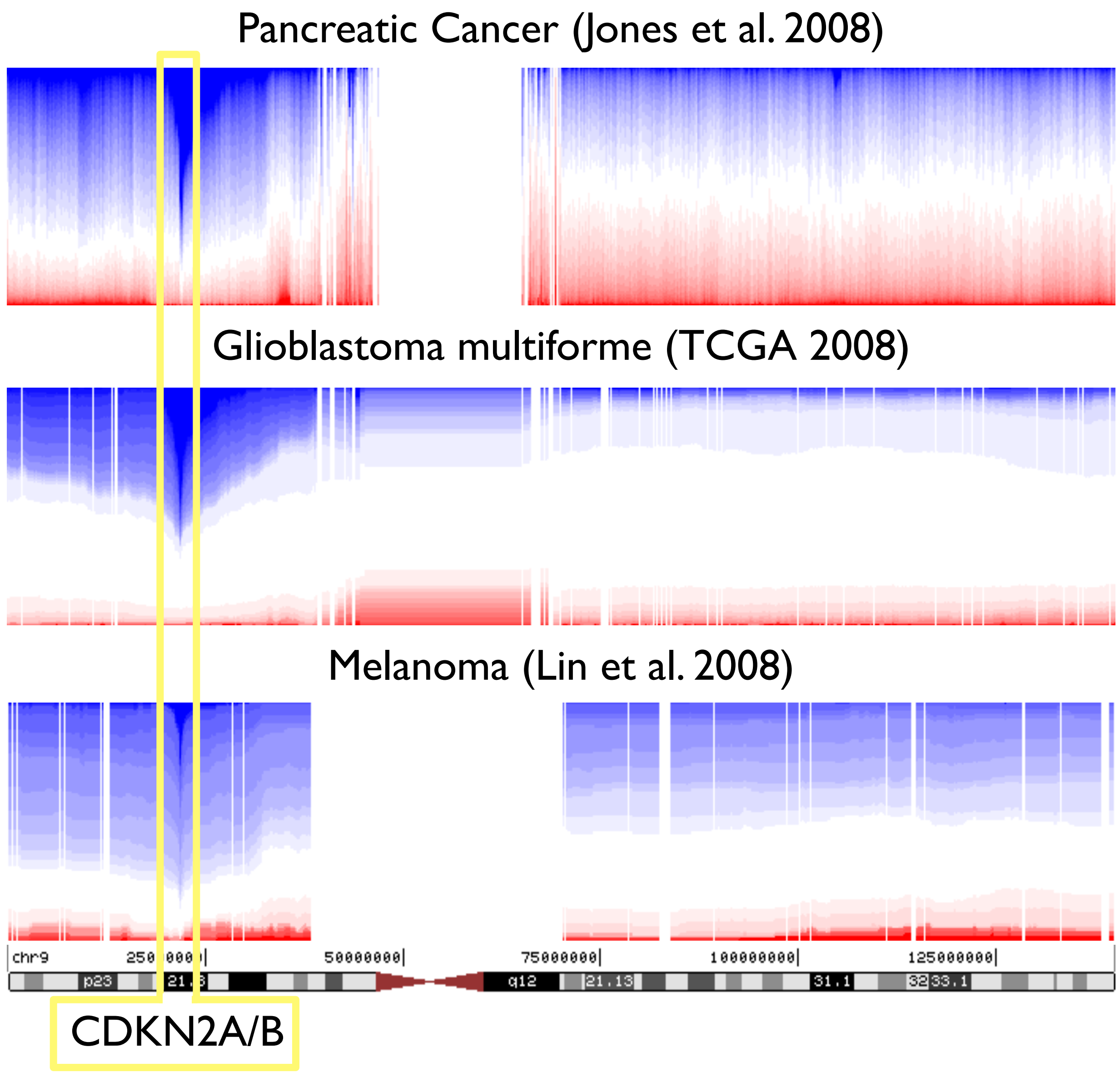


Figure 2.1: UCSC Cancer Genomics Browser proportions view of chromosome 9 across three copy number variation datasets. The yellow box represents the genomic region that is commonly deleted across all three datasets and contains the tumor suppressor CDKN2A/B.

represents concurrent copy number loss across all three datasets and is centered on the gene CDKN2A/B, a well documented tumor suppressor that has been shown to often be deleted in cancer [42]. By visualizing the data across the three datasets, it is possible to identify other regions where there is concurrent loss or gain across a large subset of samples across multiple cancers.

While the chromosome view provides a means to visualize cancer genomics data at the whole-genome or chromosomal level, it is becoming increasingly clear that functional groups of genes (such as those found in a pathway) offer a more useful view into the development of tumorigenesis. Copy number alterations or expression changes in any of several genes in the same pathway can cause equivalent disturbance of a cellular process. Pathways, therefore, provide a more robust and biologically meaningful way to summarize genomic data by grouping genes that may act in a similar fashion. The cancer browser supports a flexible geneset view that allows researchers to visualize differences within and between these pathways across multiple datasets. When we begin exploring the TCGA GBM data using this tool (Figure 2.2) we can easily recognize the systematic alterations to three main pathways across the entire cohort. Indeed, alterations in these three pathways were identified by the TCGA Research Network as being obligatory in most, if not all, glioblastomas. This type of visualization is powerful in helping researchers understand the complex relationships that occur in cancer and disease in general.

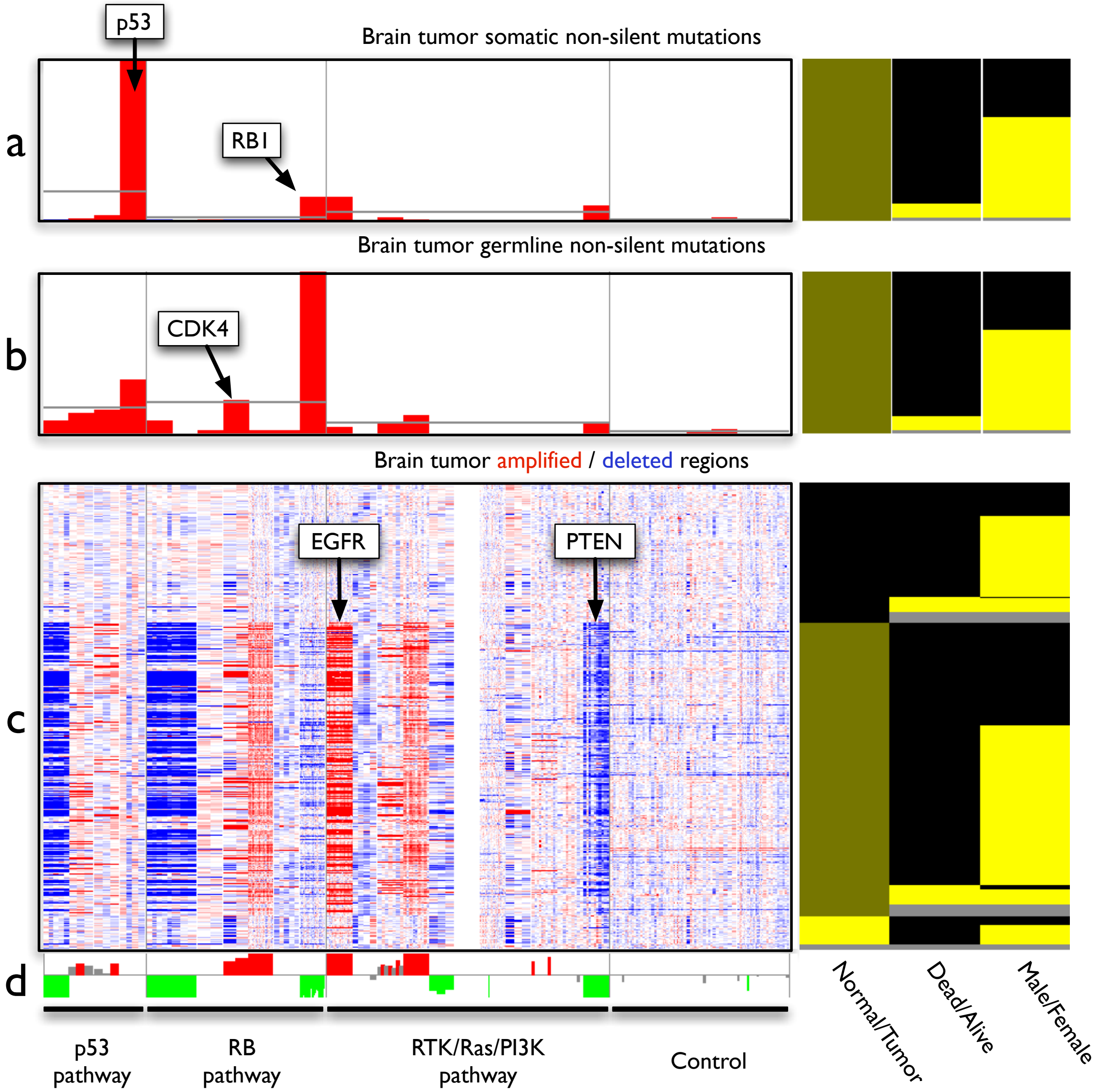
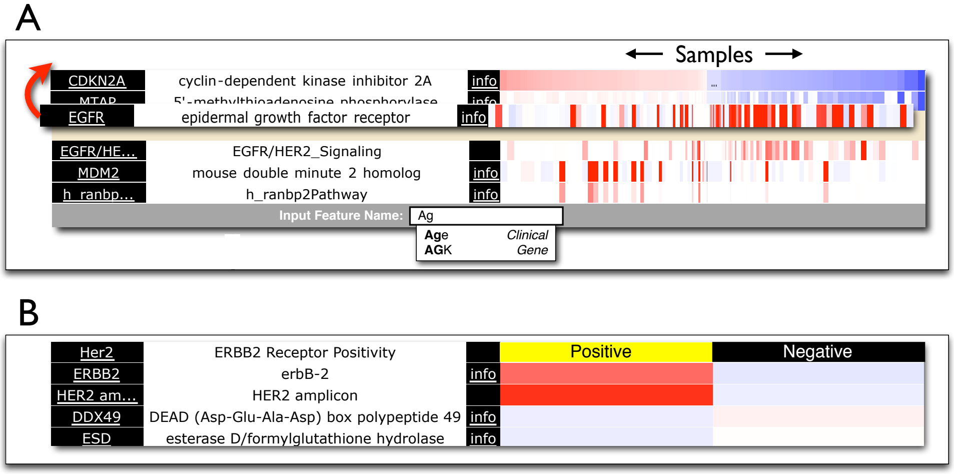


Figure 2.2: Integrative visualization of TCGA glioblastoma multiforme (GBM) genomic data across four separate pathways. (a,b) Histogram of non-silent mutations (e.g. missense, insertion, deletions, etc.) per sample per base in GBM somatic and germ line tissues, respectively. (c) Copy number variation (CNV) in GBM tumor and normal samples. Red or blue color represents amplified or deleted genes. Note that the clinical data to the right of (c) are sorted by tissue type: tumor (black) vs. normal (yellow). The vast majority of copy number alterations seen in the heatmap occur in the tumor samples. (d) A Bonferroni-corrected t-test comparing the distribution of CNV in tumor versus normal samples. A green bar represents a significantly deleted gene ($p < 0.05$, after Bonferroni correction) and a red bar represents a significantly amplified gene in the tumor samples.



CDKN2A	cyclin-dependent kinase inhibitor 2A	info
MTAP	5'-methylthioadenosine phosphorylase	info
EGFR	epidermal growth factor receptor	info
EGFR/HE...	EGFR/HER2_Signaling	
MDM2	mouse double minute 2 homolog	info
h_ranbp...	h_ranbp2Pathway	info

Input Feature Name: Ag

Age	Clinical
AGK	Gene

Her2	ERBB2 Receptor Positivity		Positive	Negative
ERBB2	erbB-2	info		
HER2 am...	HER2 amplicon			
DDX49	DEAD (Asp-Glu-Ala-Asp) box polypeptide 49	info		
ESD	esterase D/formylglutathione hydrolase	info		

Figure 2.3: BioIntegrator web interface, designed to allow users to access both raw and analytical results in an exploratory fashion. Part A highlights the ability to flexibly search for genes and clinical features, as well as drag and drop reordering. Part B shows the dynamic nature of the sorting, allowing discrete clinical features to be placed at top with the data summarized under each label, allowing intuitive visualization of differences.

2.2 BioIntegrator

While the UCSC Cancer Genomics browser is a user-friendly tool for visualization and exploratory analysis, large scale genomic datasets need a more comprehensive analysis process in order to discover complex alterations that drive tumorigenesis. The BioIntegrator is a semi-automated pipeline written in conjunction with Zack Sanborn that can process any data available on the UCSC Cancer Genomic Browser and functions in three main stages. In the first stage, gene-level perturbations are calculated by integrating data from several platforms (copy number, expression, etc.) using a variety of methods. These gene-level perturbations are fed into the second stage where they are combined with the perturbations values of other genes in genesets or structured pathways. Informative gene- and set-level perturbations are then used as features in the final stage to predict sample characteristics (e.g. outcome, response, etc.) using various machine learning techniques. The pipeline is a flexible analysis framework designed to incorporate external analytical algorithms (GISTIC, SPIA, etc.) at any stage in the pipeline, and geneset level analyses are run over thousands of genesets and pathways collected from MSigDB [70], BioCarta, KEGG [53], NCI-Nature [62], and Gene Ontology [4]. Analytical results from each stage are stored in a MySQL relational database that enables rapid access to any slice of patient data, with results being passed to a series of proven machine learning techniques for training and classification. On top of this powerful analytical pipeline, the BioIntegrator offers its own unique web-based visualization allowing researchers to view the results of each stage of the pipeline. The BioIntegrator web-based user interface (Figure 2.3) provides quick and interactive access to both raw and analytical results that allows researchers to explore their data intuitively. Both clinical and genomic data can be explored in a variety of ways, including free-text search and correlation analysis.

By using this BioIntegrator on a set of gene expression and copy number data in melanoma [47], it is possible to identify a set of genes that are modulated in response to a mutation of BRAF in patient samples. Although BRAF mutation status can be determined trivially using sequencing, by asking a machine learning algorithm to try and determine the set of features responsible for splitting those samples we can find a useful set of features associated with BRAF mutation status that may help interpret the underlying biology of BRAF activation. Because BRAF is an important predictive

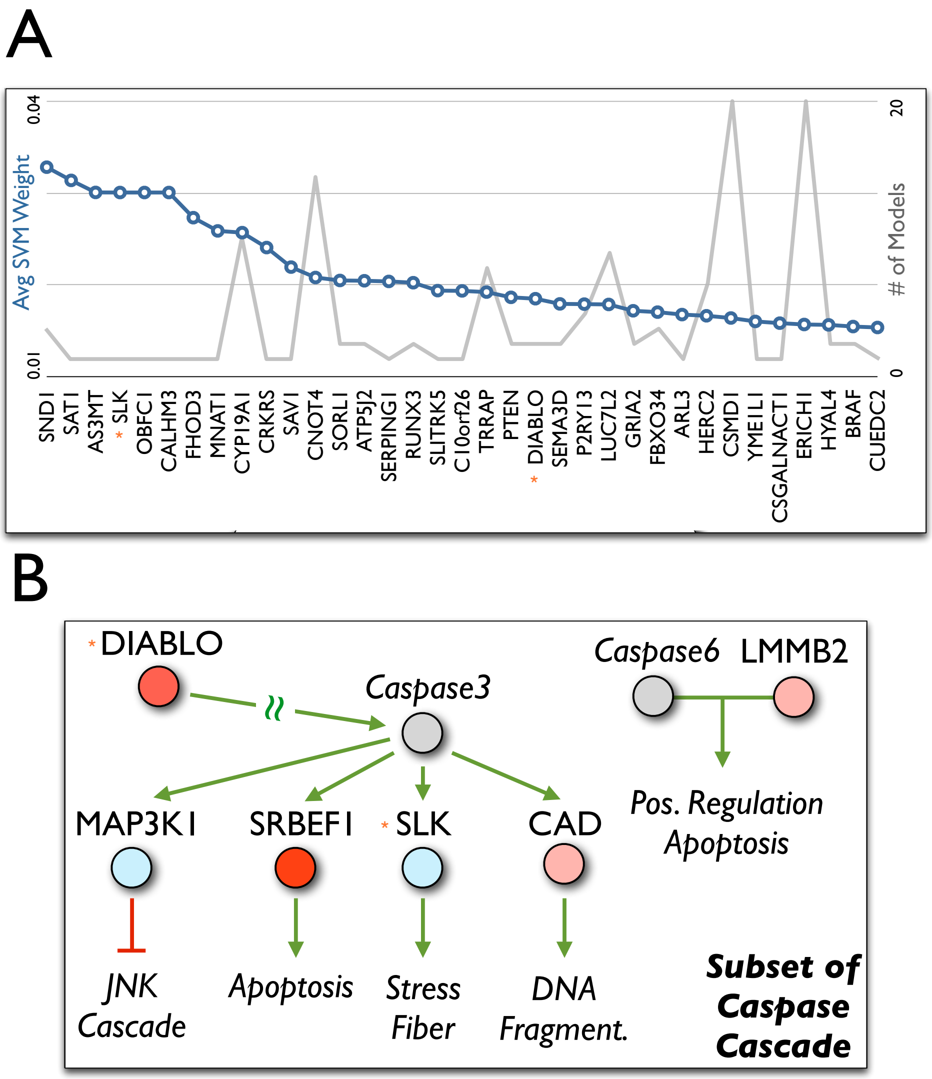


Figure 2.4: Genes were scored independent of sample labels using both gene expression and copy number data from a melanoma cohort [47] in the BioIntegrator. Samples were split 80/20 into training and test sets and feature selection was performed on the training data thresholding on correlation coefficient [15]. Samples were run through SVM-light [38] using default settings, and the top 20 models (minimum 91% accuracy) were selected for downstream analysis. Part A shows the top 35 features selected by the SVMs, with both SVM weight and frequency plotted. When those top 35 features were run through NCI's Pathway enrichment analysis, both Caspase Cascade in Apoptosis ($p < 4e\text{-}6$, Part B) and Fas Signaling Pathway (CD95) ($p < 0.001$) were found to be significantly enriched. The color of nodes in Part B illustrate the average enrichment towards the mutated class (red) or non-mutated class (blue) as calculated by the BioIntegrator gene level perturbation score. Nodes seen in both Part A and Part B are denoted with an orange asterisk.

marker for a class of targeted therapies in melanoma, understanding associated genes may provide a set of features that are functionally equivalent as well as reveal the underlying tumorigenic pathways responsible for sensitivity to the therapy. In fact, when the melanoma samples were run through SVM-light [38] and the corresponding features with the highest weights were analyzed, there was a clear functional enrichment for the caspase cascade in apoptosis ($p < 4e\text{-}6$) and Fas signaling pathway ($p < 0.001$) (Figure 2.4). This result illustrates the added pathway information that the BioIntegrator is capable of providing in an automated fashion.

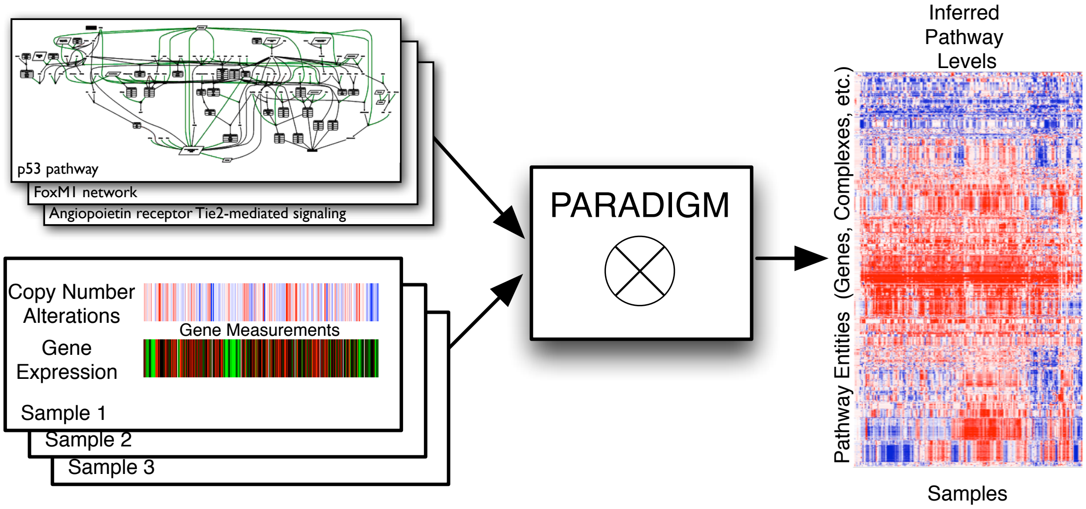


Figure 3.1: Overview of the PARADIGM method. PARADIGM uses a pathway schematic with functional genomic data to infer genetic activities that can be used for further downstream analysis.

Chapter 3

Pathway Analysis Using PARADIGM

In collaboration with Charles Vaske under the advisement of Joshua Stuart, we have developed a sample-specific probabilistic graphical model (PGM) based on factor graphs that we refer to as PARADIGM (PAthway Recognition Algorithm using Data Integration on Genomic Models). This algorithm allows us to infer the activities of genetic pathways from integrated genomic patient data, giving us a more in-depth view of pathways than was capable with tools like the BioIntegrator. Figure 3.1 illustrates the overview of the approach. Multiple genome-scale measurements on a single patient sample are combined to infer the activities of genes, products, and abstract process inputs and outputs for a single NCI pathway. PARADIGM produces a matrix of integrated pathway activities (IPAs) A where A_{ij} represents the inferred activity of entity i in patient sample j. The matrix A can then be used in place of the original constituent datasets to perform downstream analysis, including unsupervised clustering paired with survival analysis[81]. I will discuss aspects of the algorithm in more detail and some of the analysis that has been done using this method in the next subsections.

3.1 Method

The correlations seen in Figure 1.2 clearly show that useful information is encoded in the pathway structures provided by NCI's pathway database, and so we began by convertering each NCI pathway into a distinct probabilistic model that can be used by PARADIGM. A toy example of a small fragment of the p53 apoptosis pathway is shown in Figure 3.2 where we have a pathway diagram from NCI that was converted into a factor graph, including both hidden and observed states. The factor graph integrates observations on gene- and biological process-related state information with a structure describing known interactions among the entities.

To represent a biological pathway with a factor graph, we use variables to describe the states of entities in a cell, such as a particular mRNA or complex, and use factors to represent the interactions and information flow between these entities. These variables represent the *differential* state of each entity in comparison to a “control” or normal level rather than the direct concentrations of the molecular entities. This representation allows us to model many types of high-throughput datasets, such as gene expression detected with DNA microarrays, that often either directly measure the differential state of a gene or convert direct measurements to measurements relative to matched controls. It also allows for many types of regulatory relationships among genes. For example, the interaction describing MDM2 mediating ubiquitin-dependent degradation of p53 can be modeled as activated MDM2 inhibiting p53’s protein level.

The factor graph encodes the state of a cell using a random variable for each entity $X = \{x_1, x_2, \ldots, x_n\}$ and a set of m non-negative functions, or factors, that constrain the entities to take on biologically meaningful values as functions of one another. The j^{th} factor ϕ_j defines a probability distribution over a subset of entities $X_j \subset X$. The entire graph of entities and factors encodes the joint probability distribution over all of the entities as:

$$P(X) = \frac{1}{Z} \prod_{j=1}^{m} \phi_j \left(X_j\right), \tag{3.1}$$

where $Z = \prod_j \sum_{\mathbf{S} \sqsubset X_j} \phi_j(\mathbf{S})$ is a normalization constant and $\mathbf{S} \sqsubset X$ denotes that $\mathbf{S}$ is a “setting” of the variables in X.

Each entity can take on one of three states corresponding to activated, nominal, or deactivated relative to a control level (e.g. as measured in normal tissue) and encoded as 1, 0, or -1 respectively. The states may be interpreted differently depending on the type of entity (e.g. gene, protein, etc). For example, an activated mRNA entity represents overexpression, while an activated genomic copy entity represents more than two copies are present in the genome. Figure 3.2 shows the conceptual model of the factor graph for a single protein-coding gene. For each protein-coding gene G in the pathway, entities are introduced to represent the copy number of the genome (G_{DNA}), mRNA expression (G_{mRNA}), protein level ($G_{protein}$), and protein activity (G_{active}) (ovals labeled "DNA", "mRNA", "protein", and "active" in Figure 3.2). For every compound, protein complex, gene family, and abstract process in the pathway, we include a single variable with molecular type "active." While the example in Figure 3.2 shows only one process ("Apoptosis"), in reality many pathways have multiple such processes that represent everything from outputs (e.g. "Apoptosis" and "Senescence") to inputs (e.g. "DNA damage") of gene activity.

In order to simplify the construction of factors, we first convert the pathway into a directed graph, with each edge in the graph labeled with either positive or negative influence. First, for every protein coding gene G, we add edges with a label "positive" from G_{DNA} to G_{mRNA}, from G_{mRNA} to $G_{protein}$, and from $G_{protein}$ to G_{active} to reflect the expression of the gene from its number of copies to the presence of an activated form of its protein product. Every interaction in the pathway is converted to a single edge in the directed graph.

Using this directed graph, we then construct a list of factors to specify the factor graph. For every variable x_i, we add a single factor $\phi(X_i)$, where $X_i = \{x_i\} \cup \{\text{Parents}(x_i)\}$ and $\text{Parents}(x_i)$ refers to all the parents of x_i in the directed graph. The value of the factor for a setting of all values is dependent on whether x_i is in agreement with it's expected value due to the settings of $\text{Parents}(x_i)$. For this study, the expected value was set to the majority vote of the parent variables. If a parent is connected by a positive edge it contributes a vote of $+1$ times its own state to the value of the factor. Conversely, if the parent is connected by a negative edge, then the variable votes -1 times its own state. The variables connected to x_i by an edge labeled "minimum" get a single vote, and that vote's value is the minimum value of these variables, creating an AND-like connection. Similarly the variables connected to x_i by an edge labeled "maximum" get a single vote, and that vote's value is the maximum value of these variables, creating an OR-like connection. Votes of zero are treated as abstained votes. If there are no votes the expected state is zero. Otherwise, the majority vote is the expected state, and a tie between 1 and -1 results in an expected state of -1 to give more importance to repressors and deletions.

Given this definition of expected state, $\phi_i(x_i, \text{Parents}(x_i))$ is specified as:

$$\phi_i(x_i, \text{Parents}(x_i)) = \begin{cases} 1 - \epsilon & x_i \text{ is the expected state from Parents}(x_i) \\ \frac{\epsilon}{2} & \text{otherwise.} \end{cases}$$

For the results shown here, ϵ was set to 0.001, but orders of magnitude differences in the choice of epsilon did not significantly affect results.

Finally, we add observation variables and factors to the factor graph to com-

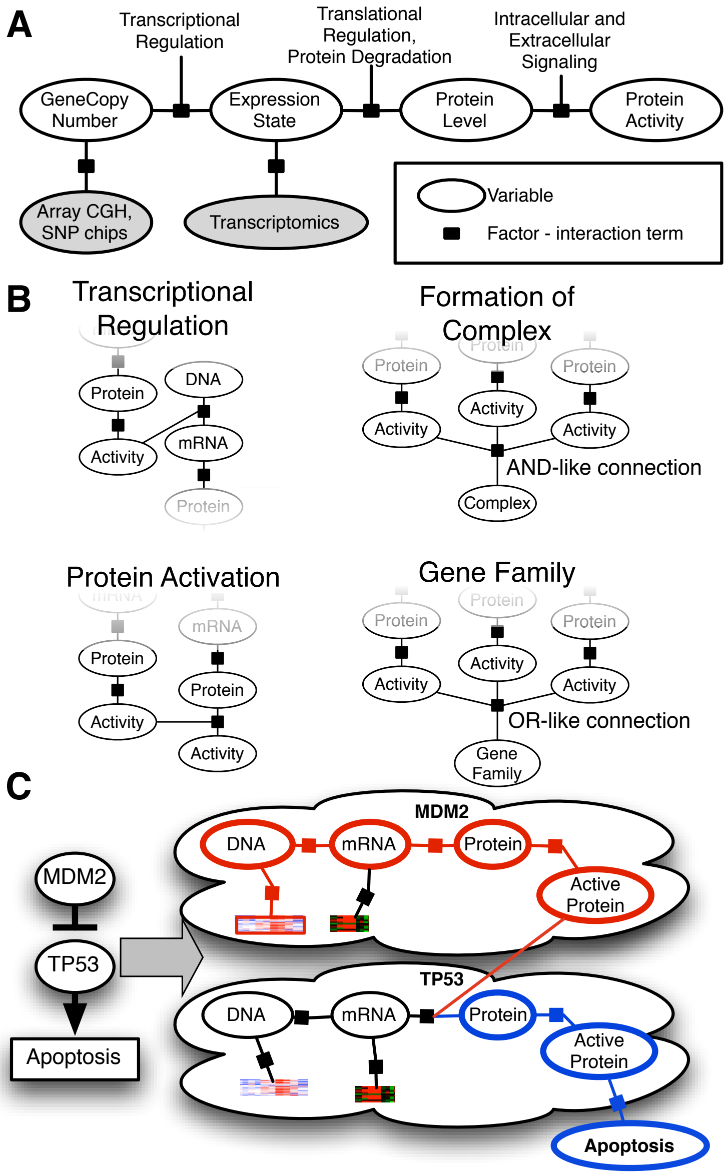


Figure 3.2: Conversion of a genetic pathway diagram into a PARADIGM model. **A**. Data on a single patient is integrated for a single gene using a set of four different biological entities for the gene describing the DNA copy number, mRNA and protein levels, and activity of the protein. **B**. PARADIGM models various types of interactions across genes including transcription factors to transcript and protein targets (upper-left), protein subunits aggregating in a complex (upper-right), protein post-translational modification (lower-left), and sets of gene products in a family performing redundant functions (lower-right). **C**. Toy example of a small sub-pathway involving P53, its inhibitor MDM2, and the high level P53 determined cell process, apoptosis as represented in the model.

plete the integration of pathway and multi-dimensional functional genomics data (Figure 3.2). Each discretized functional genomics dataset is associated with one of the molecular types of a protein-coding gene. Array CGH/SNP estimates of copy number alteration are associated with the “genome” type. Gene expression data is associated with the “mRNA” type. Though not presented in the results here, future expansion will include DNA methylation data with the “mRNA” type, and proteomics and gene-resequencing data with the “protein” and “active” types. Each observation variable is also ternary valued. The factors associated with each observed type of data are shared across all entities and the associated parameters are learned from the data using a standard Expectation Maximization (EM) procedure[19].

3.2 Comparisons

In order to assess the viability of the integrated factor graph approach, PARADIGM was compared to three competing methods across three biological scenarios, done in collaboration with Vinay Varadan, Prateek Mittal, and Charles Vaske[80]. An initial, more detailed comparison between SPIA and PARADIGM was done using two datasets spiked with decoys, designed as a more conservative comparison technique. The first two biological scenarios focus on the identification of pathways that differentiate between two pre-defined cohorts, while the third scenario identifies pathways that can best stratify patient survival. For these applications, we considered two datasets corresponding to ovarian and breast carcinoma. Copy number data (Agilent 244K CGH platform), gene expression data (Affymetrix U133A platform), and associated patient clinical information was obtained from the high-grade serous ovarian carcinoma dataset at TCGA. Overall, we used 423 ovarian cancer samples and 8 normal samples. The breast cancer data for a total of 113 patients was obtained from two public sources - copy number data derived from the publication by Chin et al. (GEO accession GPL5737) and the gene expression data for the same samples from the publication by Naderi et al. (MIAMIExpress accession E-UCon-1). Pathway information was derived from the NCI-PID database. The spiked datasets were generated from the same breast cancer data and the glioblastoma multiforme data available from TCGA. The implementations of GSEA and PathOlogist were obtained from their respective authors' websites. All of these implementations ship with inbuilt support for NCI-PID pathways. The official SPIA implementation works only with KEGG pathways, so we used a modified implementation developed by Dent Earl that supports NCI-PID pathways in order to provide a fair comparison between these approaches.

3.2.1 Distinguishing True Networks From Decoys

Because of the similarities in approach between SPIA and PARADIGM, we first asked whether the integrated activities could be obtained from arbitrary genes connected in the same way as the genes in the NCI pathways in each method. To do this, we estimated the false discovery rate and compared it to SPIA in the breast cancer and TCGA GBM cohorts. Because many genetic networks have been found to be implicated in cancer, we chose to use simulated ”decoy” pathways as a set of negative

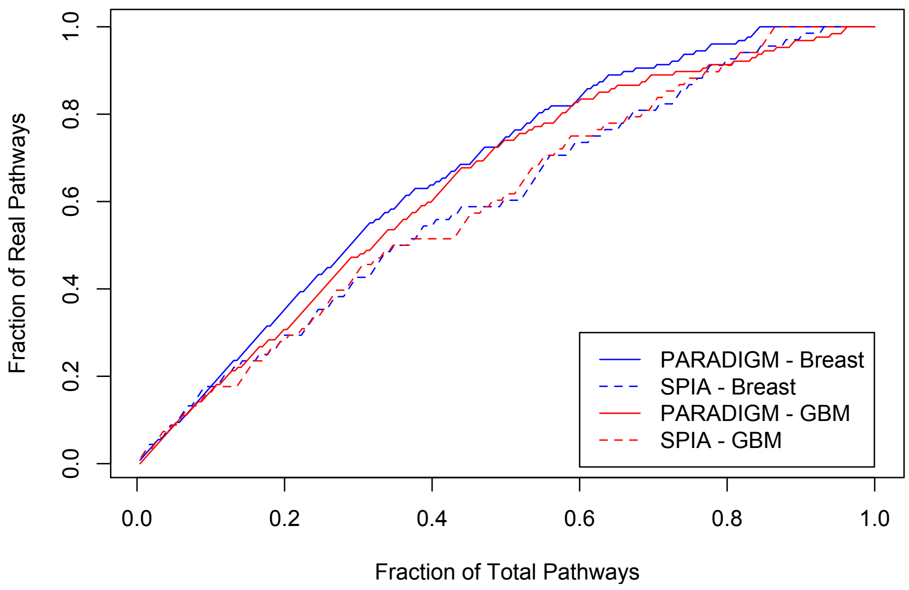


Figure 3.3: Distinguishing decoy from real pathways with PARADIGM and SPIA. Decoy pathways were created by assigning a new gene name to each gene in a pathway. PARADIGM and SPIA were then used to compute the perturbation of every pathway. Each line shows the receiver-operator characteristic (ROC) for distinguishing real from decoy pathways using the perturbation ranking. In breast cancer, the areas under the curve (AUCs) are 0.669 and 0.602 for PARADIGM and SPIA, respectively. In GBM the AUCs are 0.642 and 0.604 respectively.

controls. For each NCI pathway, we constructed a decoy pathway by connecting random genes in the genome together using the same network structure as the NCI pathway. We then ran PARADIGM and SPIA to derive IPAs for both the NCI and decoy pathways. For PARADIGM, we ranked each pathway by the number of IPAs found to be significant across the patients after normalizing by the pathway size. For SPIA, pathways were ranked according to their computed impact factor.

We found that PARADIGM excludes more decoy pathways from the top-most activated pathways compared to SPIA (Figure 3.3). For example, in breast cancer, PARADIGM ranks 1 decoy in the top 10, 2 in the top 30, and 4 in the top 50. In comparison, SPIA ranks 3 decoys in the top 10, 12 in the top 30, and 22 in the top 50. The overall distribution of ranks for NCI IPAs are higher in PARADIGM than in SPIA, observed by plotting the cumulative distribution of the ranks ($P < 0.009$, K-S test).

3.2.2 Tumor versus Normal - Pathways associated with Ovarian Cancer

One of the first applications of PARADIGM to a large-scale cohort with adjacent normal tissue was in collaboration with The Cancer Genome Atlas during the analysis of high-grade serous adenocarcinoma. In serous ovarian cancer, the FOXM1 transcription factor network was found to be differentially altered in the 330 tumor samples compared to the normal controls in the highest proportion of the patient samples (Figure 3.4). Pathways with recurrently high IPAs reveal fundamental mechanisms dysregulated in serous ovarian cancer, particularly interesting because of the aggressive

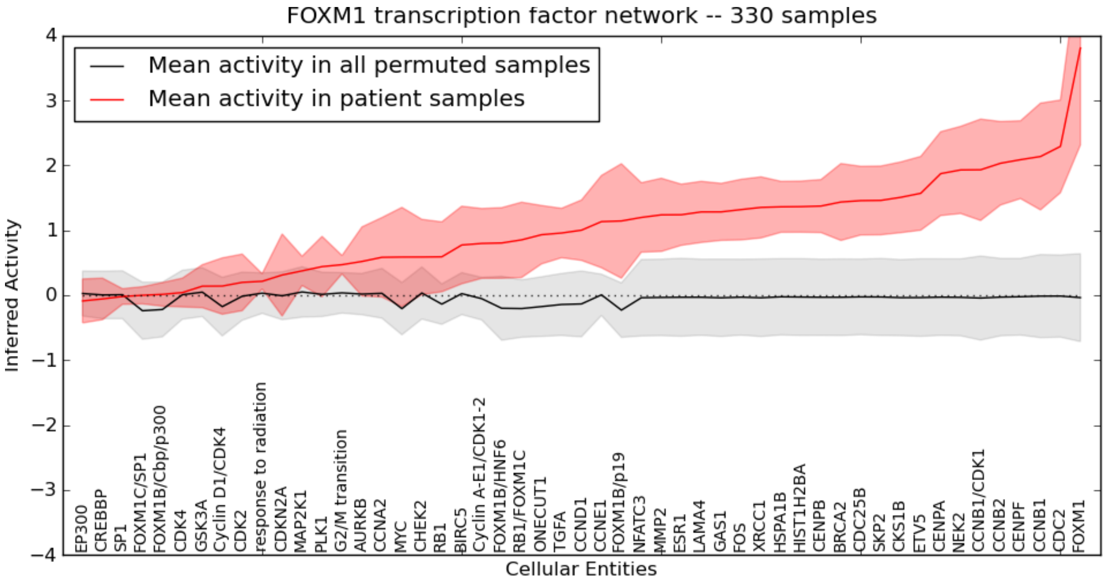


Figure 3.4: Summary of FOXM1 integrated pathway activities (IPAs) across all samples. The arithmetic mean of IPAs across tumor samples for each entity in the FOXM1 transcription factor network is shown in red, with heavier red shading indicating two standard deviations. Gray line and shading indicates the mean and two standard deviations for IPAs derived from permuted data.

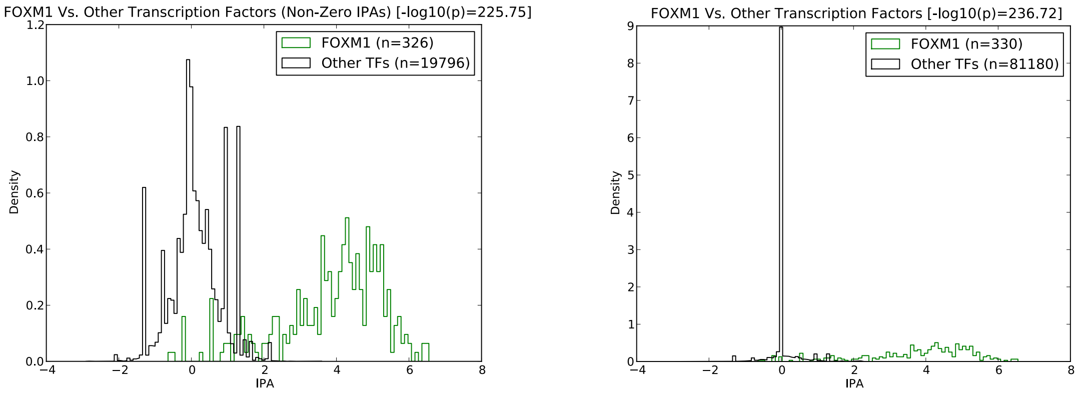


Figure 3.5: Summary of IPAs of FOXM1 and all tested transcription factors. Histograms of each sample's IPA in FOXM1 or in any other transcription factor from the PID. A. Histogram of IPAs, with IPAs of 0 removed. By a Kolmogorov-Smirnov test, FOXM1 shows a different distribution with a p-value $< 10^{-225}$. B. Histogram IPAs including 0 valued IPAs. By a KS test, FOXM1 has a different distribution of IPAs with a p-value $< 10^{-236}$.

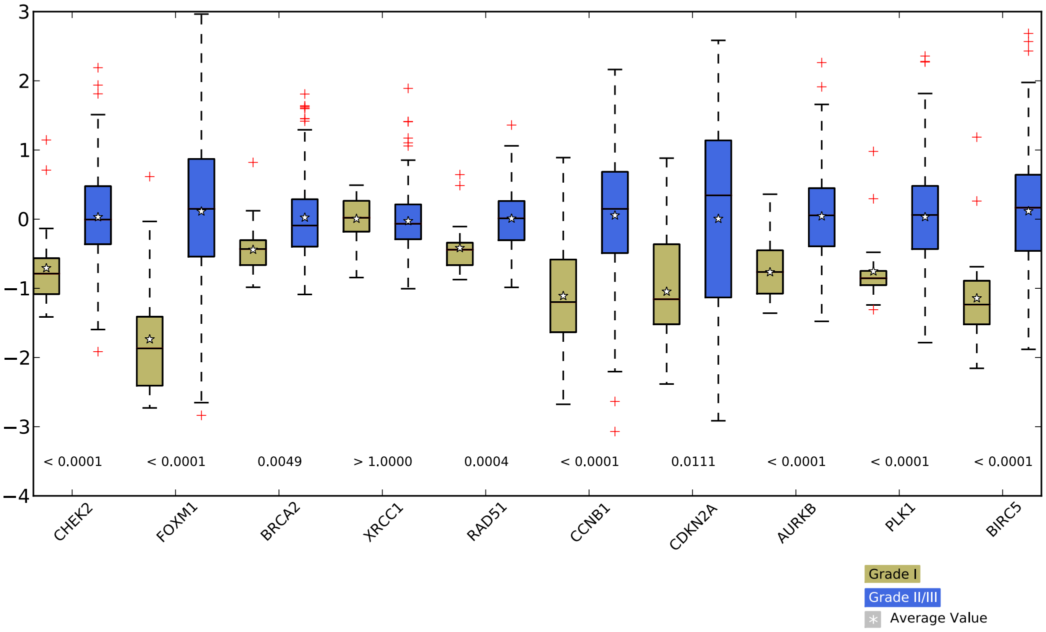


Figure 3.6: Expression of FOXM1 Transcription Factor Network genes in high grade vs. low grade carcinoma. Using data from Etemadmoghadam et al (2009) we plotted expression of genes in the FOXM1 transcription network in either low grade (I) or high grade (II/III) ovarian carcinomas. A student's t-test was used to assess the significance of differential expression between high grade and low grade, with the p-value below each pair of boxplots.

nature of the disease. In order to validate the highly activated nature of FOXM1, we asked whether the IPAs of the FOXM1 transcription factor were more highly altered than other transcription factors. We compared the FOXM1 level of activity to all of the other 202 transcription factors in the NCI-PID to validate the signature was not an artifact of the highly-connected nature of transcription factors in general. Even compared to other transcription factors in the NCI set, the FOXM1 transcription factor had significantly higher levels of activity ($p < 0.0001$) suggesting further that it may be a important signature (Figure 3.5).

Because the entire cohort for the TCGA ovarian were high-grade serous, we asked whether the FOXM1 signature was specific to high-grade serous or a more general marker of both low- and high-grade. To determine whether the signature is associated with serous ovarian cancer, we obtained the expression of FOXM1 and several of its targets from the dataset of [22] in which both low- and high-grade ovarian tumors had been transcriptionally profiled. This independent data confirmed that FOXM1 and several of its targets are significantly up-regulated in serous ovarian, even relative to low-grade ovarian cancers (Figure 3.6).

We next wanted to understand how this result compares with competing methods using the TCGA ovarian cancer dataset[13]. Using a larger set of 423 ovarian cancer samples against 8 normal samples, we attempted to identify pathways that were dysregulated in ovarian cancer using GSEA, PathOlogist, SPIA and Paradigm. We used a FDR cutoff of 0.25 to capture only the statistically significant pathways. Table 3.1 presents the top pathways that were found to be upregulated in ovarian cancer samples when compared to adjacent normal ovarian tissue. The FOXM1 transcription factor network from the NCI-PID is consistently found to be upregulated by all four methodologies. Figure 3.7 captures the level of activity in this pathway using PathOlogist's activity metric across all the ovarian samples, while all the normal samples have low activity levels, a majority of the ovarian cancer samples show high activity levels.

FOXM1's role in many different cancers, including breast and lung, has been well documented but its role in ovarian cancer has not been investigated. FOXM1 is a multifunctional transcription factor with 3 known dominant splice forms, each regulating

Table 3.1: The FOXM1 transcription factor network shows up as being upregulated in ovarian tumors compared to normal consistently across pathway methodologies.

GSEA	Pathologist		SPIA	PARADIGM
	Activity	Consistency		
Validated targets of C-MYC transcriptional activation	Signaling by Aurora Kinases	Aurora A signaling	**FOXM1 transcription factor network**	Influence of Ras and Rho proteins on G1 to S transition
E2F transcription factor network	**FOXM1 transcription factor network**	Signaling events mediated by PRL	PLK1 signaling events	IL2-Mediated signaling events
FOXM1 transcription factor network	PLK1 signaling events	PRC2 complex sets long-term gene silencing	AuroraB Signaling	IL2 signaling events mediated by PI3K
FOXO family signaling	AuroraB signaling	C-MYB transcription factor network		E2F Transcription factor network
AuroraB signaling	BARD1 signaling events	Regulation of cytoplasmic and nuclear SMAD2/3 signaling		c-MYC Pathway
PLK1 signaling events	Class I PI3K signaling events	Signaling events mediated by VEGFR1 and VEGFR2		BCR signaling pathway
	Signaling Events Mediated by VEGFR1and VEGFR2	mTOR signaling pathway		Signaling events mediated by PRL
	A6B1 and A6B4 integrin signaling	A6B1 and A6B4 integrin signaling		Signaling events mediated by TCPTP
		IGF1 pathway		**FOXM1 transcription factor network**

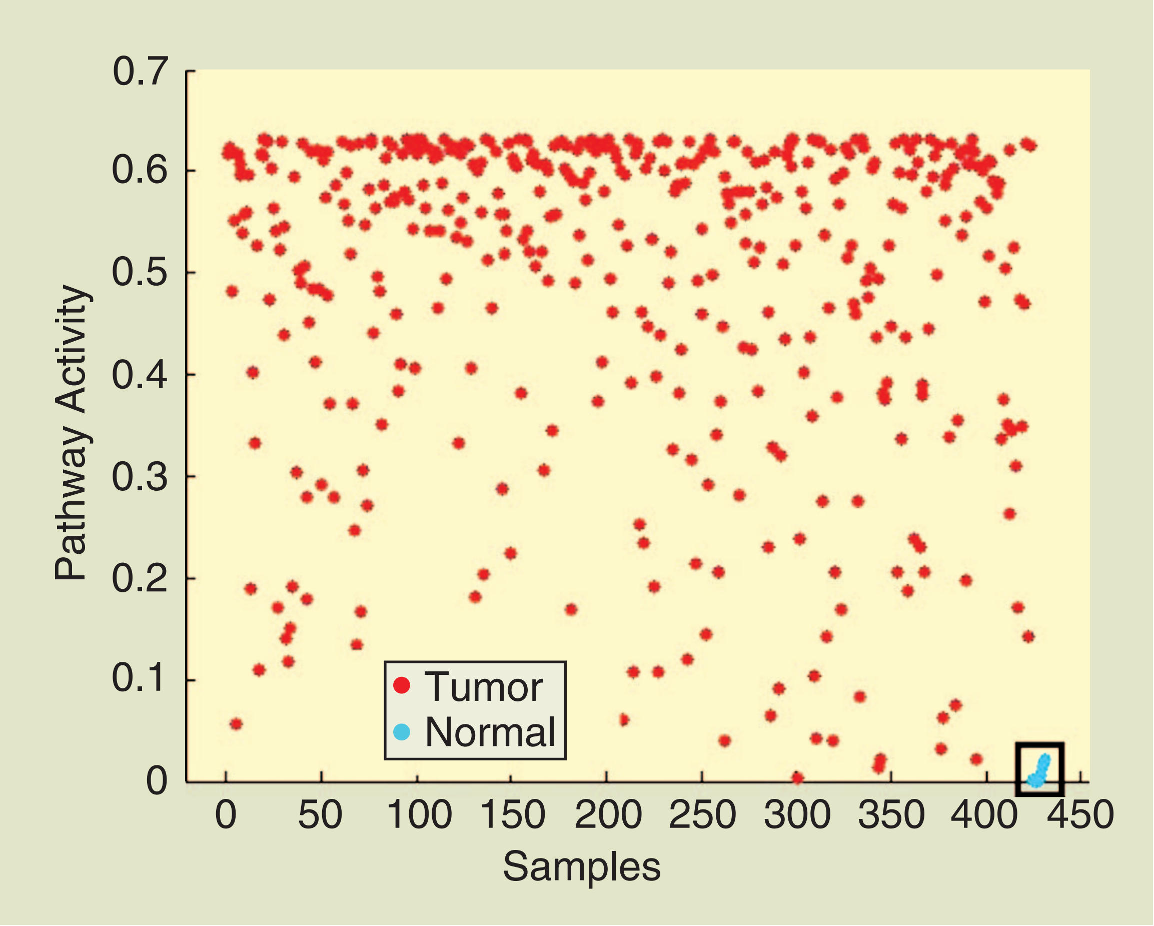


Figure 3.7: The PathOlogist approach shows significant activation of the FOXM1 pathway in tumor samples. (©2012 IEEE)

distinct subsets of genes with a variety of roles in cell proliferation and DNA repair. The FOXM1c isoform directly regulates several targets with known roles in cell proliferation including AUKB, PLK1, CDC25, and BIRC5 [48]. On the other hand, the FOXM1b isoform regulates a completely different subset of genes that include the DNA repair genes BRCA2 and XRCC1 [72]. CHEK2, which is under indirect control of ATM, directly regulates FOXM1's expression level. In addition to increased expression of FOXM1 in most of the ovarian patients, a small subset also have increase copy number amplifications detected by GISTIC ($< 5\%$). Thus the alternative splicing regulation of FOXM1 may be involved in the control switch between DNA repair and cell proliferation. Furthermore, recent evidence has shown FOXM1 is a target of p53-mediated repression [6], which is consistent with P53 mutational rates close to 100% in these samples. The observation that PARADIGM detected the highest level of altered activity centered on this transcription factor suggests that FOXM1 resides at a critical regulatory point in the cell, and the mutational rates of P53 may offer an explanation for the consistency of the signal across all the samples.

It is interesting to note here that PLK1 signaling events and AuroraB signaling were also identified as upregulated in ovarian cancers by the GSEA and PathOlogist methodologies. Since FOXM1 is known to positively regulate both PLK1 and Aurora B, this is consistent with the FOXM1 network being upregulated[57]. Another pathway consistent with the hypothesis of increased cell proliferation, the c-Myc pathway was found by both GSEA and Paradigm. Amplification of c-Myc is a common event in HGSOCs and a recent study has shown that c-Myc transformation is sufficient to induce oncogenicity in normal fallopian tube tissue[40]. Similarly Signaling events mediated by VEGFR1 and VEGFR2 and mTOR signaling was identified by PathOlogist, but not by GSEA. This is again consistent with prior studies that have reported that the VEGF pathway via VEGFR2 stimulates the AKT/mTOR pathway in ovarian cancer[76].

This kind of consistency and interpretability of the gene expression changes in tumor samples compared to normal tissues across multiple pathways highlights the explanatory power of the pathway-based methodologies. However, the fact that some pathways listed in Table 3.1 occur only in one of the methodologies illustrates the complex nature of this problem. It is an open question whether all of these pathways are indeed biologically relevant to ovarian cancer, or whether some of these pathways were false positives, identified only as a result of some assumptions underlying the specific the computational methodology.

3.2.3 Differentially Regulated Pathways in ER+ve vs ER-ve breast cancers

Our second application involves identifying pathways associated with specific subtypes of breast cancer. Estrogen receptor positivity is a major facet of luminal breast cancers and is typically associated with better prognosis and responsive to tamoxifen treatment. Understanding the molecular mechanisms that differentiate estrogen receptor positive (ER+ve) breast cancers from other subtypes could help identify potential modulators of cancer aggressiveness and novel therapeutic targets. We looked at pathways that were differentially regulated in 74 ER+ve samples when compared to the

Table 3.2: FOXA1 transcription factor network is upregulated in ER+ve tumors whereas the Hif-1-alpha transcription network and Syndecan-1 mediated signaling events are down-regulated.

GSEA	Pathologist		SPIA	PARADIGM
	Activity	Consistency		
FOXA1 transcription factor network	C-MYB transcription factor network	EPO signaling pathway	Cellular roles of Anthrax toxin	FOXA1 transcription factor network
Validated targets of C-myc transcriptional activation	EPO signaling pathway	C-MYB transcription factor network	Glucocorticoid receptor regulatory network	Syndecan-1-mediated signaling events
Hiv-1 nef: negative effector of fas and tnf	CXCR3-mediated signaling events	IL6-mediated signaling events	E2F transcription factor network	C-MYB transcription factor network
AuroraB signaling	IFN-gamma pathway	ARF6 trafficking events	Hif-1-alpha transcription factor network	Erbb receptor signaling network
Hif-1-alpha transcription factor network	Syndecan-1 mediated signaling events	PLK3 signaling events	IL4 mediated signaling events	Hif-1-alpha transcription factor network
	PLK1 signaling events	FAS signaling pathway (cd95)	ErbB1 downstream signaling	
	Class I Pi3K signaling events	FOXO family signaling		
	Caspase cascade in apoptosis	FOXA1 transcription factor network		
	TCR signaling in nave CD4+ T cells	Proteogylcan Syndecan-mediated signaling events		

39 ER-ve tumor samples in the Chin-Naderi cohort. Table 3.2 depicts the comparative analysis of the pathways identified using the four pathway-based methodologies. We can see from Table 3.2 that FOXA1 transcription factor network is identified as upregulated in ER+ve tumors by the GSEA, PathOlogist and Paradigm methodologies. FoxA1 is a key determinant of estrogen receptor function, and has been shown to influence differential interactions between ER and chromatin, thus mediating the transcriptional activity of ER and the action of tamoxifen, which is the frontline targeted therapy for ER+ve breast cancer patients[36]. FoxA1 expression is also correlated with the lumina A subtype of breast cancer[5], and among the ER-positive subgroup treated with tamoxifen, FOXA1 was found to be an independent prognostic marker whose expression was associated with low risk of recurrence[49].

3.2.4 Therapy response prediction using pathways (Platinum Free Interval in Ovarian Cancer)

In our third application, we demonstrate an example of how pathway based methodologies can be used to predict patient survival outcome. For this analysis, we use PathOlogist and Paradigm on the ovarian cancer dataset. We first compute pathway activity/consistency levels and gene IPAs for all tumor samples using PathOlogist and Paradigm respectively, and then use clustering algorithms on these metrics for each pathway to group patients. In case of PathOlogist, for each pathway, we use the k-means clustering algorithm on the patient pathway activity/consistency levels to cluster patient samples into two groups. This approach works best for the one-dimensional nature of

Table 3.3: Pathways associated with platinum-free interval in ovarian carcinoma.

Pathologist				PARADIGM	
Activity		Consistency			
Pathway	P-value (FDR)	Pathway	P-value (FDR)	Pathway	P-value (FDR)
Downstream signaling in naive cd8+ t-cells	0.018 (0.78)	Alk1 signaling events	0.023 (0.79)	Stablization and expansion of E-cadherin adherens junction	0.003 (0.21)
Alk1 signaling events	0.023 (0.78)	Hif-2-alpha transcription factor	0.036 (0.79)	Arf6 signaling events	0.020 (0.72)

PathOlogist Activity scores, which only provides a single score value per pathway. In case of Paradigm, for each pathway, we use hierarchical clustering with average linkage on all gene IPAs belonging to that pathway to cluster patients into two groups. This approach better accounts for the two-dimensional nature of the Paradigm results which provide a score per feature per sample within each pathway. Platinum free survival data was available for a total of 113 samples in the ovarian dataset and was associated with the pathway-based clustering of the patients. Pathways that led to clusters with less than 30 samples were eliminated from consideration for survival analysis. Kaplan-Meier curves were estimated for each cluster of the selected pathways and the differences between survival curves were estimated using the Mantel-Haenszel test.

Stratifying ovarian cancer patients into responders or non-responders using genomic features is known to be an extremely difficult problem, with no reliable predictor currently available in the clinic. This is reflected even in the pathway analysis, as can be seen from Table 3.3. PathOlogist did not find any significant pathways after correcting for multiple testing, while Paradigm found just one pathway that had marginal

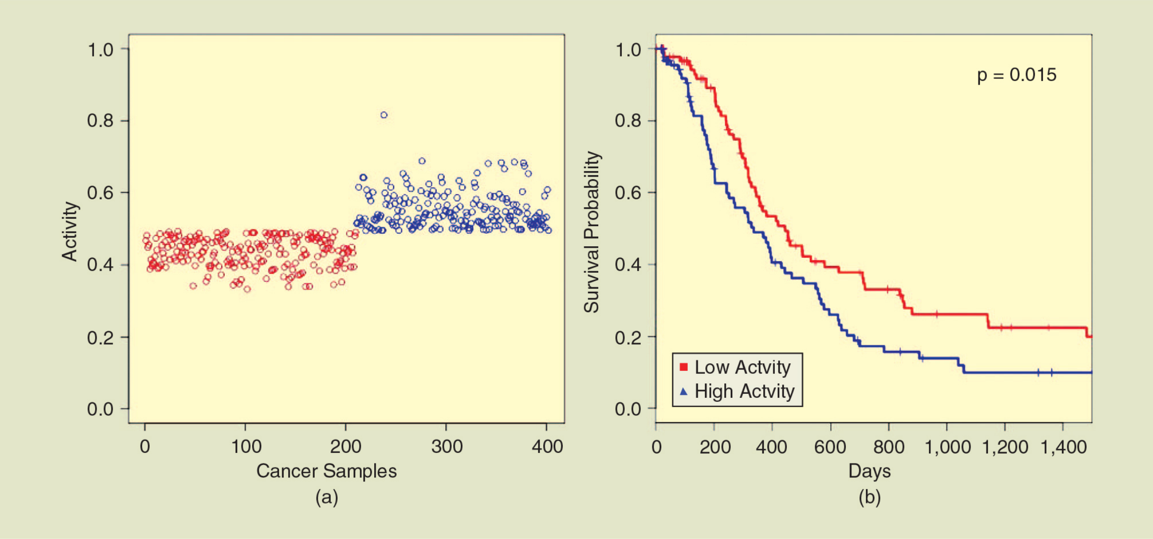


Figure 3.8: PathOlogist reveals that patients with higher activity levels in the Downstream signaling in naive cd8+t cells pathway are associated with improved response to platinum therapy. (©2012 IEEE)

stratifying capability in terms of differences in the platinum free interval. Figure 3.8 captures PathOlogist's top pathway activity level across all patients, and the corresponding k-means clustering of these samples into two clusters. Figure 3.8B depicts the Kaplan-Meier survival curves associated with these clusters. The relative differences in platinum free interval is quite small, as already highlighted in Table 3.3.

Paradigm successfully found a single pathway as being significant after multiple testing to identify patient clusters with different response outcomes to platinum therapy - the E-cadherin adherens junction pathway. Figure 3.9 depicts the dendrogram corresponding to the hierarchical clustering analysis and Figure 3.9 shows the Kaplan-Meijer curves for these clusters. We find that patients with lower IPAs in the genes associated with the E-cadherins adherens junction pathway have a better response to

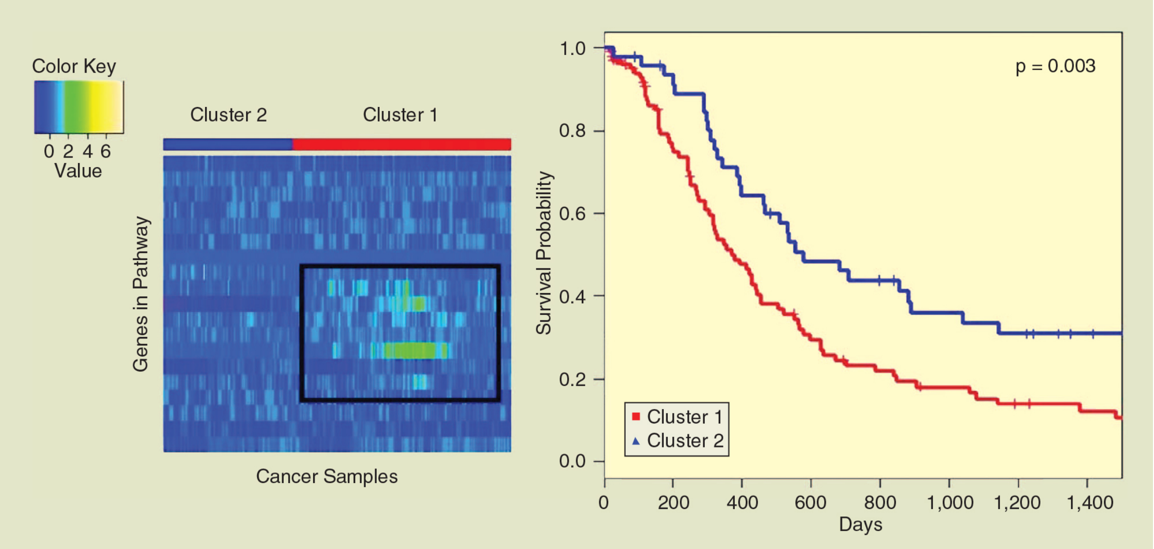


Figure 3.9: Clustering analysis of ovarian samples using Paradigm shows that that patients with lower IPA's for genes in the E-cadherin adherens junction pathway are associated with better response to platinum therapy (p=0.003). (©2012 IEEE)

platinum therapy (p= 0.003). E-cadherin has been rather controversial as a prognostic factor in serous ovarian cancer with some studies identifying it as being only marginally associated with prognosis[83], while other studies pointing to a deeper role suggestive of a novel subtype of serous ovarian carcinoma harboring a mesenchymal phenotype[75]. This particular subtype of serous ovarian carcinoma was associated with slightly improved relapse-free survival, which would correspond to our finding of slightly better response to platinum therapy. These results suggest that a pathway-level framework is likely to provide deeper insights on mechanisms underlying clinically-relevant subtypes when compared to evaluating the expression levels of just one or more genes, even if they were chosen from within the same pathway.

3.3 Unsupervised Stratification of Cancer Patients by Pathway Activities

While finding consistently altered pathways across sample cohorts can provide clues to globally misregulated processes within known cancer subtypes, it does not help us find new pathways that might successfully stratify patients into more useful prognostic subgroups based on survival or treatment response. Gene expression data has been used successfully to define molecular subtypes for various cancers, and cancer subtypes have been found that correlate with different clinical outcomes such as drug sensitivity and overall survival. We asked whether we could identify new, informative subtypes for GBM and Serous Cystadenocarcinoma using PARADIGM IPAs rather than the raw expression data. The advantage of using IPAs is they provide a summarization of copy number, expression, and known interactions among the genes and may therefore provide more robust signatures for elucidating meaningful patient subgroups.

We first determined all IPAs that were at least moderately recurrently activated across the GBM samples and found that 1,755 entities had IPAs of 0.25 in at least 75 of the 229 samples. We collected all of the IPAs for these entities in an activity matrix. The GBM samples and entities were then clustered using hierarchical clustering with uncentered Pearson correlation and centroid linkage (Figure 3.10). Visual inspection revealed four obvious subtypes based on the IPAs with the fourth subtype clearly distinct from the first three. The Serous Cystadenocarcinoma samples contained less obvious clusters and thus were determined by running HOPACH with uncentered

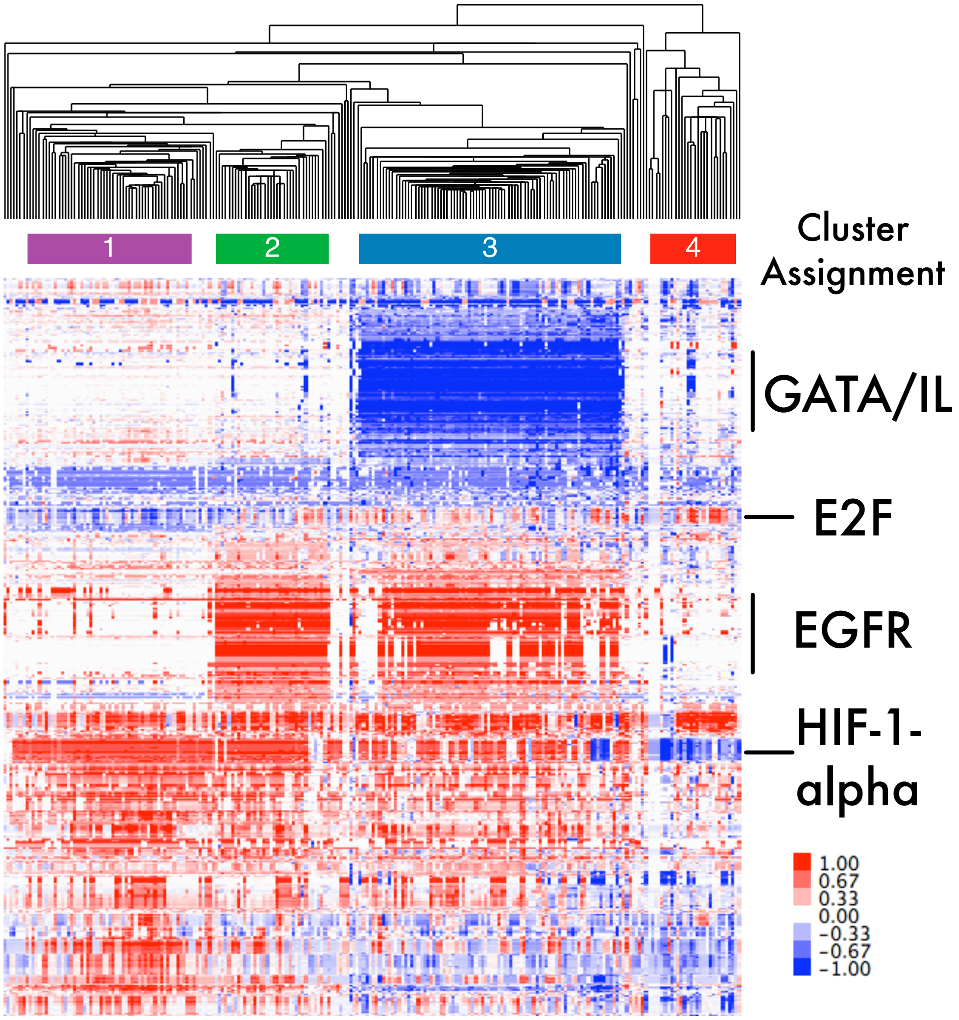


Figure 3.10: Clustering of IPAs for TCGA GBM. Each column corresponds to a single sample, and each row to a biomolecular entity. Color bars beneath the hierarchical clustering tree denote clusters used for Figure 3.12.

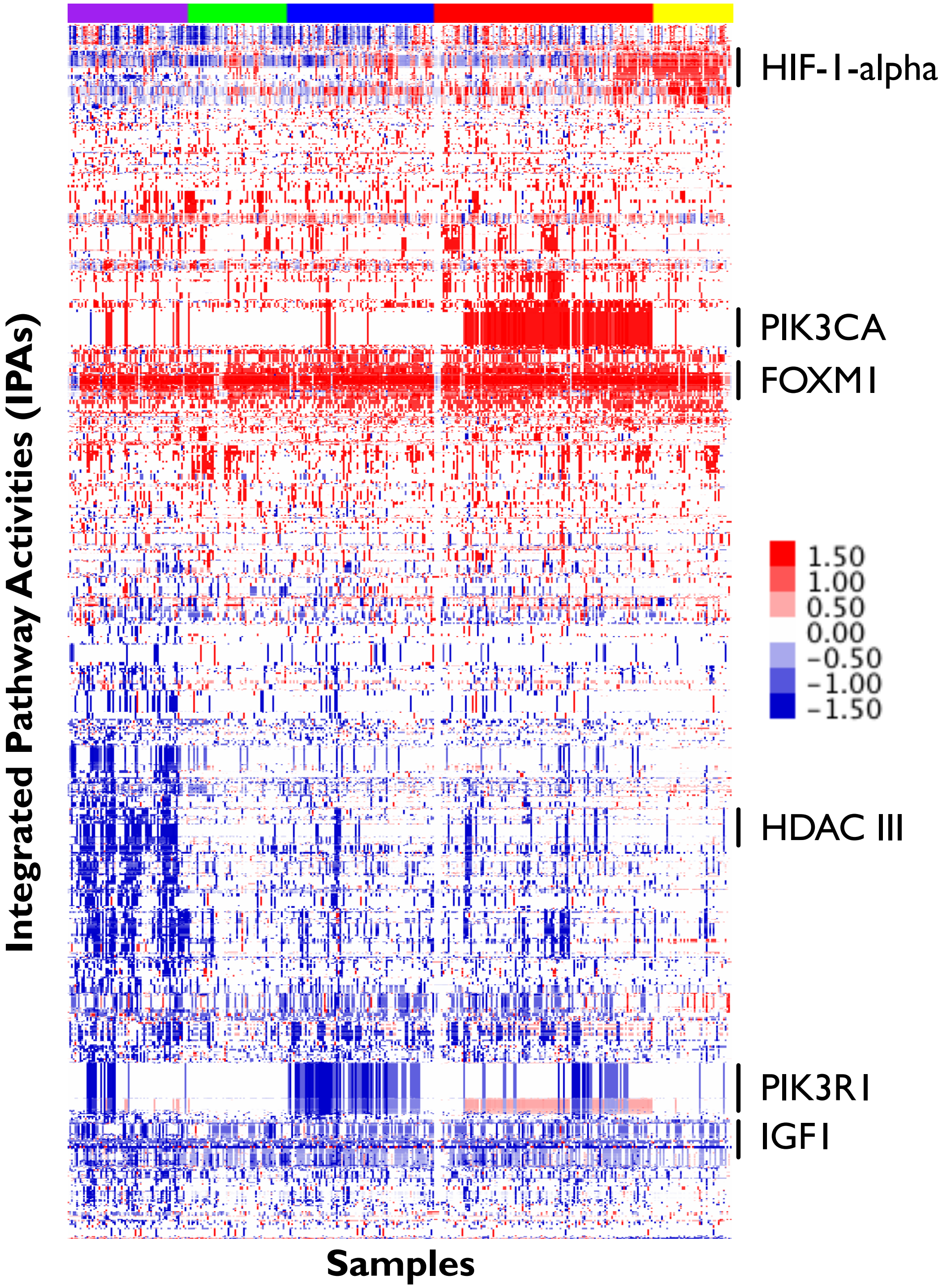


Figure 3.11: Clustering of IPAs for TCGA OV. Each column corresponds to a single sample, and each row to a biomolecular entity. Color bars above the heatmap denote clusters used for Figure 3.13.

Pearson correlation, finding five consistent clusters across the samples (Figure 3.11).

The fourth cluster in GBM exhibits clear downregulation of HIF-1-alpha transcription factor network as well as overexpression of the E2F transcription factor network (also shared by cluster 3). HIF-1-alpha is a master transcription factor involved in regulation of the response to hypoxic conditions. In contrast, two of the first three clusters have elevated EGFR signatures and an inactive MAP kinase cascade involving the GATA interleukin transcriptional cascade. Interestingly, mutations and amplifications in EGFR have been associated with high grade gliomas as well as glioblastomas [44]. Amplifications and certain mutations can create a constitutively active EGFR either through self stimulation of the dimer or through ligand-independent activation. The constitutive activation of EGFR may promote oncogenesis and progression of solid tumors. Gefitinib, a molecule known to target EGFR, is currently being investigated for its efficacy in other EGFR-driven cancers.

In OV, it is clear the red cluster is defined by high PIK3CA levels, an upstream kinase of AKT2 for which therapies are currently being explored in ovarian cancer [68]. The yellow cluster shows high HIF-1-alpha related activity, for which there is increasing evidence that high levels of HIF-1-alpha is prognostic for lower overall survival [17]. Overall, the purple cluster was mostly defined by lower activities in histone deacetylase class III (NAD-dependent) related proteins, in particular EP300, a protein which has previously been shown to be mutated in a subset of epithelial cancers [30]. In light of these well defined clusters, qualitatively they appeared to be honing in on biologically meaningful themes that can stratify patients.

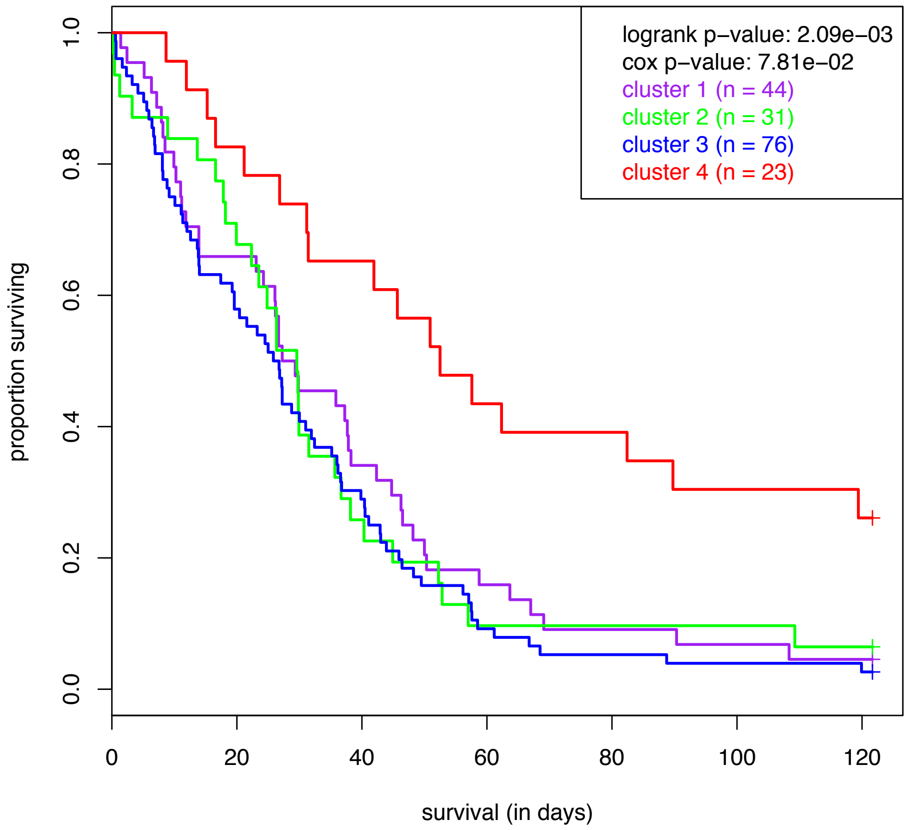


Figure 3.12: Kaplan-Meier survival plots for the GBM clusters from Figure 3.10.

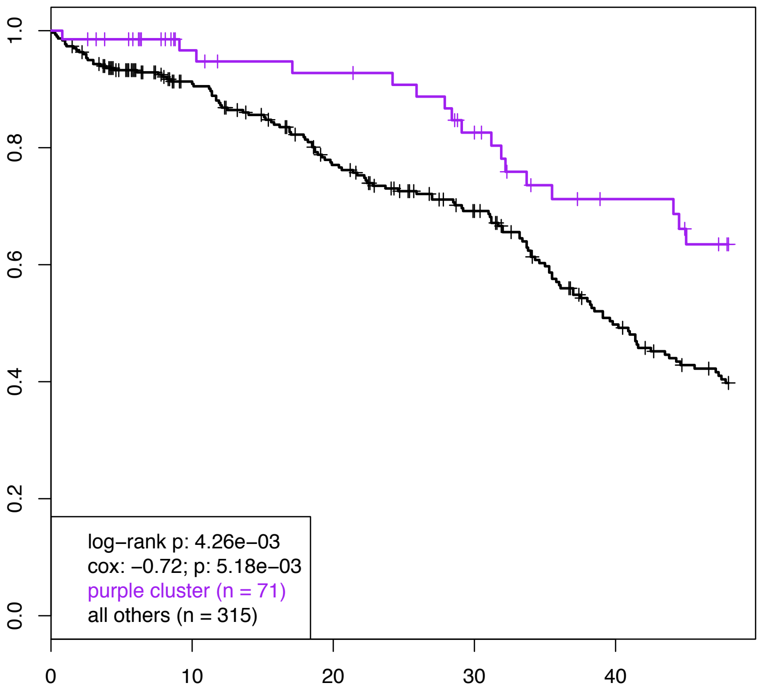


Figure 3.13: Kaplan-Meier survival plots for the OV purple cluster versus the rest from Figure 3.11.

To quantify these observations, we asked whether the different GBM and OV subtypes identified by PARADIGM coincided with different survival profiles. We calculated Kaplan-Meier curves for each of the sets of cluster clusters by plotting the proportion of patients surviving versus the number of months after initial diagnosis. We plotted Kaplan-Meier survival curves for each of the sets of clusters to see if any cluster associated with a distinct IPA signature was predictive of survival outcome (Figure 3.12). The fourth cluster in GBM is significantly different from the other clusters ($P < 2.11 \times 10^{-5}$; Cox proportional hazards test). Half of the patients in the first three clusters survive past 18 months; the survival is significantly increased for cluster 4 patients where half survive past 30 months. In addition, over the range of 20 to 40 months, patients in cluster 4 are twice as likely to survive as patients in the other clusters. In OV, when the purple cluster is compared to the remaining samples, we see a significantly better overall survival curve even when corrected for all possible single-cluster comparisons ($p < 0.0213$ with bonferroni correction). On average, approximately 20% more patients are alive in the purple cluster at any one time-point during the four years following diagnosis.

The survival analysis in GBM revealed that the patients in cluster 4 have a significantly better survival profile. Cluster 4 was associated with an inactivity of the HIF-1-alpha transcription factor. The inactivity in the fourth cluster may be a marker that the tumors are more oxygenated, suggesting that they may be smaller, newer, or simply better vascularized. These vascularized tumors with increased E2F may have allowed chemotherapy to more effectively reduce the aerobic state, suppressing the tumor's high E2F and proliferation, effectively slowing its growth and allowing for longer patient survival. In Ovarian, as previously mentioned, the purple cluster is defined by low HDAC-III signaling-related activities, in particular EP300. EP300 has been shown to have truncating mutations in many epithelial cancers, including ovarian, and is a known tumor suppressor. Because these ovarian patients were given platinum-based treatments, having reduced tumor suppressor levels (and higher proliferation levels) may have resulted in increased effectiveness of the treatment. It is clear that PARADIGM IPAs are pointing to novel and clinically meaninfgul pathways that can now be experimentally evaluated for their relation to treatment response and/or patient outcome.

To confirm the novelty of pathways identified by PARADIGM, we also attempted to cluster the patients in GBM using only expression data or CNA data to derive patient subtypes. No obvious groups were found from clustering using either of these data sources, consistent with the findings in the original TCGA analysis of this dataset [74] (Figure 3.14). This suggests that the interactions among genes and resulting combinatorial outputs of individual gene expression may provide a better predictor of such a complex phenotype as patient outcome.

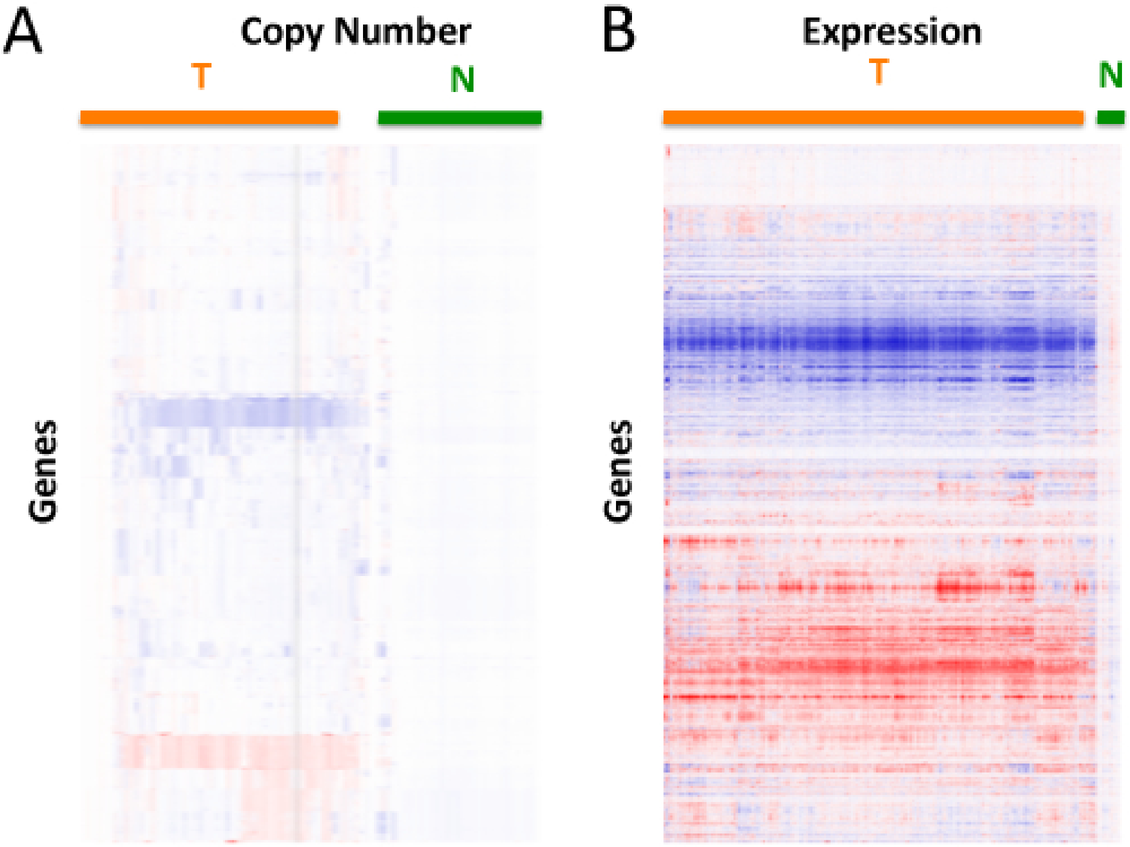


Figure 3.14: Clustering glioblastoma multiform (GBM) datasets with HOPACH did not reveal any obvious patient sub-types when using either gene-level copy number or expression data. Patient samples (columns) were clustered according to gene-level copy number or expression (rows). Clustering may be dominated by the nearly patient-wide amplifications and deletions (and correlated high and low expression) for a subset of the genes in the genome (large blue and red swaths in the heatmaps). A. Copy number estimates from competitive genome hybridization for 17,508 probes across 267 tumor (left) and 170 normal (right). B. Microarray gene expression data for 11,240 probes across 243 tumor (left) and 10 normal (right) samples.

Chapter 4

SuperPathway - A Global Pathway Model for Cancer

The original implementation of PARADIGM is an important tool for interpreting multi-dimensional datasets at a single-sample resolution. However, that algorithm heavily relied on a single pathway source (NCI PID) and two input datatypes (genomic copy number variations and transcriptome data). In order to more accurately reflect the underlying biology in the system, additional pathway representations are critical. Currently NCI PID contains pathways central to the mechanisms of cancer, including functional processes such as DNA repair, cell cycle regulation, and angiogenesis. There are large classes of pathways that are missing from this database that we must also model, such as metabolic and cellular differentiation processes. Although NCI PID is one of the highest quality pathway databases available, many other pathway databases that represent additional cellular mechanisms could be utilized by PARADIGM to successfully infer activities and stratify patient populations. In particular, Biocarta and Reactome [39] offer large databases of interactions from both curation and experimental assays.

This raises an important question about the fundamental information being modeled in PARADIGM. In it's original form, each pathway was modeled as an independent collection representing a set of cellular entities related to each other in a particular process. Because of this independence, a particular protein or complex with multiple functions may be represented separately in each pathway. This results in the model inferring multiple values for certain proteins that appear in more than one pathway in the database. While modeling these pathways independently allows us to deconvolute differentially activated processes within cancer, multifunctional proteins are intrinsically tied to all their roles within the cell, and each pathway is not acting in its own compartment. This naturally raised the question of whether of not it is possible for PARADIGM to model a more "global" pathway model in the cell, representing thousands of genes and interactions in a singular context that would more accurately represent a snapshot of a cancer cell.

Other groups have successfully built databases designed to combine pathways sources, such as Pathway Commons. Pathways Commons is a collaborative database powered by cPath [14] and run by MSKCC and University of Toronto that consolidates nine pathway databases onto a single cellular network. As of May 2010, they represent 1,400 pathways containing 420,000 interaction and 88,500 cellular entities across 440 organisms. Unfortunately, the approach taken by Pathway Commons is focused

Table 4.1: Features and Interactions in SuperPathway (July 2011 version)

Concepts		Interactions	
Protein	6906	Protein Activation	7269
Complex	7345	Protein Inhibition	1005
Family	1449	Transcriptional Activation	1963
Abstract	582	Transcriptional Inhibition	386
miRNA	15	Complex Formation	23132
RNA	55	Family Membership	6559
Total	**16352**	**Total**	**40314**

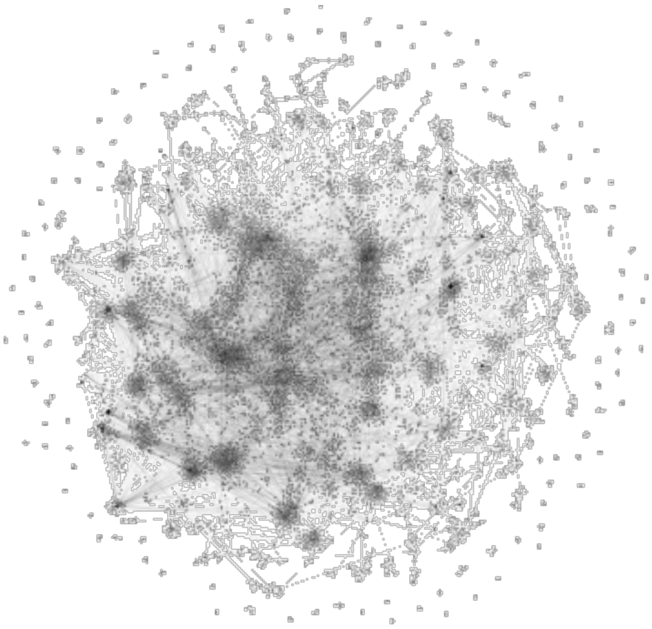


Figure 4.1: Topographical visualization of the SuperPathway.

primary on integration, and thus their representation of the cellular network contains only undirected interactions. Because of this limitation, their database is unsuitable for a method that relies on direction and sign of interactions to propagate information. Furthermore, there are huge computational implications in representing and modeling thousands of interactions and entities simultaneously.

Because of the incomplete nature of Pathway Commons, I built a superimposed version of the collection of pathways we have termed the SuperPathway. 1321 Pathways were obtained on July 25th, 2011 in BioPax Level 2 format from NCI-PID, Reactome and BioCarta. Genes, complexes, families and abstract processes (e.g. "cell cycle" and "apoptosis") were unified by Uniprot ID (genes) or name across the three sources, created a definitive list of what we call "pathway concepts." Across the three sources, this list contained 16,352 total concepts, including 6,906 genes, 7,345 complexes, 1,449 families and 582 processes and 40,314 interactions as outlined in Table 4.1. Links were combined if they consisted of the same parent, child and interaction type, and the overall topology of the network can be see in Figure 4.1.

Across the initial set of pathways, an additional concept type of small molecule was present and represented over 800 unique concepts. The small molecule type is reserved for biochemical products that are byproducts of enzymatic or kinase activities, including molecules such as H_2O and ATP/ADP. While maintaining these elements in the pathway might increase our ability to understand the metabolic flux that is occurring in the tumor, these elements have a disproportionately higher degree than other concepts in the pathway. As a result, these concepts tend to connect aspects of the network that would normally be several links away, causing inappropriate propagation of belief to nodes. Although future versions of PARADIGM may attempt to model these reactions using techniques such as flux-balance analysis [23], that work is outside the scope of the original intention of PARADIGM and so these nodes were removed from the final SuperPathway. This results in a loss of just over 1500 concepts across all the categories due to them only being connected in the pathway through the small molecules that were removed. The net effect resulted in a much more computationally tractable pathway that forms the basis for the studies in this chapter.

In order to efficiently compute activities across this network, a number of optimization techniques were required beyond the removal of small molecules. Of particular interest, exact inference was no longer possible due to a number of cycles and contractions within the network. Instead, loopy belief propagation can be used to compute probabilities with a tolerance of 1e-9 and a maximum of 10,000 iterations. Because this method can be run on cycles and with contradictory information, all interactions in the network were considered and no attempt to resolve them was made. The final SuperPathway was built by performing a breadth-first traversal starting from the concept with the highest number of interactions, which resulted in capturing a majority of the concepts, effectively capturing the largest possible connected network.

4.1 SuperPathway in Ovarian Cancer

Our previous studies had shown the importance of FOXM1 in the TCGA Ovarian Cancer cohort, so as a first test of the SuperPathway I asked if we could recapture this information given the new overall pathway structure. Using both the SNP 6.0 copy number data and Agilent 244k mRNA expression data from the 316 high grade serous cystadenocarcinoma samples found in the TCGA publication, I computed pathway IPAs across the SuperPathway. Because of the reduced I/O and number of

Table 4.2: Top TCGA Ovarian Concepts from SuperPathway

Concept	OV Avg	OV StdDev	NA Avg	NA StdDev	-Log(TTest)
FOXM1	3.7861	1.0077	0.3327	0.9395	209.5321
SKP2	3.2782	0.9818	0.5409	0.8146	169.2913
DSP	3.2318	1.9157	1.0678	2.0729	52.5983
MAPK	3.2180	1.9049	1.0429	2.0814	53.3704
E2F3	3.1120	0.7703	0.5761	0.8801	211.2003
BCAT1	3.0954	0.8581	0.5739	0.8689	184.4879
CKS1B	3.0590	0.7361	0.4401	0.6965	217.3218
CENPF	2.9807	0.7310	0.4242	0.6916	214.0751
NEK2	2.9700	0.7703	0.4466	0.6898	199.4847
MINA	2.7870	1.2480	0.5661	0.8791	102.9586
		...			
PIK3C2A	-1.3275	0.4235	-0.3690	0.6195	137.2294
chromatin remodeling	-1.3311	0.5115	-0.4604	0.6970	93.0525
Alpha v beta 3 : Vitronectin complex	-1.3901	0.4417	-0.3998	0.6296	136.1474
Alpha v beta 5 : Vitronectin complex	-1.3901	0.4417	-0.3998	0.6296	136.1474
Alpha v beta 8 : Vitronectin complex	-1.3901	0.4417	-0.3998	0.6296	136.1474
WNT7B / FZD10 / LRP5	-1.3950	0.7056	-0.4736	0.7383	66.4470
FOXA1	-1.4200	0.8062	-0.2365	0.8047	79.5173
WNT5A / FZD4 / LRP5	-1.4550	0.7294	-0.4981	1.1383	57.9418
WNT5A / ROR2	-1.4550	0.7294	-0.4981	1.1383	57.9418
VTN	-1.5199	0.5141	-0.3677	0.7677	134.8298

processes required, the overall run time was more efficient at approximately 5 minutes per sample versus the approximately 1 minute per pathway per sample across 1300 pathways. Due to there only being a single pathway, the traditional analysis methods of comparing individual pathways no longer apply in an analogous fashion. Instead, features can be compared with a more traditional Student's T-Test for the real samples versus the cohort of permuted samples (generated as outlined earlier).

Table 4.2 lists the top features found in the SuperPathway versus the null model, where activity is ranked according to the average IPA value across the real samples. Again, FOXM1 appears as the top entry, confirming that despite large scale structural changes of the underlying pathway data, PARADIGM can identify consistently overactive features across a cohort. SKP2, the second highest activity across all the samples, has recently been indicated as a prognostic factor in ovarian adenocarcinoma[64]. The appearance of WNT signaling towards the bottom of the activity list is interesting, given there are increasing numbers of papers implicating WNT signaling and the differences between normal ovarian and cancerous tissue, despite the lack of mutations[29]. PIK3C2A, a critical subunit of the PI3K Complex has been clearly implicated in ovarian cancer, however as a oncogene[29] - a feature that appears to contradict the appearance towards the bottom of the activity list. However, primary alteration of PI3K is through mutational alterations, a feature that was not provided to PARADIGM in this context. The discrepancy may be explained as a result of parent nodes attempted to downregulate PIK3C2A activity as the cell detects overactive PI3K levels.

Although looking at top concepts across the patient cohort can offer insight into

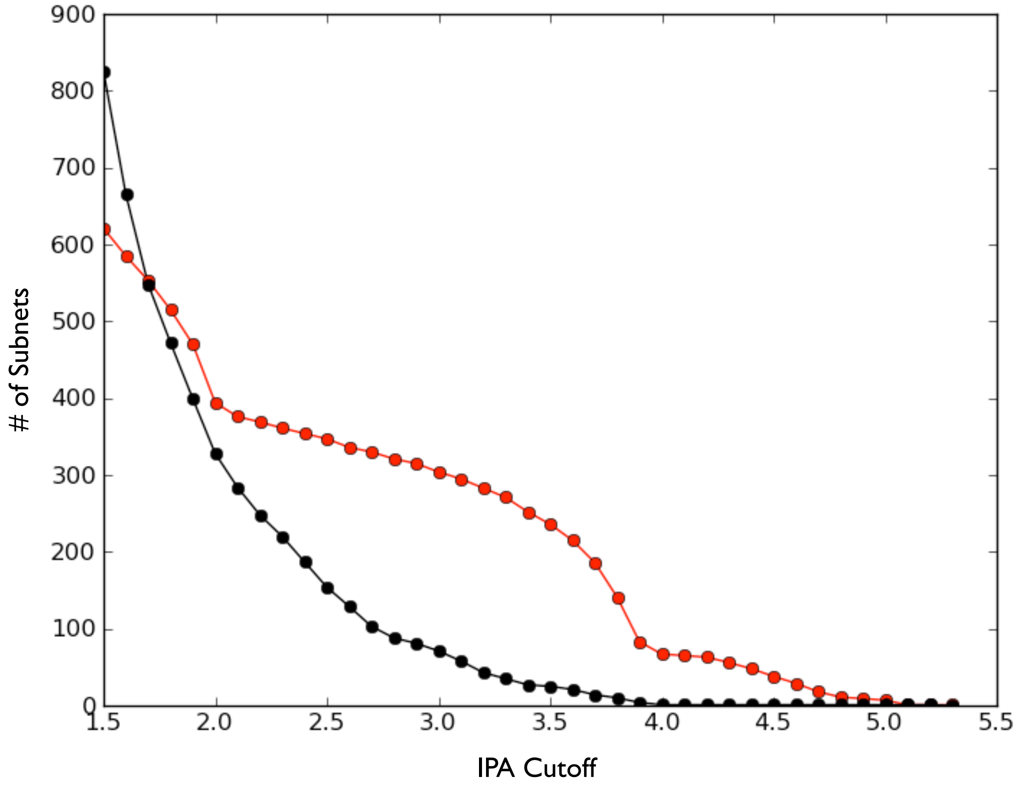


Figure 4.2: Number of subnets found for a variety of IPA cutoffs. The red line represents actual samples, with the black line representing the permuted samples. The max difference of 156 subnets between these two populations was found at IPA cutoff of 3.3.

Table 4.3: Top pathway concepts consistently found in subnets across the OV samples using IPA cutoff 3.3. Only HSP90AA1 has more null samples than real, indicating a likely false positive.

Concept	NA Sum	OV Sum
FOXM1	9	157
SKP2	8	154
MYC/Max	13	150
CKS1B	6	148
NEK2	7	138
CENPF	5	137
BCAT1	2	116
E2F3	6	116
MINA	7	115
DDX18	7	107
PFKM	8	105
PEG10	5	93
BMI1	4	92
CCNB2	4	85
MYC/Max/RPL11	6	82
MTDH	5	81
TAF4B	7	72
HSP90AA1	*93*	*57*
PRDX3	7	52
BRCA2	3	49

highly over- or under-active pathway processes, this approach doesn't take advantage of the connected nature of the resulting pathway concepts. One alternative approach is to reconstruct small subnetworks within the larger pathway context, retaining the information regarding interactions and taking advantage of the existence of concepts that span cellular contexts to obtain a better understanding of the interconnected nature of the cell. As an initial approach to a cohort-level subnet analysis, I looked for regions of the network above a certain IPA score continuing to expand as long as connected components met the cutoff. Subnets were only retained if they contained at least 5 concepts. Figure 4.2 shows the number of subnets found for each IPA cutoff possible, both for real samples and permuted samples.

At the optimal cutoff of abs(IPA) $>=$ 3.3, there were 156 more subnets in the real samples versus the permuted samples. Every concept was then ranked across all the samples to compute the most prevalent features within the subnets, the results of which appear in Table 4.3. Again, FOXM1 appears as the most consistent concept across approximately half of the ovarian samples, showing that the high IPA of this concept is supported by data from interactions in at least half of the cases. Reassuringly, many of the same concepts from this subnet analysis are also seen in the previous, single concept analysis. However, some of the concepts that appeared to have high overall pathway activity do not appear, such as DSP and MAPK. These concepts also have lower t-test p-values than the other members in the list, indicating they are less likely to be pathway specific drivers and instead may be due to technical artifacts from the arrays that measured them. From this study it is clear the SuperPathway provides an exciting new view of oncogenic pathway activities while still recapitulating previously established biology, and presents a framework that allows us to build new analysis methods to understand differential activities in a global scope.

4.2 SuperPathway in Breast Cancer

4.2.1 Chin-Naderi Cohort

The success of the SuperPathway to identify key pathway activities and subnets across a cohort of samples provides a useful framework for finding key differential activities within a subset of a cohort. To explore this idea, we asked if we could use the activities generated in the SuperPathway to identify key subnetworks that correspond with a particular clinical feature. We re-ran the previously analyzed Chin-Naderi breast cancer cohort on the SuperPathway, and used the clinical annotation of ER+ve and ER-ve, as provided in the publications. Using the non-parametric Wilcoxon rank sum test, we assigned each node in the network a differential score of the negative log10 p-value, signed by the group with the higher mean. We then removed any node with an absolute negative log10 value below 1.3 as a method to trim activities to only those found significant. This resulted in a series of subnets, the largest consisting of 141 nodes and 188 edges, which were visualized using Cytoscape[66].

Figure 4.3 captures the output of the SuperPathway subnet analysis on the Chin-Naderi breast cancer, showing the regulatory networks of genes that are significantly differentially expressed in the ER+ve compared to the ER-ve subgroups of breast

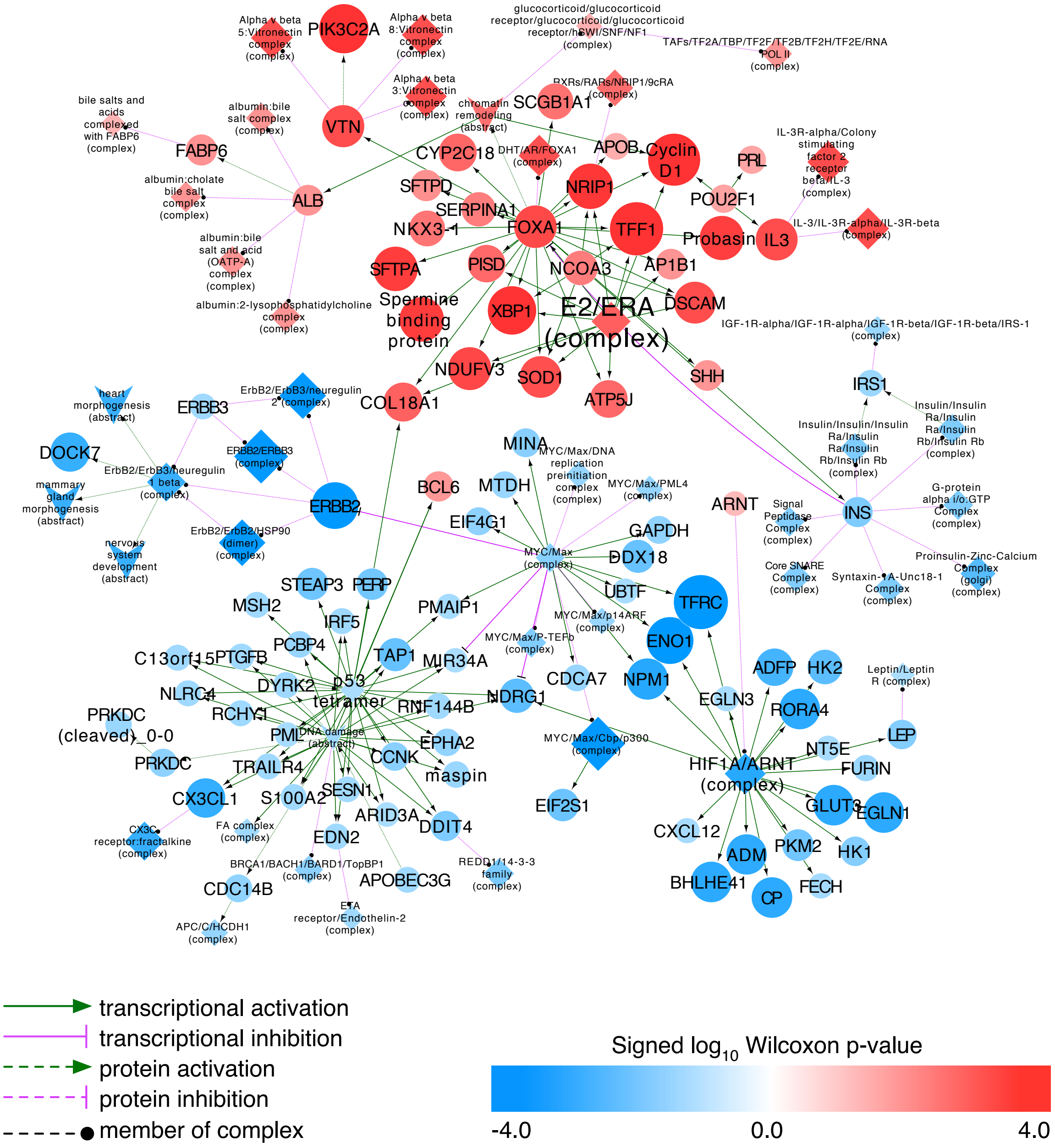

Alpha v beta 5:Vitronectin complex (complex)
PIK3C2A
Alpha v beta 8:Vitronectin complex (complex)
glucocorticoid/glucocorticoid receptor/glucocorticoid receptor/hSWI/SNF/NF1 (complex)
TAFs/TF2A/TBP/TF2F/TF2B/TF2H/TF2E/RNA POL II (complex)
Alpha v beta 3:Vitronectin complex (complex)
RXRs/RARs/NRIP1/9cRA (complex)
bile salts and acids complexed with FABP6 (complex)
albumin:bile salt complex (complex)
VTN
chromatin remodeling (abstract)
SCGB1A1
FABP6
CYP2C18
DHT/AR/FOXA1 (complex)
APOB
Cyclin D1
IL-3R-alpha/Colony stimulating factor 2 receptor beta/IL-3 (complex)
PRL
albumin:cholate bile salt complex (complex)
ALB
SFTPD
NRIP1
POU2F1
SERPINA1
NKX3-1
FOXA1
TFF1
Probasin
IL3
IL-3/IL-3R-alpha/IL-3R-beta (complex)
albumin:bile salt and acid (OATP-A) complex (complex)
SFTPA
PISD
NCOA3
AP1B1
albumin:2-lysophosphatidylcholine complex (complex)
Spermine binding protein
XBP1
E2/ERA (complex)
DSCAM
IGF-1R-alpha/IGF-1R-alpha/IGF-1R-beta/IGF-1R-beta/IRS-1 (complex)
NDUFV3
SOD1
SHH
ErbB2/ErbB3/neuregulin 2 (complex)
ATP5J
IRS1
heart morphogenesis (abstract)
ERBB3
COL18A1
Insulin/Insulin Ra/Insulin Ra/Insulin Rb/Insulin Rb (complex)
Insulin/Insulin/Insulin Ra/Insulin Ra/Insulin Rb/Insulin Rb (complex)
DOCK7
ERBB2/ERBB3 (complex)
MINA
MYC/Max/DNA replication preinitiation complex (complex)
ErbB2/ErbB3/neuregulin 1 beta (complex)
mammary gland morphogenesis (abstract)
BCL6
MTDH
MYC/Max/PML4 (complex)
G-protein alpha i/o:GTP Complex (complex)
ERBB2
ARNT
Signal Peptidase Complex (complex)
INS
EIF4G1
ErbB2/ErbB2/HSP90 (dimer) (complex)
nervous system development (abstract)
GAPDH
MYC/Max (complex)
DDX18
Core SNARE Complex (complex)
Syntaxin-1A-Unc18-1 Complex (complex)
Proinsulin-Zinc-Calcium Complex (golgi) (complex)
STEAP3
PERP
UBTF
TFRC
MSH2
PMAIP1
IRF5
MYC/Max/p14ARF (complex)
ENO1
PCBP4
TAP1
MYC/Max/P-TEFb (complex)
C13orf15
PTGFB
MIR34A
ADFP
HK2
Leptin/Leptin R (complex)
DYRK2
p53 tetramer
NLRC4
CDCA7
NPM1
NDRG1
EGLN3
RCHY1
RNF144B
RORA4
PRKDC (cleaved)_0-0
PML
DNA damage (abstract)
LEP
MYC/Max/Cbp/p300 (complex)
EPHA2
NT5E
PRKDC
TRAILR4
CCNK
HIF1A/ARNT (complex)
FURIN
maspin
CX3CL1
S100A2
SESN1
GLUT3
EGLN1
EIF2S1
ARID3A
CX3C receptor:fractalkine (complex)
FA complex (complex)
DDIT4
CXCL12
EDN2
PKM2
HK1
ADM
BRCA1/BACH1/BARD1/TopBP1 (complex)
REDD1/14-3-3 family (complex)
APOBEC3G
BHLHE41
FECH
CDC14B
CP
ETA receptor/Endothelin-2 (complex)
APC/C/HCDH1 (complex)
transcriptional activation
transcriptional inhibition
protein activation
protein inhibition
member of complex
Signed $\log_{10}$ Wilcoxon p-value
-4.0
0.0
4.0

Figure 4.3: Paradigm reveals the regulatory network underlying estrogen receptor positivity in the Chin-Naderi Cohort.

cancer. The FoxA1 network is clearly seen as highly differentially expressed, followed by the Hif-1-alpha transcription factor network, which was also identified by SPIA and GSEA as being downregulated in ER+ve tumors. This is concordant with recently published evidence that that Hif-1-alpha represses the transcription of the estrogen receptor gene, ESR1, in breast cancer cell lines and thus could play a role in ER-ve breast cancers[59]. This is particularly important given that Hif-1-alpha is an important regulator of the cellular response to hypoxia and has been shown to be an independent prognostic factor in breast cancer[11].

4.2.2 TCGA Breast Cancer

Using the same approach as in the Chin-Naderi cohort, I wanted to validate these findings on an independent cohort using a more prognostic clinical feature, in hopes to understand the molecular basis of prognosis in breast cancers. Using 463 samples in the TCGA breast cancer dataset, PARADIGM was run on the SuperPathway utilizing the SNP 6.0 copy number and Agilent 244k gene expression platforms. Resulting features were ranked using the same approach as in the Chin-Naderi cohort, this time using the intrinsic expression subtypes (as previously calculated by the Perou group) of Luminal A and Basal as the clinical feature of interest. Because of the large number of features found differential by this clinical attribute, the subnet creation was modified to a more stringent approach to ease with visualization. Links were only kept if the parent of the link had at least 10 children concepts in the SuperPathway and the child of the link had at least 5 children concepts in the SuperPathway. Effectively termed "hubs",

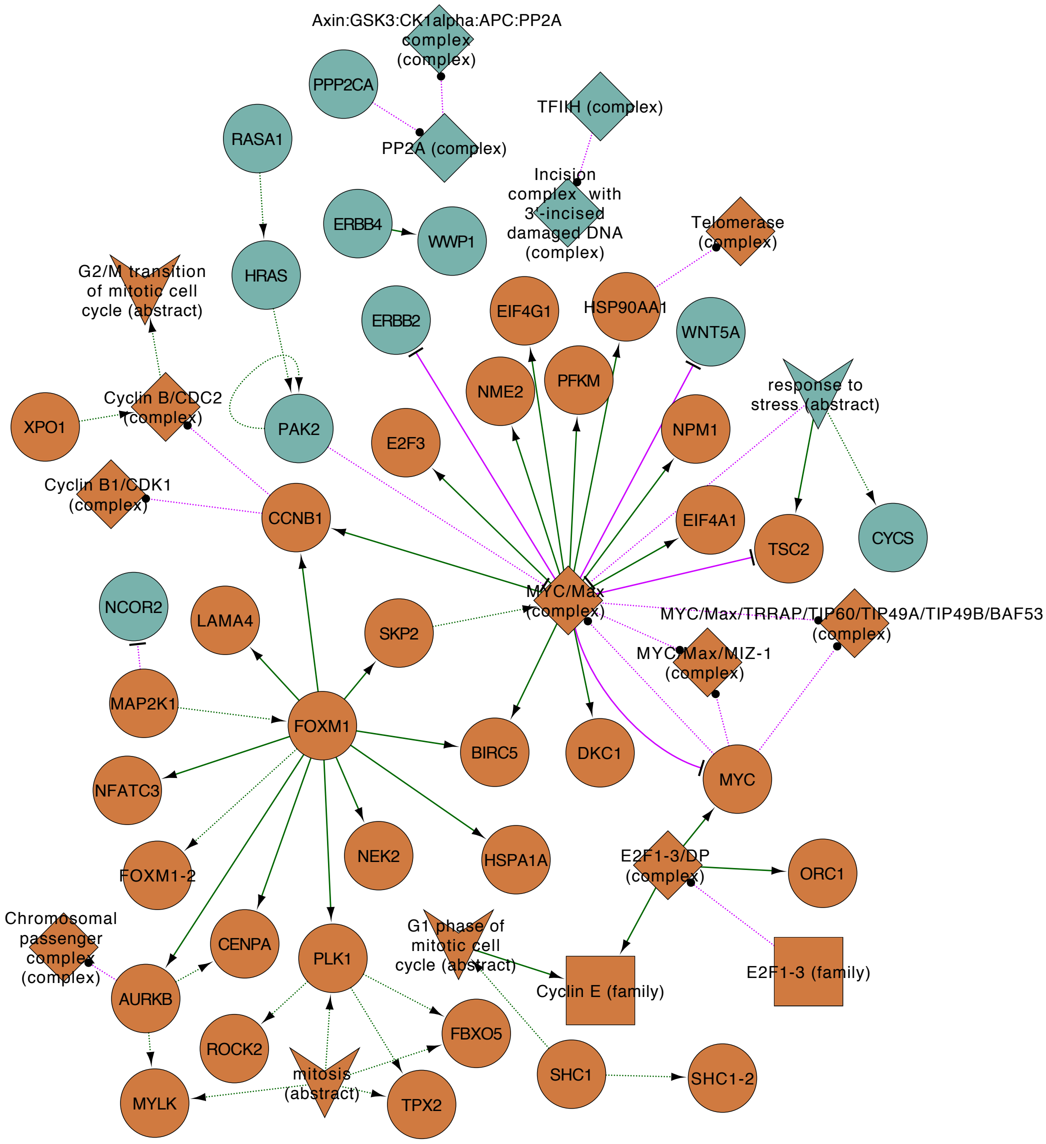


Figure 4.4: Paradigm concept hubs separating Luminal and Basal in the TCGA Breast Cancer Cohort. In order for a link to appear in the subnetwork, the parent node needed at minimum 10 children nodes in the overall network. In addition, the child node needed at minimum 5 children nodes in the overall network. Concepts are colored orange if their activity is higher in Basal samples, and cyan if their activity is higher in Luminal samples.

these concepts are key features in the SuperPathway that are responsible for regulating a disproportionally high number of other nodes, given the average number of neighbors of 4.864 and network density of close to 0 across the entire pathway. Figure 4.4 shows these hubs colored by the class with the higher mean activity across the samples (orange for Basal, cyan for Luminal). Features of note include basal enrichment in the proliferative markers FOXM1, MYC/Max, AURKB and the mitosis abstract process. Overall a much more limited set of concept hubs appear to have increased activity in the luminal samples, consistent with the known aggressiveness of basal tumors compared to luminals[12].

It is then interesting to ask if it is possible to use the same analysis to distinguish between mutations within a single protein that occur in different domains. There is increasing evidence of the importance of PI3K mutations in breast cancer, and recent studies have suggested that the effect of the mutation is highly dependent on the protein domain that is altered[85]. Using the same subnetwork hub analyses, TCGA breast cancer samples were split on kinase versus non-kinase domain mutants. The resulting subnetwork with the mutational visualization can be seen in Figure 4.5. Although less significant than the split done with the intrinsic subtypes, the kinase domain mutants appear to be enriched for proliferation-related activities, while non-kinase domain mutants appear to be enriched in proteins responsible for cellular migration (ACTN1) and adhesion (E-cadherin). Mutation distribution was independent of subtype and TP53 status, indicating these results are unlikely to be confounded by alternative factors. Although the clinical implications of these results are unclear, this methodology provides

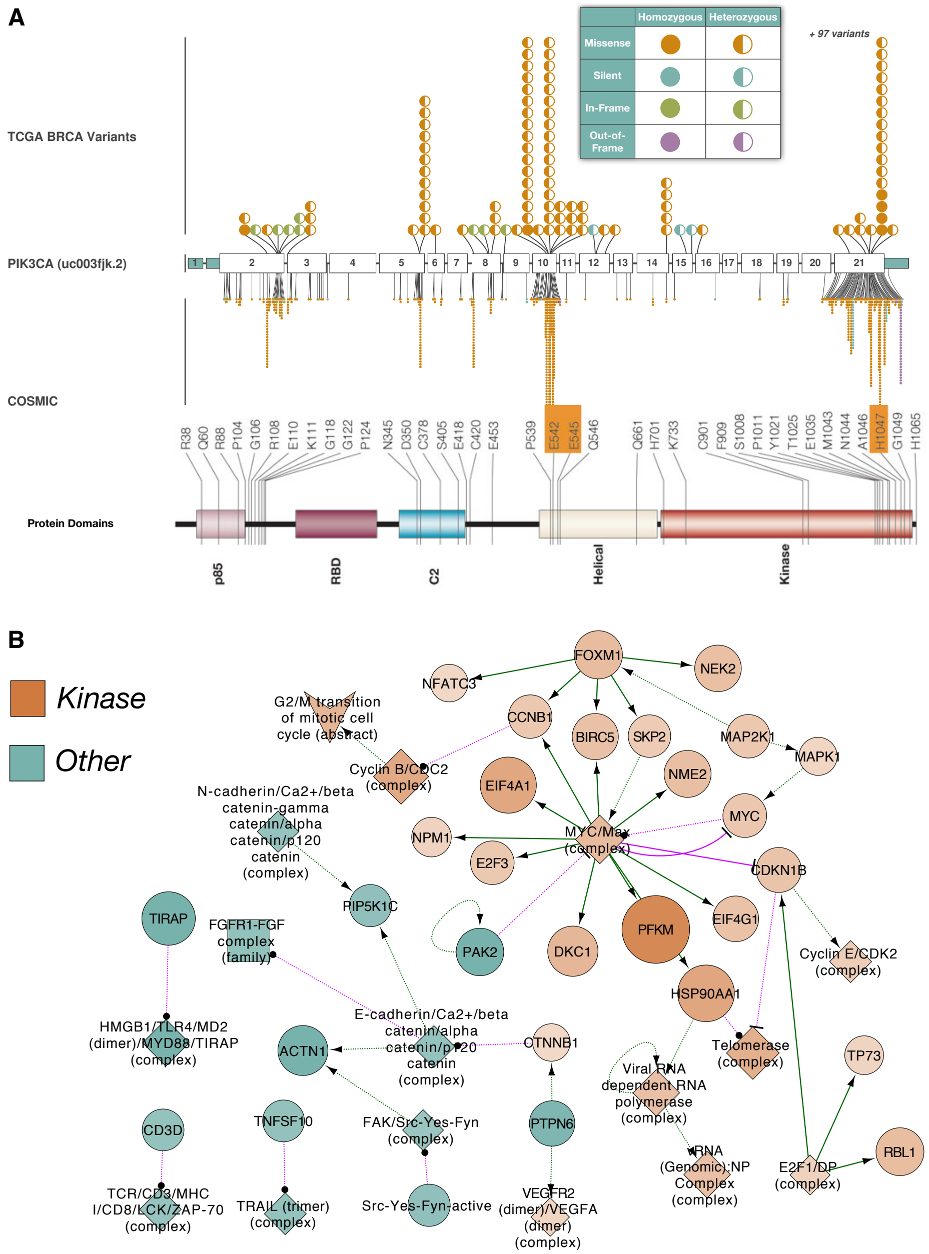


Figure 4.5: Mutations in different protein domains show differential pathway activity. A: Mutational comparison of TCGA breast cancer samples versus mutations seen in the COSMIC mutation database. B: Pathway hubs that show differential activity between kinase and non-kinase mutations, colored and sized by significance.

a useful mechanism for understanding even subtle sample differences that can effect regions of the pathway, allowing researchers to focus on subnetworks that are more easily interpretable.

4.3 Cross-Cancer SuperPathway

Recurrent pathways that are commonly activated in subsets of tissue-specific cohorts are particularly interesting, as these may indicate shared molecular mechanisms for oncogenesis. If common pathways can be detected and correlated with response to known therapies, it may be an indication that the molecular profile of those tissues is more important than the tissue of origin. Diseases such as breast cancer are good candidates for this type of analysis, as there are well documented molecular heterogeneity and survival-based subtypes describing this heterogeneity [67]. As the TCGA has promised to produce expression and copy-number data for over twenty thousand patients across more than twenty tissues of origin, that sample population represents a powerful dataset to explore the molecular origins of cancer. The consistency of datatypes and platforms that are used for measuring mRNA expression and copy-number alterations (Agilent 244K arrays and Affymetrix SNP 6.0 arrays, respectively) limits the effect of batch effects that can confound cross-cohort analyses. However, these samples have stringent purity and size requirements which may result in other unknown confounding effects.

In order to ask if molecular mechanisms resulting in the observed subtypes in breast cancer can be found in other tissues, I first ran PARADIGM across the entire

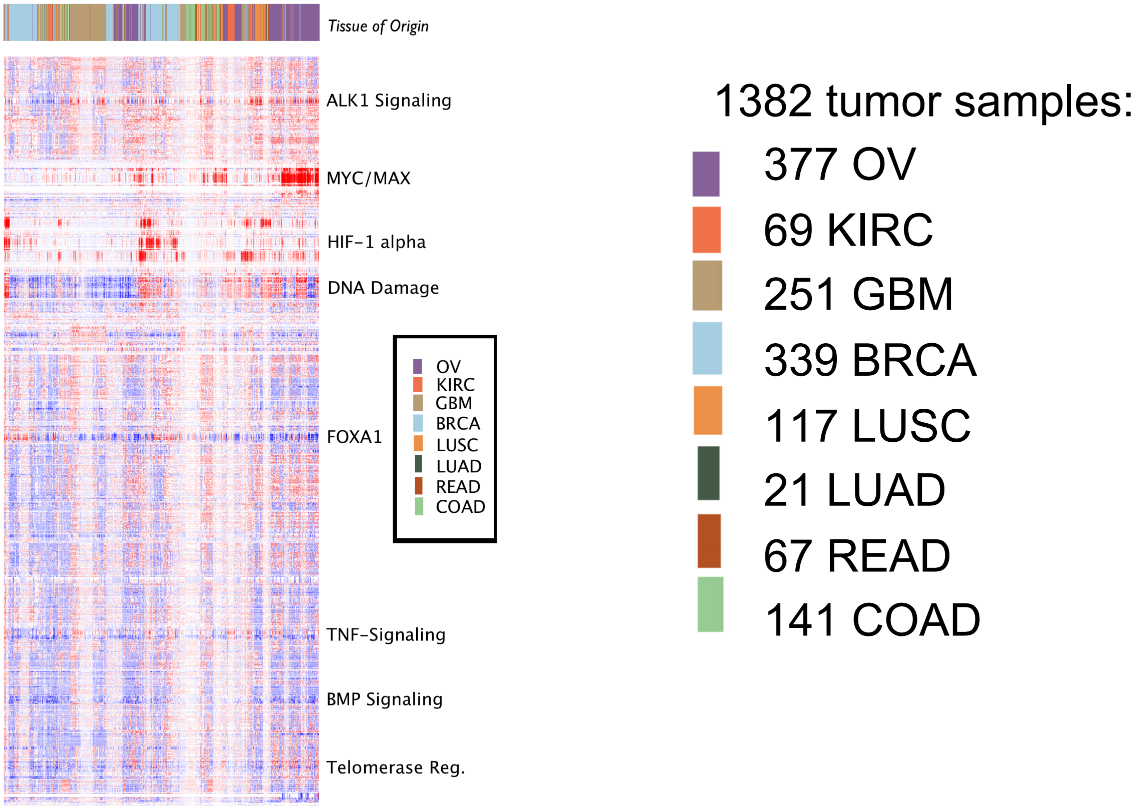


Figure 4.6: PARADIGM activities computed across 1382 TCGA samples from 8 tissues of origin with matched copy-number and median-centered gene expression data.

TCGA cohort. Expression data was median centered across all tissues for each gene to assure a common reference when comparing samples, and copy-number data remained normalized to blood normal for each sample. Figure 4.6 visualizes the resulting activity heatmap for close to 1400 samples across the 8 available tissues of origin at the time of analysis. Hierarchical clustering for the samples was performed, resulting in a surprising mixture of tissues across the dataset. Of particular interest, the breast appears to cluster into two distinct groups, each having similarities with a lung subset.

Because of the similarity in features between the breast and lung cancer samples within TCGA, I decided to focus my analysis on comparing breast intrinsic subtypes with the two classes of lung cancer (squamous and adenocarcinoma). Intrinsic subtypes for breast cancer were obtained from the Perou lab using their previously published methods, and were the same as used in the previous breast cancer analysis. Only samples labelled Luminal A were considered as the luminal class, and all basal labelled cancers were used, resulting in 250 samples. The two lung subtypes contained a total of 138 samples. Based on conversations with Dr. Eric Collison of UCSF and other collaborators, I decided to focus this analysis by asking if a machine learner can distinguish subtypes regardless of tissue of origin.

Using SVM-Light[38], I trained a linear SVM using default parameters on the entire set of breast cancer samples, where Basal samples were labelled "-1" and Luminal samples were labelled "+1". I then labelled lung squamous "-1" and lung adenocarcinoma "+1" and asked the SVM to classify these samples. Using 63 support vectors, the SVM classified 130 out of 138 lung samples correctly, resulting in an accuracy of 94.2%,

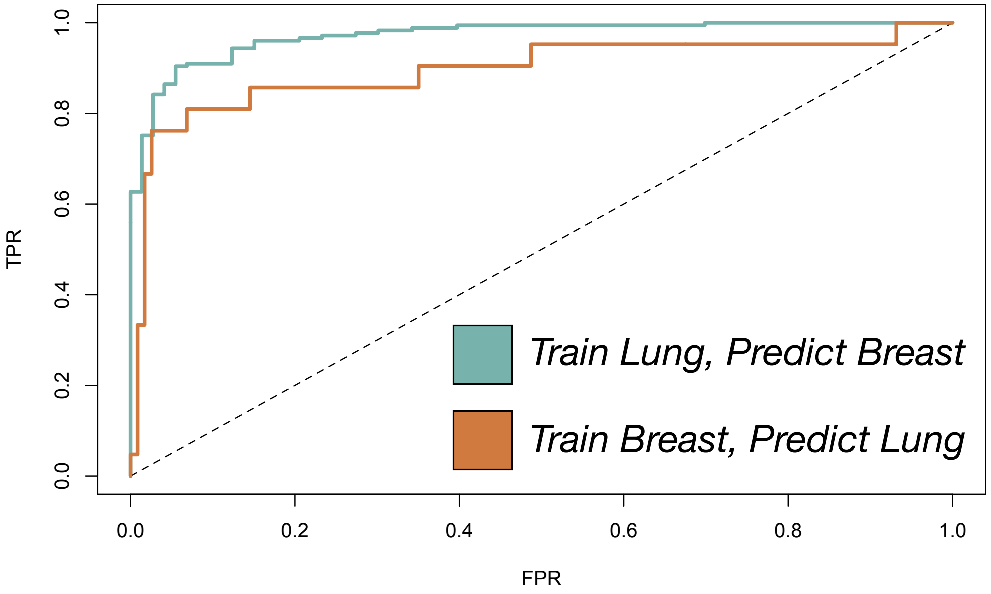


Figure 4.7: Receiver operating characteristics of SVM-Light classifications done across tissue types. The Cyan ROC indicates a linear SVM that was trained on lung squamous versus adenocarcinoma and given breast samples labelled basal or luminal for classification (AUC=0.971751). The orange ROC indicates the performance of a linear SVM that was trained on breast basal versus luminal and tested on lung squamous versus adenocarcinoma (AUC=0.894994).

precision of 84.21% and recall of 76.19%. A plot of the resulting ROC can be seen in Figure 4.7 and the resulting AUC was 0.894994.

I then reversed the experiment, asking if breast samples can be distinguished by training on lung. I passed all 138 lung samples to a linear SVM light trainer with default parameters, and asked it to classified the 250 breast samples. Using 53 support vectors, the SVM was only capable of classifying 173 of the 250 samples correctly, resulting in an accuracy of 69.2%, however with a precision of 100% and recall of 56.5%. A plot of the resulting ROC can also be seen in Figure 4.7 with a surprising AUC was 0.971751. The large AUC with corresponding low accuracy likely indicates a non-optimal cut point was chosen by the SVM, something that could likely be improved with additional training or parameters.

Given the classifier was capable of achieving high AUCs in both classification tasks, it is interesting to explore the features that were used by the SVM to help understand what might best distinguish the two subtypes. Figure 4.8 is a subnetwork analysis using SVM-weights assigned by training on the lung subtypes (the same weights that were used to classify on the breast samples). A concept was required to have a weight greater than 2 standard deviations from the mean of all weights in order to appear in the resulting network, in addition to the hub restrictions. A majority of concepts pulled out through this analysis have positive weights, as expected, with the most discriminative features being P53 (arguably the most famous tumor associated protein), HGF, VDR, SERPINE1 and HIF1alpha. The common basal / luminal markers of FOXA1, MYC/Max, and DNA Damage also appear. This analysis clearly shows that

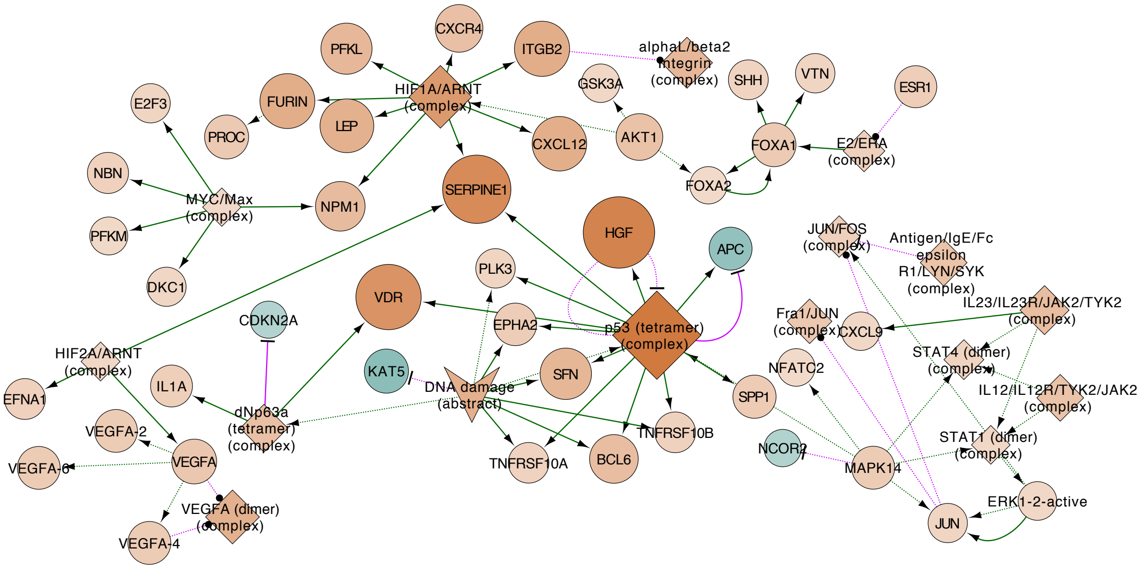


Figure 4.8: Paradigm concept hubs found to have high SVM weights when trained on Lung Adenocarcinoma versus Lung Squamous. In order for a link to appear in the subnetwork, the parent node needed at minimum 10 children nodes in the overall network. In addition, the child node needed at minimum 5 children nodes in the overall network. Concepts are colored orange if their SVM weight is positive, meaning they contribute to the resulting classification, and cyan if they have a negative SVM weight, meaning they offer no discriminative power.

there are corresponding molecular subtypes that can help explain origin of disease better than tissue of origin. As molecular classification becomes more prevalent in the clinic, it will be increasingly important to understand treatment from the aspect of molecular features instead of tissue of origin, and PARADIGM provides a powerful framework for that type of analysis.

Chapter 5

Pathway Analysis of Drug Effects

Detecting commonly perturbed pathways either across an entire cohort or that can successfully stratify patients is critical to understanding the underlying misregulation within a cancer type. However, we ultimately need to understand how these misregulated pathways can be fixed in order to provide clinically actionable information to healthcare providers. By examining the pathway activities in conjunction with drug responses in a model system such as cell lines, it is possible to identify key pathway activities that correspond with treatment response. These key pathways and activities can be used for *in silico* knockout experiments to find combinational therapies that will increase drug response within the subset of cell lines that did not respond. PARADIGM can provide a unique insight into the possible functional implications of pathway alterations, and may provide the information necessary to model these combination therapies.

5.1 SuperPathway on Breast Cell Lines

In order to asses the ability of pathway activities to provide a novel mechanism to study drug effects, it is critical to show both that pathway activities accurately capture the molecular signatures of samples, and that these molecular signatures can also be mirrored in a model system allowing interrogation with therapeutic agents. Cell line studies done in collaboration with Dr. Joe Gray, Laura Heiser, Ted Goldstein, Sam Ng, and Dr. Josh Stuart allow us to establish the ability of PARADIGM to accurate model the molecular profile of breast cancer cells. Using combined copy-number profiles and exon expression measures, PARADIGM IPAs were calculated for 46 breast cancer cell lines, including lines from each intrinsic subtype. Using the PARADIGM IPAs, cell lines were clustered together with TCGA tumor samples to determine if cell lines were similar to tumor samples of the same subtype. Well-studied areas of the SuperPathway contain genes with many interactions (hubs) and large signaling chains of many intermediate complexes and abstract processes for which no direct data is available. In order to prevent noise introduced from the integration of multiple sources of arrays, pathway concepts with highly correlated vectors (Pearson correlation coefficient > 0.9) across both the cell line and tumor samples were unified into a single vector prior to clustering. This unification resulted in 2351 non-redundant vectors from the original 8768 pathway concepts.

Samples were clustered using the resulting set of non-redundant concepts. The matrix of inferred pathway activities for the 46 breast cancer cell lines and 183 TCGA

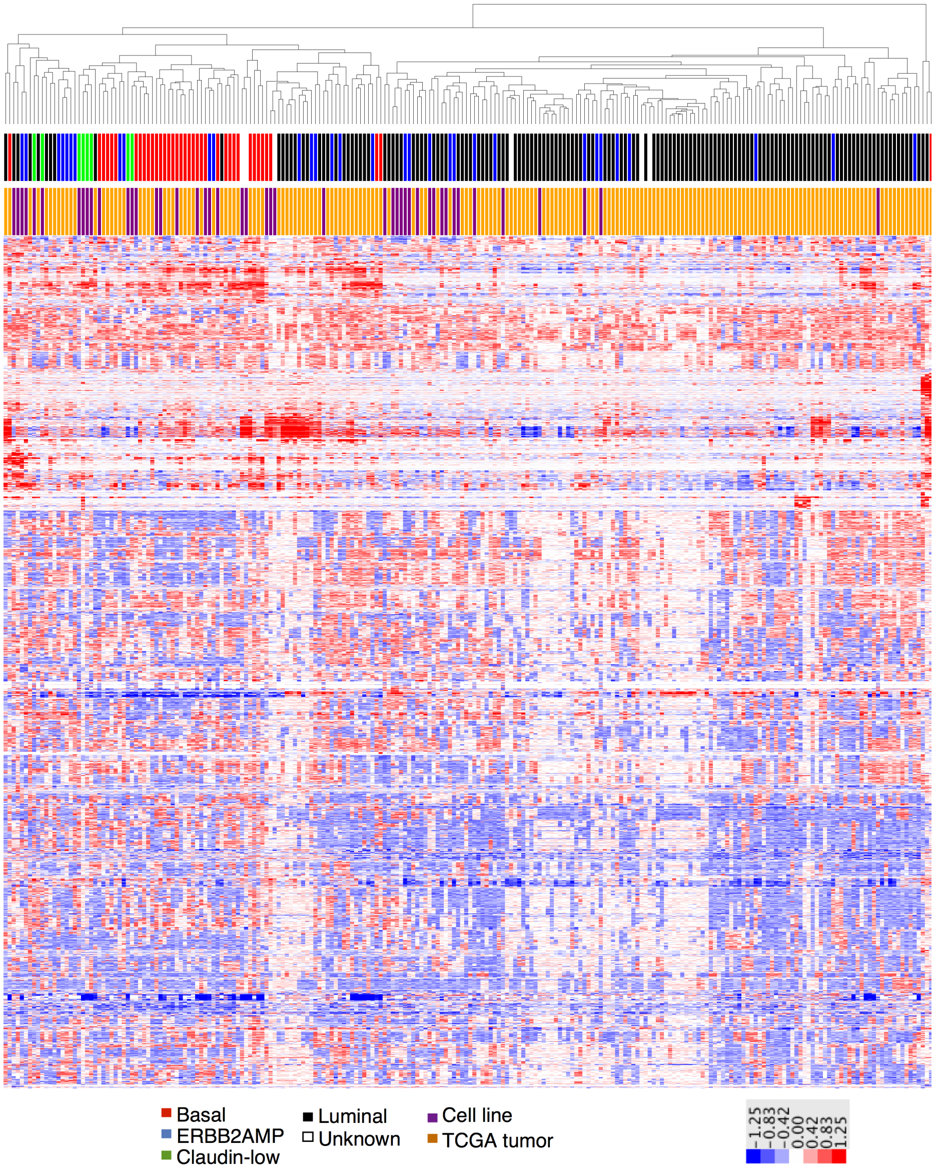


Figure 5.1: Clustering of TCGA samples with breast cancer cell lines.

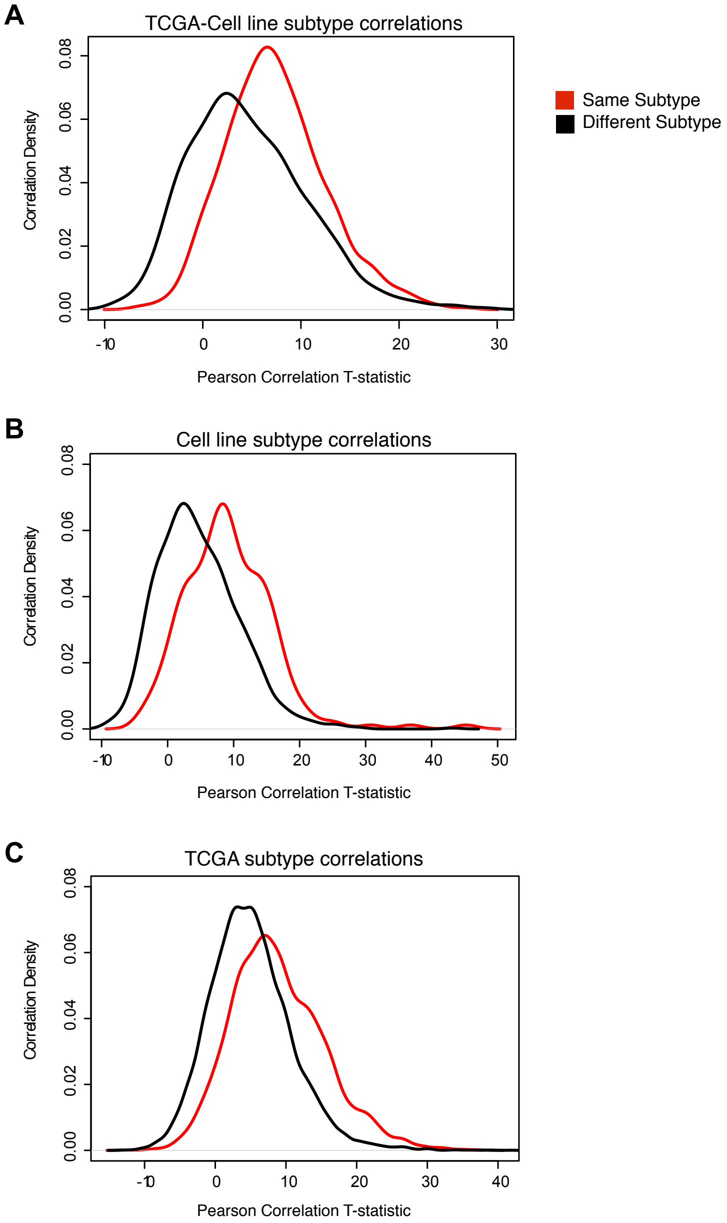


Figure 5.2: Uncentered Pearson correlation of pathway concepts show samples and cell lines from the same subtype are more correlated, regardless of the origin of the sample.

tumor samples was clustered using complete linkage hierarchical agglomerative clustering implemented in the Eisen Cluster software package version 3.0(2). Uncentered Pearson correlation was used as the metric for the pathway concepts and Euclidean distance was used for sample metric (Figure 5.1).

To quantify the degree to which cell lines clustered with tumor samples of the same subtype, we compared two distributions of t-statistics derived from Pearson correlations (Figure 5.2). Let Cs be the set of cell lines of subtype s. Similarly, let Ts be the set of TCGA tumor samples of subtype s. For example, Cbasal and Tbasal are the set of all basal cell lines and basal tumor samples respectively. The first distribution was made up of t-statistics derived from the Pearson correlations between every possible pair containing a cell line and tumor sample of the same subtype; i.e. for all subtypes s, every pairwise correlation t- statistic was computed between a pair (c, t) such that c $\in$ Cs and t $\in$ Ts. The second distribution was made of correlation t-statistics between cell lines of different subtypes; i.e. computed over pairs (c, c$'$) such that c $\in$ Cs and c$'$ $\in$ Cs$'$ and s $\neq$ s$'$. We performed a Kolmogorov-Smirnov test to compare the distributions. We repeated this analysis using samples from the same source (cell line or tumor) to verify that cells of the same subtype have overall pathway activities that are more similar than cells of different subtypes. As above, the first distribution was made up of t-statistics between pairs of samples of the same subtype and the same origin (cell line or tumor). The second distribution was made of correlation t-statistics between samples of different subtypes again from the same origin.

Focusing on the collection of breast cancer cell lines, it was possible to detec-

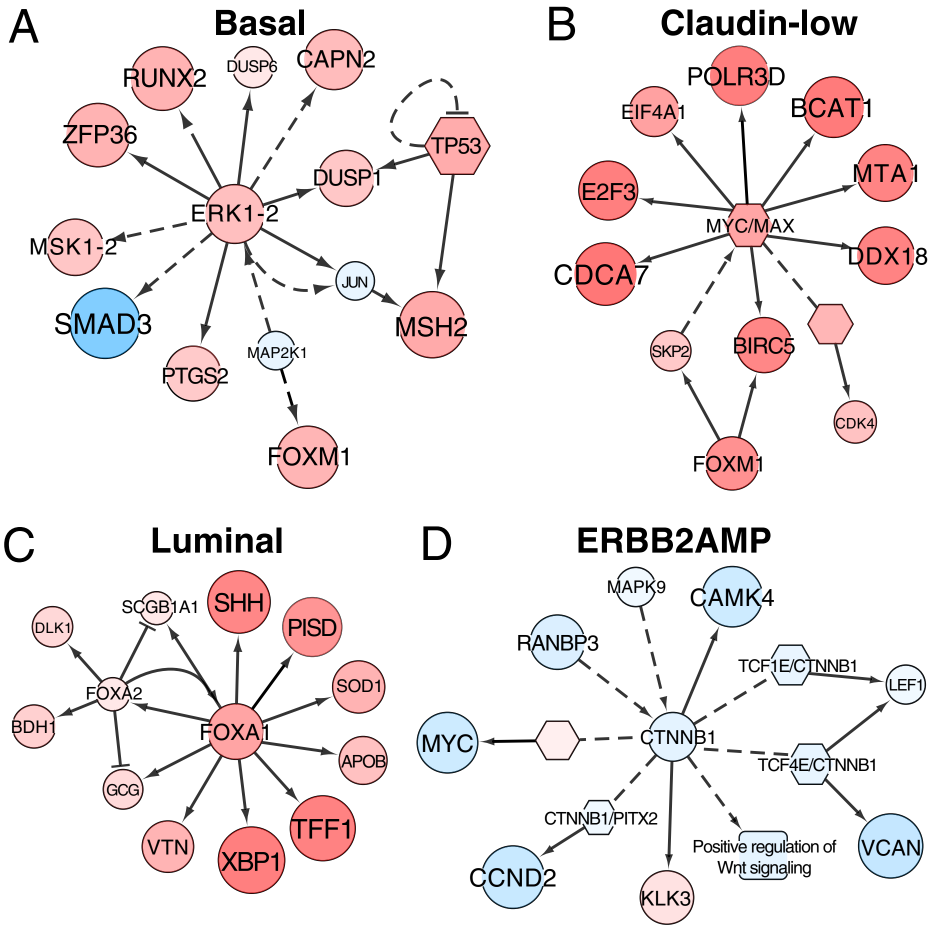


Figure 5.3: Cell-line subtypes have unique SuperPathway network features. In all panels, each node represents a pathway concept corresponding to a protein (circle), a multimeric complex (hexagon), or an abstract cellular process (square). Node sizes are drawn in proportion to the DA score; larger nodes correspond to concepts more correlated with a particular subtype than with all other subtypes. Color indicates whether the concept is correlated positively (red) or negatively (blue) with the subtype of interest. Lines represent interactions, including proteinprotein interactions (dashed lines) and transcriptional interactions (solid lines). Interactions are included if they connect concepts whose absolute level of DA is higher than the mean absolute level. Labels on some nodes are omitted for clarity. (A) An ERK1/2 subnet preferentially activated in basal breast cancer cell lines. (B) A MYC/ MAX network activated in claudin-low cell lines. (C) A FOXA1/FOXA2 net- work up-regulated in the luminal subtype. (D) A CTNNB1 subnet down-regulated in the ERBB2AMP subtype.

tion subnets that differed in activity between transcriptional subtypes. As an example, comparison of subnet activities between basal cell lines and all others in the collection identified a network comprised of 965 nodes connected by 941 edges, where nodes represent proteins, protein complexes, or cellular processes and edges represent interactions, such as protein phosphorylation, between these elements, including several subnetworks that were up- or down-regulated. Figure 5.3A, for example, shows up-regulation of the MYC/MAX subnetwork associated with metabolism, proliferation, angiogenesis, and oncogenesis[58]; and up-regulation of the ERK1/2 subnetwork controlling cell cycle, adhesion, invasion, and macrophage activation[37]. The FOXM1 and DNA damage subnetworks also were markedly up-regulated in the basal cell lines. The claudin-low network showed up-regulation of many of the same subnetworks as well as up-regulation of the beta-catenin (CTNNB1) network in Figure 5.3B, a network already implicated in tumorigenesis and associated with poor prognosis[9, 32]. Comparison of the luminal cell lines with all others showed down-regulation of an ATF2 network, which inhibits tumorigenicity in melanoma[5], and up-regulation of FOXA1/FOXA2 networks that control transcription of ER-regulated genes (Figure 5.3C) and are implicated in good prognosis luminal breast cancers[50, 46]. ERBB2AMP subnetworks were similar to those for luminal cells, which is not surprising because most ERBB2AMP cells also are classified as luminal. However, Figure 5.3D also shows down-regulation centered on RPS6KBP1 in ERBB2AMP cell lines.

SuperPathway analysis of differential drug response among the cell lines also revealed subnet activities that provide information about mechanisms of response. For

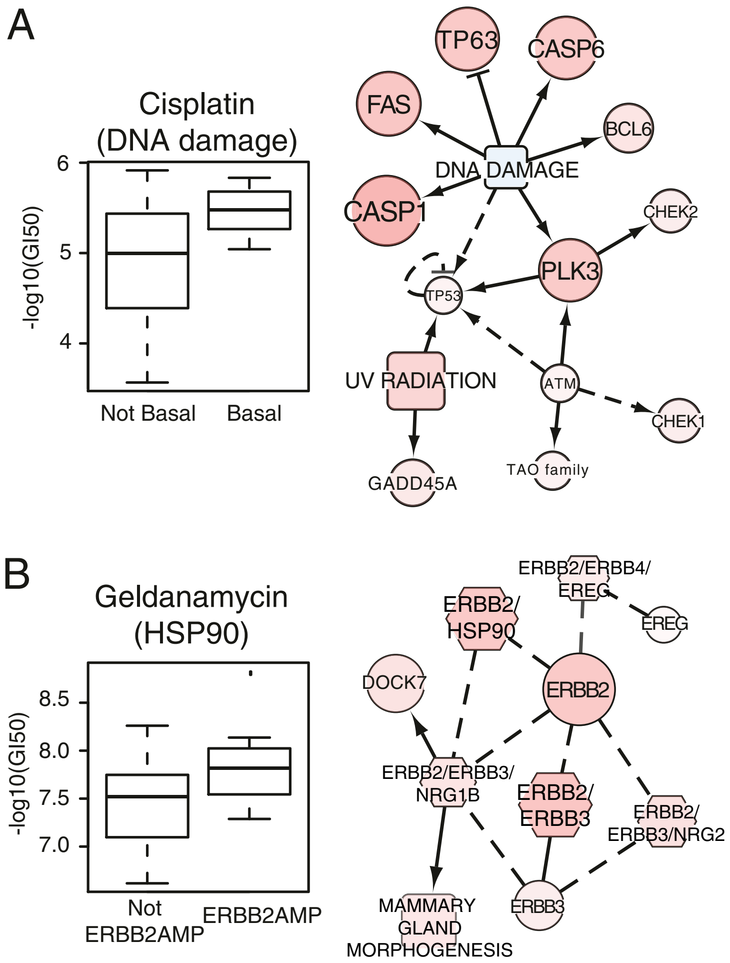


Figure 5.4: Pathway diagrams can be used to predict response to therapies. (A) (Left) Basal breast cancer cell lines respond preferentially to the DNA-dam- aging agent cisplatin. Each boxplot represents the distribution of drug response data for basal (right) and non-basal (left) cell lines. (Right) Basal cell lines show enhanced pathway levels in a subnetwork associated with the DNA-damage response, providing a possible mechanism by which cisplatin acts in these cell lines. (B) (Left) ERBB2AMP cell lines are sensitive to the HSP90 inhibitor geldanamycin. (Right) The ERBB2HSP90 network is up-regulated in ERBBP2AMP cell lines. Conventions are as in Figure 5.3.

example, basal cell line sensitivity to the DNA damaging agent, cisplatin, was associated with up-regulation of a DNA-damage response subnetwork that includes ATM, CHEK1 and BRCA1, key genes associated with response to cisplatin[65] (Figure 5.4A). Likewise, ERBB2AMP cell line sensitivity to geldanamycin (HSP90 inhibitor) was associated with up-regulation of an ERBB2-HSP90 subnetwork (Figure 5.4B). This is consistent with the known ERBB2 degradation induced by geldanamycin binding[10, 7].

The potential clinical utility of these findings is supported by concordance of in vitro-derived molecular predictors of response to therapeutic compounds and clinical results. For example, ERBB2-amplified cell lines are preferentially sensitive to ERBB2-targeted agents, and basal subtype cell lines are preferentially sensitive to platinum salts, as observed clinically. That said, additional work remains before the signatures reported in this study can be used to select patients for clinical trials. Such future work would include the development of robust and reliable molecular assays that can be applied to clinical samples, establishment of predictive algorithms with decision-making thresholds optimized for clinical use, and validation of predictive power in multiple independent studies. To initiate this process, we suggest that the response-associated signatures identified in this study be developed into standardized assays that can be assessed for clinical predictive power in early-stage clinical trials and used to design trials that are properly powered to detect the responses in the clinical subsets predicted by the in vitro studies. Assays that show positive predictive power in early clinical trials then can be locked down and tested for predictive power in follow-on clinical trials.

5.2 Culling Targets Using Machine Learning

Although the SuperPathway and subnet analysis allow us to identify potential networks that might correspond with response or sensitively, attempting to find the optimal location for combinatorial interventions requires a more automated approach. Supervised learning methods such as support vector machines (SVM) or non-negative matrix factorization (NMF) can be designed to identify these key activities, and would intrinsically provide a mechanism for assessing how well these activities map to response. By attempting to predict response to each therapy in a cross validation setting, these machine learners will attempt to identify activities that can best separate the responders and non-responders, learning weights for each feature given. By examining the weights each machine learner has assigned, and evaluating each learning in a cross-validation setting using a receiver operating characteristic (ROC) curve and calculating an area under the curve (AUC), it is possible to assign a score both to how well the learner is modeling the data corresponding to drug response, and which features were optimal for separation beween the classes. The top features within the top classifiers are ones that offer a clear separation between the two classes (responders and non-responders), and can be passed to PARADIGM in a series of knockout simulations.

5.3 Simulating Knockdowns in PARADIGM

Figure 5.5 shows the approach taken to simulate siRNA knockdowns within PARADIGM. In brief, once EM has converged a final run of paradigm is performed first

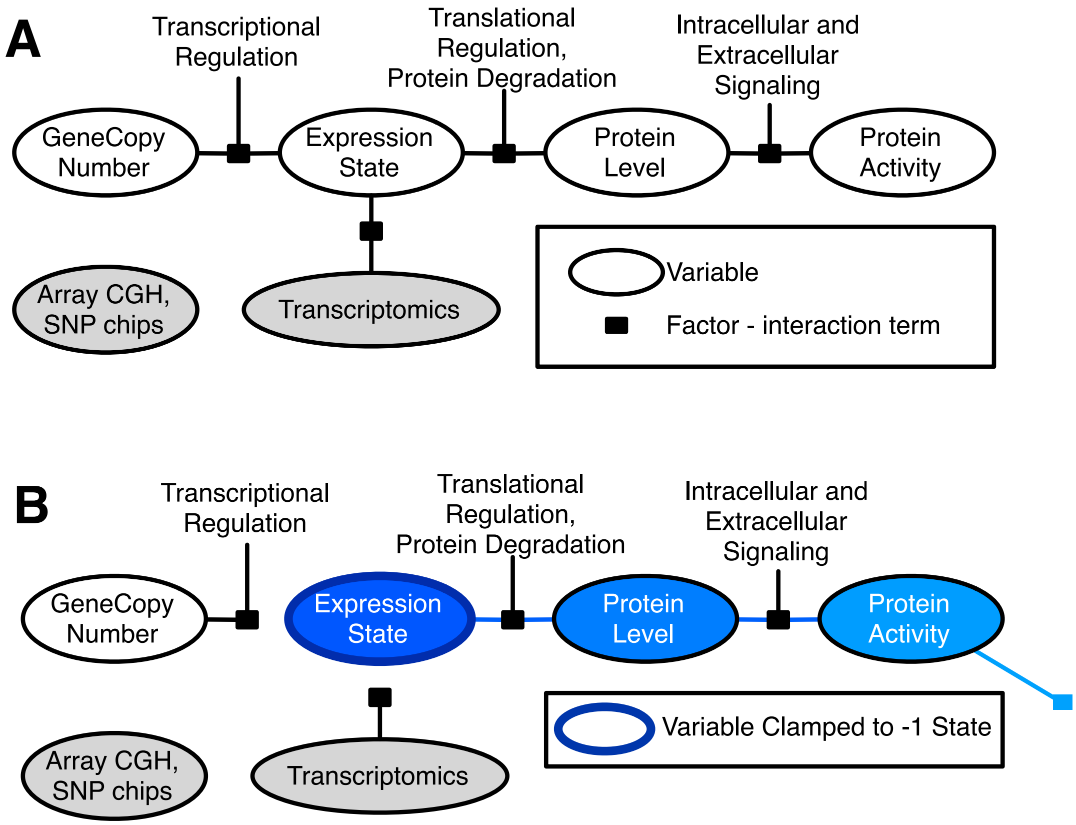

A
Transcriptional Regulation
Translational Regulation, Protein Degradation
Intracellular and Extracellular Signaling
GeneCopy Number
Expression State
Protein Level
Protein Activity
Array CGH, SNP chips
Transcriptomics
Variable
Factor - interaction term
B
Transcriptional Regulation
Translational Regulation, Protein Degradation
Intracellular and Extracellular Signaling
GeneCopy Number
Expression State
Protein Level
Protein Activity
Array CGH, SNP chips
Transcriptomics
Variable Clamped to -1 State

Figure 5.5: Schematic indicating how knockdown is performed within the factor graph central dogma. Part A represents the standard factor graph dogma model provided by PARADIGM. Part B represents the modified dogma used for the proteins that are being knocked out, with an example of how signal might propagate post KO. Of note is the disconnection of considering the mRNA evidence node and transcriptional regulation, with the resulting signal propagation to the protein node and any interactions that may occur with the knocked down node.

without the knockdown. The factor graph is then modified to clamp the mRNA node internal to the dogma structure for the specified protein node. This causes a fluctuation of probabilities in the vicinity of the node, and belief propagation is performed one final time. Although this falls short of a true causative model, the corresponding nodes in the network are sufficiently perturbed to assess the initial viability of this approach.

In order to assess if these changes correspond with actual siRNA knockdowns down in model systems, gene expression data from House et al [35] was obtained from the authors containing GFP-controls and post-knockdown expression for colon cancer cell lines treated with siRNAs covering 29 genes. The expression data was normalized for PARADIGM as previously described, and all analysis used the SuperPathway model generated from BioPAX Level 3 pathways obtained on Feb 27, 2012 from NCI-PID, BioCarta and Reactome pathway databases. Inference and EM learning were performed consistent with the previously described methods using the SuperPathway, and learning parameters converged in 5 iterations. GFP-control cell line paradigm values were recomputed using a simulated knockdown model for each of the five tier 3 gene knockdowns that had pathway context available. Samples were only kept if they had a .75 tau or larger correlation to the medoid of the other knockdowns for that same gene in the original expression data. The simulated knockdown model was computed by clamping the mRNA to the -1 state within the central dogma and regenerating inference and IPL values.

Integrated Pathway Levels (IPLs) were computed for each element in the pathway for GFP-control (GFPc), GFP-simulated-siRNA (GFPssi) and true-siRNA (siR-

Table 5.1: Correlations between simulated knockdowns and true siRNA knockdowns for a series of genes in a colon cancer cell line.

GFP	KO	KO Effectiveness %	Real	Null	Diff
GFP_scrCvsHR	GNAI3_06_1vsHR	97.61%	0.5544	0.5526	0.0018
GFP_scrAvsHR	GNAI3_06_1vsHR	95.02%	0.5475	0.5472	0.0003
GFP_scrCvsHR	GNAI3_05_2vsHR	97.13%	0.4330	0.4288	0.0043
GFP_scrCvsHR	UBE2L6_05_1vsHR	78.23%	0.4258	0.4258	0.0000
GFP_scrAvsHR	SEC24D_10_2vsHR	78.58%	0.4188	0.4188	0.0000
GFP_scrCvsHR	SEC24D_10_2vsHR	79.30%	0.4103	0.4103	0.0000
GFP_scrAvsHR	GNAI3_05_2vsHR	95.41%	0.4084	0.4058	0.0026
GFP_scrAvsHR	UBE2L6_05_1vsHR	77.94%	0.4058	0.4058	0.0000
GFP_scrCvsHR	SEC24D_12_1vsHR	77.94%	0.3744	0.3744	0.0000
GFP_scrAvsHR	GNAI3_06_2vsHR	90.69%	0.3744	0.3740	0.0003
GFP_scrCvsHR	GNAI3_06_2vsHR	93.53%	0.3703	0.3693	0.0009
GFP_scrAvsHR	SEC24D_12_1vsHR	76.97%	0.3698	0.3698	0.0000
GFP_scrAvsHR	EIF2AK2_11_1vsHR	94.73%	0.3611	0.3601	0.0010
GFP_scrCvsHR	SEC24D_12_2vsHR	74.53%	0.3425	0.3425	0.0000
GFP_scrCvsHR	EIF2AK2_11_1vsHR	93.35%	0.3242	0.3236	0.0005
GFP_scrCvsHR	SEC24D_10_1vsHR	73.88%	0.3104	0.3104	0.0000
GFP_scrAvsHR	SEC24D_12_2vsHR	75.51%	0.3093	0.3093	0.0000
GFP_scrAvsHR	SEC24D_10_1vsHR	75.31%	0.2721	0.2721	0.0000
GFP_scrCvsHR	UBE2L6_05_2vsHR	73.19%	0.1488	0.1488	0.0000
GFP_scrAvsHR	UBE2L6_06_1vsHR	73.43%	0.1361	0.1361	0.0000
GFP_scrCvsHR	SCN5A_AvsHR	75.75%	0.1299	0.1299	0.0000
GFP_scrAvsHR	UBE2L6_05_2vsHR	73.28%	0.1291	0.1291	0.0000
GFP_scrAvsHR	SCN5A_AvsHR	74.74%	0.1246	0.1246	0.0000
GFP_scrCvsHR	SCN5A_BvsHR	76.75%	0.1128	0.1128	0.0000
GFP_scrCvsHR	UBE2L6_06_1vsHR	75.02%	0.1062	0.1062	0.0000
GFP_scrAvsHR	SCN5A_BvsHR	75.92%	0.1051	0.1051	0.0000
GFP_scrAvsHR	EIF2AK2_11_2vsHR	92.97%	0.0823	0.0814	0.0009
GFP_scrCvsHR	EIF2AK2_11_2vsHR	91.15%	0.0807	0.0802	0.0005
GFP_scrCvsHR	SCN5A_CvsHR	74.02%	0.0523	0.0523	0.0000
GFP_scrAvsHR	SCN5A_CvsHR	74.79%	0.0365	0.0365	0.0000
GFP_scrCvsHR	UBE2L6_06_2vsHR	73.46%	-0.0116	-0.0116	0.0000
GFP_scrAvsHR	UBE2L6_06_2vsHR	75.31%	-0.0433	-0.0433	0.0000

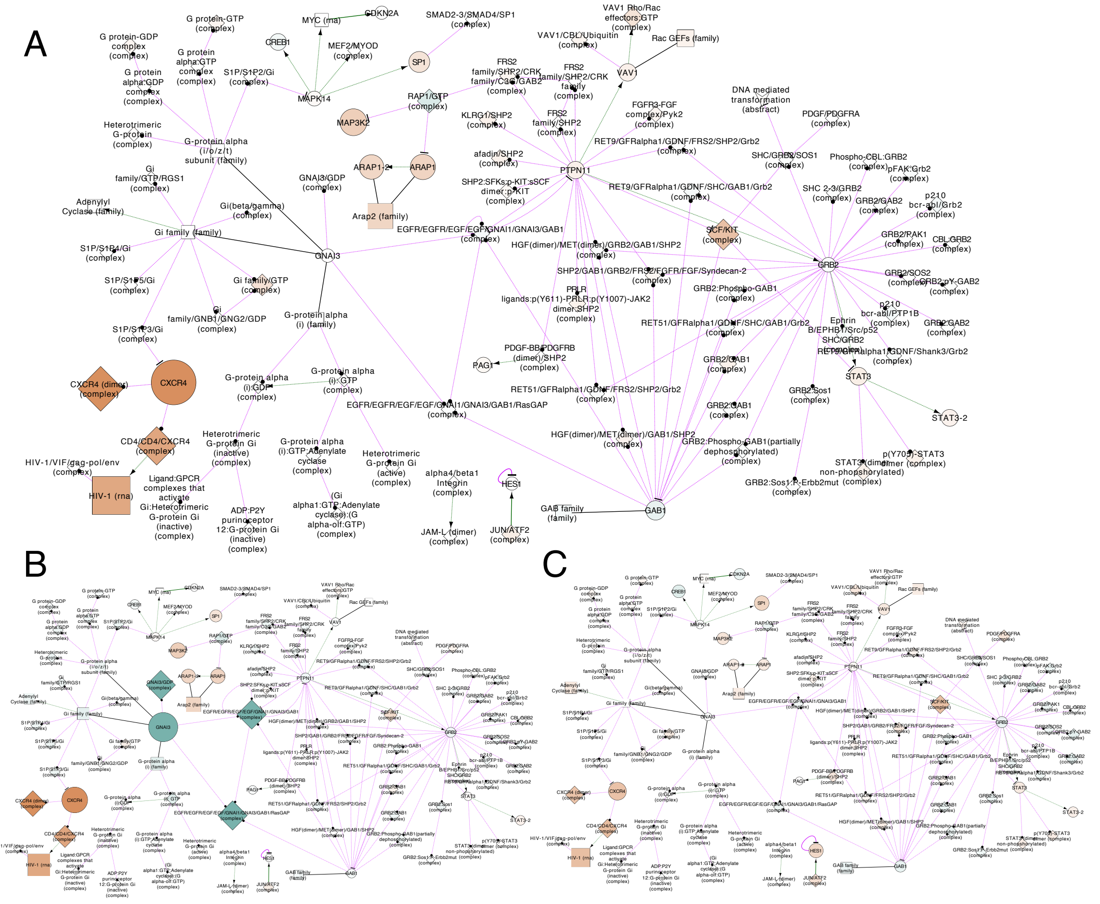

A
B
C
G protein-GTP (complex)
MYC (rna)
CDKN2A
SMAD2-3/SMAD4/SP1 (complex)
VAV1 Rho/Rac effectors/GTP (complex)
G protein-GDP complex (complex)
CREB1
MEF2/MYOD (complex)
VAV1/CBL/Ubiquitin (complex)
Rac GEFs (family)
G protein alpha:GTP complex (complex)
SP1
FRS2 family/SHP2/CRK family/C3G/GAB2 (complex)
G protein alpha:GDP complex (complex)
S1P/S1P2/Gi (complex)
MAPK14
FRS2 family/SHP2/CRK family (complex)
VAV1
RAP1/GTP (complex)
DNA mediated transformation (abstract)
FGFR3-FGF complex/Pyk2 (complex)
Heterotrimeric G-protein (complex)
MAP3K2
KLRG1/SHP2 (complex)
FRS2 family/SHP2 (complex)
PDGF/PDGFRA (complex)
G-protein alpha (i/o/z/t) subunit (family)
RET9/GFRalpha1/GDNF/FRS2/SHP2/Grb2 (complex)
ARAP1:2
ARAP1
SHC/GRB2/SOS1 (complex)
Phospho-CBL:GRB2 (complex)
Gi family/GTP/RGS1 (complex)
afadin/SHP2 (complex)
PTPN11
pFAK:Grb2 (complex)
GNAI3/GDP (complex)
SHP2:SFKs:p-KIT:sSCF dimer:p-KIT (complex)
RET9/GFRalpha1/GDNF/SHC/GAB1/Grb2 (complex)
SHC 2-3/GRB2 (complex)
p210 bcr-abl/Grb2 (complex)
Adenylyl Cyclase (family)
Gi(beta/gamma) (complex)
Arap2 (family)
GRB2/GAB2 (complex)
EGFR/EGFR/EGF/EGF/GNAI1/GNAI3/GAB1 (complex)
GRB2/PAK1 (complex)
Gi family (family)
SCF/KIT (complex)
CBL/GRB2 (complex)
S1P/S1P4/Gi (complex)
GNAI3
HGF(dimer)/MET(dimer)/GRB2/GAB1/SHP2 (complex)
GRB2
GRB2/SOS2 (complex)
GRB2:pY-GAB2 (complex)
S1P/S1P5/Gi (complex)
Gi family/GTP (complex)
SHP2/GAB1/GRB2/FRS2/FGFR/FGF/Syndecan-2 (complex)
GRB2:Phospho-GAB1 (complex)
PRLR ligands:p(Y611)-PRLR:p(Y1007)-JAK2 dimer:SHP2 (complex)
p210 bcr-abl/PTP1B (complex)
Gi family/GNB1/GNG2/GDP (complex)
G-protein alpha (i) (family)
Ephrin B/EPHB1/Src/p52 SHC/GRB2 (complex)
GRB2:GAB2 (complex)
S1P/S1P3/Gi (complex)
RET51/GFRalpha1/GDNF/SHC/GAB1/Grb2 (complex)
RET9/GFRalpha1/GDNF/Shank3/Grb2 (complex)
PDGF-BB/PDGFRB (dimer)/SHP2 (complex)
STAT3
CXCR4 (dimer) (complex)
CXCR4
PAG1
GRB2/GAB1 (complex)
G-protein alpha (i):GDP (complex)
G-protein alpha (i):GTP (complex)
RET51/GFRalpha1/GDNF/FRS2/SHP2/Grb2 (complex)
GRB2:Sos1 (complex)
EGFR/EGFR/EGF/EGF/GNAI1/GNAI3/GAB1/RasGAP (complex)
GRB2:GAB1 (complex)
STAT3-2
CD4/CD4/CXCR4 (complex)
Heterotrimeric G-protein Gi (inactive) (complex)
G-protein alpha (i):GTP:Adenylate cyclase (complex)
HGF(dimer)/MET(dimer)/GAB1/SHP2 (complex)
GRB2:Phospho-GAB1(partially dephosphorylated) (complex)
p(Y705)-STAT3 (dimer) (complex)
HIV-1/VIF/gag-pol/env (complex)
Heterotrimeric G-protein Gi (active) (complex)
STAT3 (dimer) non-phosphorylated) (complex)
HIV-1 (rna)
Ligand:GPCR complexes that activate Gi:Heterotrimeric G-protein Gi (inactive) (complex)
alpha4/beta1 Integrin (complex)
HES1
GRB2:Sos1:P-Erbb2mut (complex)
ADP:P2Y purinoceptor 12:G-protein Gi (inactive) (complex)
(Gi alpha1:GTP:Adenylate cyclase):(G alpha-olf:GTP) (complex)
GAB family (family)
GAB1
JAM-L (dimer) (complex)
JUN/ATF2 (complex)

Figure 5.6: Visualization of IPLs for the region around GNAI3 for the best Knockdown effectiveness from Table 5.1: A) GFPc of GFP_scrCvsHR, B) GFPssi of GFP_scrCvsHR and C) siRNAt of GNAI3_06_1vsHR

NAt) samples. Using a Kendall Correlation of the top 2500 variable genes across both GFPc and siRNAt samples, it was possible to assess the knockdown models ability to simulate IPLs that more closely resemble the IPLs in the true knockdowns (siRNAt). All pairwise-correlations were calculated between siRNAt and GFPc IPL vectors and directly compared to the corresponding pairwise-correlations between siRNAt and GFPssi IPL vectors, the result of which can be seen in Table 5.1. A one-sided pairwise t-test between the two populations of correlations was computed to assess the effectiveness of the simulated knockdown, which resulted in a p-value of 0.008133. Knockdown effectiveness (KE) for each simulated knockdown is calculated by simulating knockdowns for all possible genes in the pathway and measuring the change in correlation for each. The correlation changes for the simulated knockdowns were ranked among the true gene correlation change and the effectiveness is calculated as the percentage of genes resulting in worse correlations than the intended target.

Many of the knockdowns resulted in no detectable change to the resulting correlations, however for many of these the KE was still relatively high, indicating that for many genes a simulated knockdown reduces the correlation to the true siRNA activities. The top result as ranked by KE is visualized as subnetworks in figure 5.6. From the visualization, it appears as though the PTPN11 down regulation that's seen in the siRNAt sample for GNAI3 accounts for a large part of the positive correlation shift. This result illustrates that the knockdowns are able to capture a small result of the total overall changes between the networks in the immediate area of the target, highlighting the need for accurate interaction information to successfully model the total effect. Another explanation for the lack of larger correlation increases may be due to insufficient down regulation signaling by key regulatory paths, which might benefit from the implementation of interaction strengths providing more accurate signal propagation. Overall this method warrants testing on additional data and should be able to prioritize the order of targets for testing in a laboratory setting.

5.4 Breast Cell Line Simulated Knockdowns

In a series of initial experiments preparing for validation, paradigm activities were generated for the breast cell line panel using the previously discussed copy number, gene expression data, and drug sensitivity measure for 58 drugs. Four feature selection methods, 10 classification algorithms, and 2 methods for subgrouping samples into binary classes (responders vs. non-responders), were combinatorially used to build a maximum of 80 fully-trained models for each drug response prediction problem. Due to limited sampling in some drugs there was insufficient data to build the full compliment of 80 models, so in these cases only a subset of these models were assessed. All models were assessed using average accuracy in 5x5 fold cross-validation.

To create ranked lists of top candidates from these models we extracted feature weights from the most accurate linear-kernel models in TopModel, a machine learning system developed by Christopher Szeto that evaluates multiple machine learning algorithms to find the best for a particular prediction problem. These included support-vector machines [38], squential minimal optimization models [33] and NMF-based models. We justify using only linear models by showing that, overall, these linear-kernel models perform as well as higher-order models in our cross-validation experiments, and are more easily interpreted biologically. For each drug, we selected the 200 most highly weighted features (or fewer where applicable) from the most accurate linear-kernel models as candidates for likely optimal intervention points.

PARADIGM produced a new set of pathway activities that correspond to the estimated levels if the protein of each knocked down feature. To establish the effectiveness of these simulations, the vectors of resulting IPLs are applied back to the classifier for sensitivity predictions. The shift in the classification value (resistance score) is calculated; giving a score we refer to as Resistance Shift. The significance of this shift is calculated by building a background distribution from classifying the IPLs derived from knockdowns of genes picked for other drug classifications and not selected for the drug being assessed.

The top results of the classification tasks and knockdowns can be found in Table 5.2. Figure 5.8 highlights the top prediction score shift from resistant to sensitive, found for the combination of QNZ (an NF-KB inhibitor) and E2F3. Previous studies have shown that NF-KB may activate E2F family transcription factors [3], indicating a mechanism by which the cells may have developed resistance to an NF-KB therapy. Additionally, it appears as though treatments such as bortezomib, which function partially by E2F3 downregulation, act independently of NF-KB activity [61] indicating that this pair of targets may offer effectiveness in certain settings. It appears as though all resistant lines may show some shift towards sensitivity with a knockdown like this,

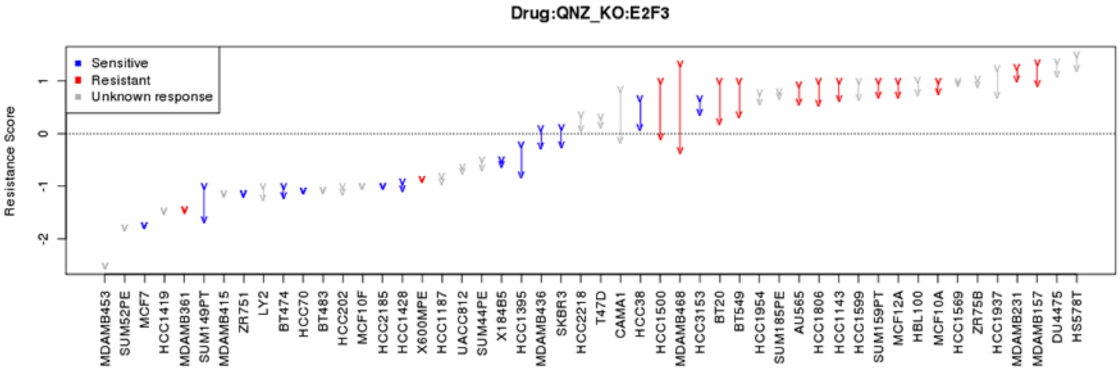


Figure 5.7: Classification of response to QNZ (an NF-KB inhibitor) in the breast cell lines, before and after simulated knockdown experiments to E2F3. Tail of the arrow represents the predicted resistance score using the pathway activities, and the head represents the prediction after simulated knockdown was performed. Blue arrows indicate cell lines that are labelled sensitive and red arrows indicate resistant cell lines.

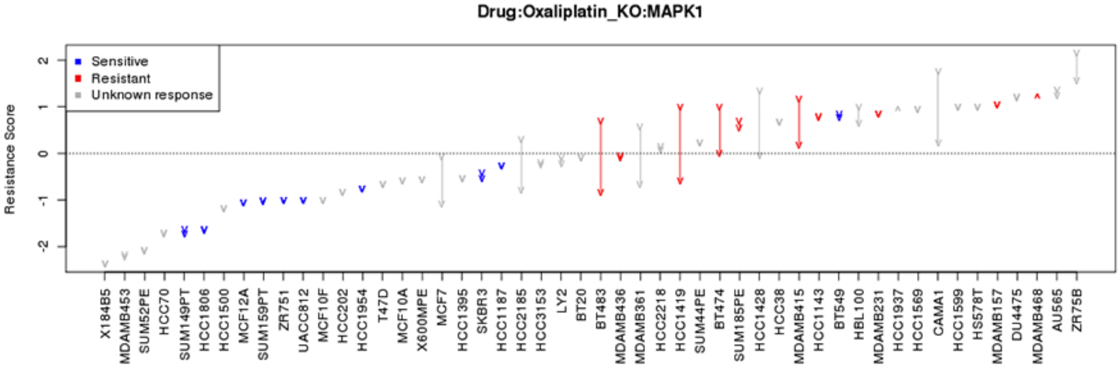


Figure 5.8: Classification of response to Oxaliplatin in the breast cell lines, before and after simulated knockdown of MAPK1.

Table 5.2: Resitance shift and p-values for the top 20 simulated KO and drug combinations ranked by shift score.

Drug	KO	Sample	PreScore	PostScore	Shift	P-Value
QNZ	E2F3	MDAMB468	1.3335	-0.327	1.6604	9.17E-144
Oxaliplatin	MAPK3	HCC1419	1	-0.5927	1.5927	1.72E-132
Oxaliplatin	FOXM1	CAMA1	1.745	0.2027	1.5423	2.01E-124
Oxaliplatin	MAPK1	BT483	0.6869	-0.8431	1.53	1.76E-122
Oxaliplatin	GREB1	ZR75B	2.1429	0.6866	1.4563	3.13E-111
Oxaliplatin	MAPK3	HCC1428	1.3238	-0.0655	1.3893	1.78E-101
Oxaliplatin	MAPK1	MDAMB361	0.5662	-0.6577	1.2239	2.41E-79
Oxaliplatin	FOXM1	HCC2185	0.2991	-0.8675	1.1667	2.48E-72
Oxaliplatin	E2F3	MDAMB468	1.2147	0.1002	1.1146	3.12E-66
Baicalein	CITED2	LY2	1	-0.0792	1.0792	2.98E-62
Oxaliplatin	FOXA1	BT474	1	-0.0738	1.0738	1.19E-61
QNZ	E2F3	HCC1500	0.9999	-0.0672	1.067	1.10E-25
QNZ	GAPDH	ZR75B	1.0292	-0.0292	1.0584	5.73E-60
Oxaliplatin	MAPK1	MCF7	-0.0621	-1.0766	1.0144	2.73E-55
Baicalein	GREB1	HCC1419	2.0221	1.0433	0.9788	1.22E-51
Oxaliplatin	MAPK1	MDAMB415	1.1615	0.1834	0.9781	1.44E-51
QNZ	E2F3	CAMA1	0.8331	-0.1257	0.9588	7.07E-05
Oxaliplatin	MYC	HBL100	1	0.0662	0.9338	3.22E-47
JH175	GAPDH	ZR75B	0.29	-0.6367	0.9267	1.55E-46
Oxaliplatin	AP1B1	LY2	-0.0936	-0.9751	0.8814	2.52E-42

however MDAMB468 shows a shift of 1.66, the largest across the entire experiment. When compared to a null model of other knockdowns and a p-value is computed, this shift is highly significant (9.18e-144).

Many of the top hits across the cell lines and treatments are specific to treatment with Oxaliplatin, a DNA damage agent. Literature searches indicate the importance of the MAPK cascade in sensitivity to platinum-based treatment across multiple tissues[34, 18], a relationship illustrated in the simulated knockdowns for 6 of the top 20 that were ranked. In addition, some of the common markers of pathway activities that differentiate basal and luminal subtypes in breast cancer appear with Oxaliplatin (FOXM1, FOXA1, etc), a result that is not surprising given the preferred sensitivity of platinum-based therapies towards basal tumors. Overall this top list of targets illustrates a good place to start experimental validation of combinational therapies, and should assist in searching an extremely large number of drug pairs in a more systematic fashion.

Chapter 6

Conclusion

6.1 Conclusions

As technological advances have increased the amount of data produced by single experiments to thousands of gigabytes, it's becoming increasingly clear the bottleneck in making clinically relevant discoveries is distilling this information to understandable levels. The overall approach of PARADIGM aims to provide an intuitive and scalable framework to understand these data in biologically relevant contexts in a patient-specific manner. Combining the PARADIGM engine with the power of the SuperPathway's representation of the cellular interactome offers an unprecedented ability to understand the heterogeneity that exists within cancer. This method also moves us away from preconceived curated pathway models that contain large amounts of study bias and towards models that can accommodate the complexities of cellular signaling and transcription networks. This is particularly critical in cancer where much of the

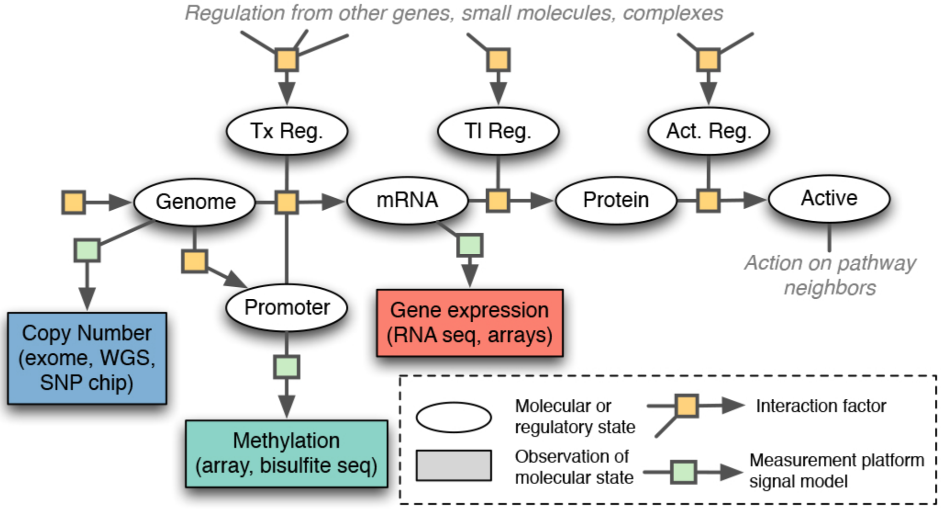

Regulation from other genes, small molecules, complexes
Tx Reg.
Tl Reg.
Act. Reg.
Genome
mRNA
Protein
Active
Action on pathway neighbors
Promoter
Copy Number (exome, WGS, SNP chip)
Gene expression (RNA seq, arrays)
Methylation (array, bisulfite seq)
Molecular or regulatory state
Interaction factor
Observation of molecular state
Measurement platform signal model

Figure 6.1: An example expanded PARADIGM dogma model that accounts for promoter states and allows attachment of Methylation data. (©IEEE 2012)

misregulation occurs by bypassing standard cellular protective mechanisms, the understanding of which is necessary for understanding which therapies may be effective. The PARADIGM simulated knockdown model will aid our thinking about combinational therapies in help refine pathway models as functional testing begins. Overall PARADIGM's greatest strength is the extremely powerful and highly extensible engine that is provided to model biology in intuitive ways that reduce the time it takes to test biological phenomenon in a computational framework.

6.2 Future Expansions

Despite all of the work done to date, there are many enhancements to PARADIGM that should increase our modeling resolution to provide more accurate results. The incorporation of the additional measurements currently taken about a particular cancer is still necessary to move to a more accurate model within PARADIGM. These include DNA Methylation from the Illumina methylation arrays, mutational information from sequencing, microRNA levels, and proteomics assays. Each of these datatypes presents unique challenges to modeling that must be overcome, and one example of how we might model methylation data can be seen in Figure 6.1.

For example, “driver” mutations within cancer are typically viewed as hitting two types of proteins, tumor suppressors and oncogenes. If a mutation hits a tumor suppressor, the typical result is to inactivate the protein so that it can no longer perform its function. When we go to model these types of interactions in PARADIGM, we will want to assume mutations to tumor suppressors are large negatively activating observations. However, when mutations hit oncogenes, they typically hit a consistent domain across multiple patients and tumors, and generally cause the protein to be constitutively activated. This would require modeling a large positively activation observation for the protein that contains the mutation. In addition, there are a vast number of “passenger” mutations that will be carried clonally in the tumor genome with the small number of “driver” mutations. It may be possible to distinguish these two classes of mutations using PARADIGM by sorting passengers into irrelevant pathways, with drivers only appearing in cancer-dependent pathways.

MicroRNAs (miRNAs) present another technical challenge to modeling within PARADIGM. It is clear that a single miRNA has the potential to modulate the transcript abundance of hundreds of different proteins. Simply attempting to add miRNA levels to the model will likely result in highly linked networks that are no longer computationally tractable, and it may be necessary to find functional subsets of miRNA targets that can be grouped together.

With the recent advances in sequencing technology, it's rapidly becoming affordable to sequence the entire genome and transcriptome of a cancer sample. Although pathway modeling is generally not restricted to a particular type of data used for observational variables, all of the work to this point has been done on array-based technologies. These technologies provide measurements of RNA and DNA abundance in relative light intensities, allowing researchers to compare these between samples after standard normalization techniques. Moving away from array-based measurements into sequencing-based measurements is going to require modifying these normalization techniques, but will ultimately provide a higher resolution of the biology and should be capable of better informing the model. More importantly, sequencing-based measurements allow modeling of gene fusions, splice variants, and allele-specific alterations which may modify the underlying pathway structure. For example, despite our ability to identify FOXM1 as a critical transcription factor in ovarian cancer, it is clear that we will never be able to fully elucidate the mechanism of misregulation without isoform-specific transcript levels provided by RNA-seq.

Sequencing data also may enable the ability to model allele-specific variations and potential gene fusions that result in dysregulated pathways. Linking the results of sequence analysis pipelines like BamBam, which can be used to detect allele specific variations, to the pathway model may provide interesting evidence that helps probabilistically explain discrepancies in the pathway. In order to model each allele, it will

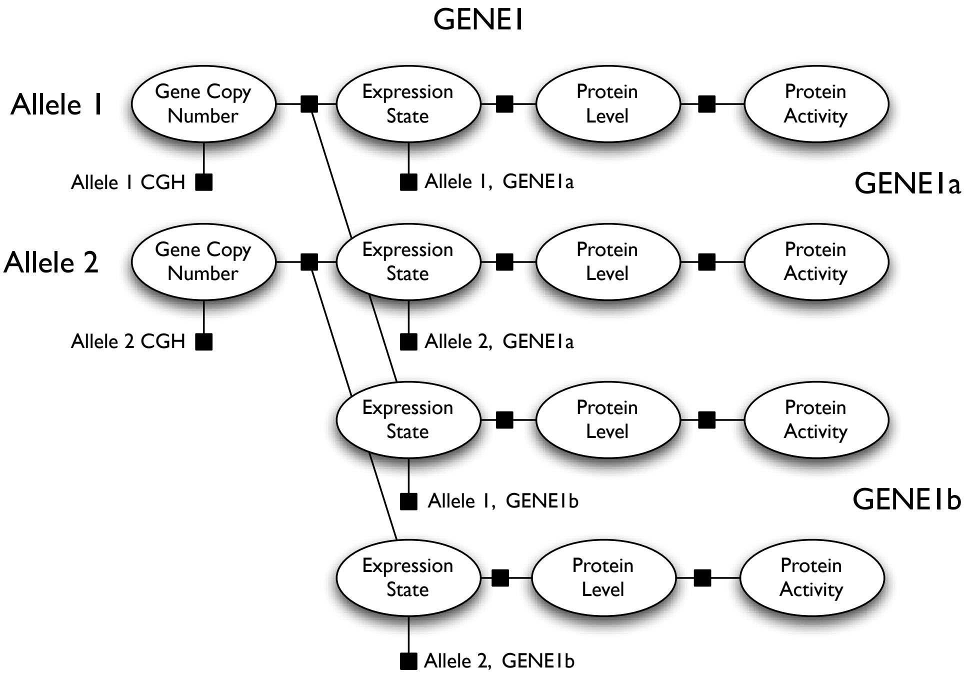


Figure 6.2: Allele and Isoform Specific Factor Graph Representation. An example of how we might represent a gene ("GENE1") with allele-specific data and two isoforms.

be necessary to expand our current "central dogma" representation to an allele-specific dogma. This new model will have genome, mRNA, and protein nodes specific to each allele represented. An example of what this representation might look like is shown in figure 6.2. If BamBam identifies an allele specific mutation, PARADIGM can dynamically expand the standard dogma to this new variant, allowing inference on each allele and eventually trying to consolidate those alleles at the active protein node. By linking this information with functional mutation algorithms being developed by collaborators, PARADIGM can further expand this to remove or create direct interactions depending on the mutational type. For example, if a mutation is detected that might disable a substrate binding site, PARADIGM could disconnect the appropriate interactions in an attempt to consolidate the data and increase the likelihood of the model.

Beyond allele-specific variations, transcriptome sequencing data will allow PARADIGM to identify gene fusions and splice-specific isoforms. As a simplistic initial analysis, we can take the genes fused and form a new network with the union of the two pathways in which each gene appears. This new pathway will represent the potential chimeric function of this new protein, and we can assess the validity of this pathway compared to each of the individual pathways by comparing the likelihood of the data given the pathway structure. In addition, since splice variants are already annotated in our pathway models, having splice-specific data we can attach will allow us to more accurately represent the observational nodes for those isoforms. The combination of these new types of data will hopefully build on PARADIGM's ability to identify critical pathways within each tumor.

6.3 Final Thoughts

As we begin to unravel the molecular origin of each cancer, it becomes clear that despite their shared hallmarks, no two cancers are misregulated using the same combination of genes. Establishing molecular fingerprints for each tumor from pathway models such as Paradigm is the only way to sanely approach treatment as we move towards the future of personalized medicine. Despite the early stage, having the computational power to simulate response to therapy on an individual basis should prove to vastly increase the speed at which therapies are developed and tested. This excitement of a new era where discovery and treatment are more closely coupled may prove to be the method by which the war on cancer is finally won.

Bibliography

[1] A A Alizadeh, M B Eisen, R E Davis, C Ma, I S Lossos, A Rosenwald, J C Boldrick, H Sabet, T Tran, X Yu, J I Powell, L Yang, G E Marti, T Moore, J Hudson, L Lu, D B Lewis, R Tibshirani, G Sherlock, W C Chan, T C Greiner, D D Weisenburger, J O Armitage, R Warnke, R Levy, W Wilson, M R Grever, J C Byrd, D Botstein, P O Brown, and L M Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–11, Feb 2000.

[2] David B Allison, Xiangqin Cui, Grier P Page, and Mahyar Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*, 7(1):55–65, Jan 2006.

[3] K Araki, K Kawauchi, and N Tanaka. IKK/NF-kappaB signaling pathway inhibits cell-cycle progression by a novel Rb-independent suppression system for E2F transcription factors. *Oncogene*, 27(43):5696–5705, September 2008.

[4] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics, Published online: 02 November 2008; — doi:10.1038/ng.259*, 25(1):25–9, May 2000.

[5] Sunil Badve, Dmitry Turbin, Mangesh A Thorat, Akira Morimiya, Torsten O Nielsen, Charles M Perou, Sandi Dunn, David G Huntsman, and Harikrishna Nakshatri. FOXA1 expression in breast cancer–correlation with luminal subtype A and survival. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 13(15 Pt 1):4415–4421, August 2007.

[6] A M Barsotti and C Prives. Pro-proliferative foxm1 is a target of p53-mediated repression. *Oncogene*, 28(48):4295–305, Dec 2009.

[7] Jose Baselga and Sandra M Swain. Novel anticancer targets: revisiting ERBB2 and discovering ERBB3. *Nature reviews Cancer*, 9(7):463–475, July 2009.

[8] Michael A Beer and Saeed Tavazoie. Predicting gene expression from sequence. *Cell*, 117(2):185–98, Apr 2004.

[9] Anindita Bhoumik, Nic Jones, and Ze'ev Ronai. Transcriptional switch by activating transcription factor 2-derived peptide sensitizes melanoma cells to apoptosis and inhibits their tumorigenicity. *Proceedings of the National Academy of Sciences of the United States of America*, 101(12):4222–4227, March 2004.

[10] M V Blagosklonny. Hsp-90-associated oncoproteins: multiple targets of geldanamycin and its analogs. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K*, 16(4):455–462, April 2002.

[11] Reinhard Bos, Petra van der Groep, Astrid E Greijer, Avi Shvarts, Sybren Meijer, Herbert M Pinedo, Gregg L Semenza, Paul J van Diest, and Elsken van der Wall. Levels of hypoxia-inducible factor-1alpha independently predict prognosis in patients with lymph node negative breast carcinoma. *Cancer*, 97(6):1573–1581, March 2003.

[12] Ana Bosch, Pilar Eroles, Rosa Zaragoza, Juan R Viña, and Ana Lluch. Triple-negative breast cancer: molecular features, pathogenesis, treatment and current lines of research. *Cancer treatment reviews*, 36(3):206–215, May 2010.

[13] Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, June 2011.

[14] Ethan G Cerami, Gary D Bader, Benjamin E Gross, and Chris Sander. cpath: open source software for collecting, storing, and querying biological pathways. *BMC Bioinformatics 2009 10:203*, 7:497, Jan 2006.

[15] H Chai and C Domeniconi. An evaluation of gene selection methods for multi-class microarray data classification. *... European Workshop on Data Mining and Text ...*, Jan 2004.

[16] Goodarz Danaei, Stephen Vander Hoorn, Alan D Lopez, Christopher J L Murray, Majid Ezzati, and Comparative Risk Assessment collaborating group (Cancers). Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. *Lancet*, 366(9499):1784–93, Nov 2005.

[17] Alexandros Daponte, Maria Ioannou, Ilias Mylonis, George Simos, Marcos Minas, Ioannis E Messinis, and George Koukoulis. Prognostic significance of hypoxia-inducible factor 1 alpha(hif-1 alpha) expression in serous ovarian cancer: an immunohistochemical study. *BMC Cancer*, 8:335, Jan 2008.

[18] M A de la Cruz-Morcillo, M L L Valero, J L Callejas-Valera, L Arias-González, P Melgar-Rojas, E M Galán-Moya, E García-Gil, J García-Cano, and R Sánchez-Prieto. P38MAPK is a major determinant of the balance between apoptosis and autophagy triggered by 5-fluorouracil: implication in resistance. *Oncogene*, 31(9):1073–1085, August 2011.

[19] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, January 1977.

[20] Sandrine Dudoit and Jane Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol*, 3(7):RESEARCH0036, Jun 2002.

[21] S Efroni, CF Schaefer, and KH Buetow. Identification of key processes underlying cancer phenotypes using biological pathway analysis. *PLoS ONE*, 2(5), 2007.

[22] Dariush Etemadmoghadam, Anna Defazio, Rameen Beroukhim, Craig Mermel, Joshy George, Gad Getz, Richard Tothill, Aikou Okamoto, Maria B Raeder, Paul Harnett, Stephen Lade, Lars A Akslen, Anna V Tinker, Bianca Locandro, Kathryn Alsop, Yoke-Eng Chiew, Nadia Traficante, Sian Fereday, Daryl Johnson, Stephen Fox, William Sellers, Mitsuyoshi Urashima, Helga B Salvesen, Matthew Meyerson, David Bowtell, and AOCS Study Group. Integrated genome-wide dna copy number and expression analysis identifies distinct mechanisms of primary chemoresistance in ovarian carcinomas. *Clin Cancer Res*, 15(4):1417–27, Feb 2009.

[23] D A Fell and J R Small. Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *The Biochemical journal*, 238(3):781–786, September 1986.

[24] Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, Feb 2004.

[25] Nir Friedman and Moises Goldszmidt. Sequential update of bayesian network structure. *In Proc. 13th Conference on Uncertainty in Artificial Intelligence (UAI'97)*, pages 165–174, Feb 1997.

[26] Irit Gat-Viks and Ron Shamir. Refinement and expansion of signaling pathways: The osmotic response network in yeast. *Genome Research*, 17(3):358–367, 2007.

[27] Irit Gat-Viks, Amos Tanay, Daniela Raijman, and Ron Shamir. The factor graph network model for biological systems. *RECOMB*, pages 31–47, 2005.

[28] Irit Gat-Viks, Amos Tanay, Daniela Raijman, and Ron Shamir. A probabilistic methodology for integrating knowledge and experiments on biological networks. *Journal of Computational Biology*, 13(2):165–181, March 2006.

[29] T A Gatcliffe, B J Monk, K Planutis, and R F Holcombe. Wnt signaling in ovarian tumorigenesis. *International journal of gynecological cancer : official journal of the International Gynecological Cancer Society*, 18(5):954–962, August 2008.

[30] S A Gayther, S J Batley, L Linger, A Bannister, K Thorpe, S F Chin, Y Daigo, P Russell, A Wilson, H M Sowter, J D Delhanty, B A Ponder, T Kouzarides, and C Caldas. Mutations truncating the ep300 acetylase in human cancers. *Nature*

Genetics, Published online: 02 November 2008; — doi:10.1038/ng.259, 24(3):300–3, Mar 2000.

[31] T R Golub, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, M L Loh, J R Downing, M A Caligiuri, C D Bloomfield, and E S Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–7, Oct 1999.

[32] Hany Onsy Habashy, Desmond G Powe, Emad A Rakha, Graham Ball, Claire Paish, Julia Gee, Robert I Nicholson, and Ian O Ellis. Forkhead-box A1 (FOXA1) expression in breast cancer and its prognostic significance. *European journal of cancer (Oxford, England : 1990)*, 44(11):1541–1551, July 2008.

[33] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, November 2009.

[34] Javier Hernández Losa, Carlos Parada Cobo, Juan Guinea Viniegra, Victor Javier Sánchez-Arevalo Lobo, Santiago Ramón y Cajal, and Ricardo Sánchez-Prieto. Role of the p38 MAPK pathway in cisplatin-based therapy. *Oncogene*, 22(26):3998–4006, June 2003.

[35] C D House, C J Vaske, A M Schwartz, V Obias, B Frank, T Luu, N Sarvazyan, R Irby, R L Strausberg, T G Hales, J M Stuart, and N H Lee. Voltage-Gated Na+ Channel SCN5A Is a Key Regulator of a Gene Transcriptional Network That Controls Colon Cancer Invasion. *Cancer research*, 70(17):6957–6967, August 2010.

[36] Antoni Hurtado, Kelly A Holmes, Caryn S Ross-Innes, Dominic Schmidt, and Jason S Carroll. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nature Genetics*, 43(1):27–33, January 2011.

[37] Nancy E Hynes and Tina Stoelzle. Key signalling nodes in mammary gland development and cancer: Myc. *Breast cancer research : BCR*, 11(5):210, 2009.

[38] T Joachims. Making large-scale support vector machine learning practical. *Advances in Kernel Methods*, Jan 1999.

[39] G Joshi-Tope, M Gillespie, I Vastrik, P D'Eustachio, E Schmidt, B de Bono, B Jassal, G R Gopinath, G R Wu, L Matthews, S Lewis, E Birney, and L Stein. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*, 33(Database issue):D428–32, Jan 2005.

[40] Alison M Karst, Keren Levanon, and Ronny Drapkin. Modeling high-grade serous ovarian carcinogenesis from the fallopian tube. *Proceedings of the National Academy of Sciences of the United States of America*, 108(18):7547–7552, May 2011.

[41] M K Kerr, M Martin, and G A Churchill. Analysis of variance for gene expression microarray data. *J Comput Biol*, 7(6):819–37, Jan 2000.

[42] S Knuutila, Y Aalto, K Autio, A M Björkqvist, W El-Rifai, S Hemmer, T Huhta, E Kettunen, S Kiuru-Kuhlefelt, M L Larramendy, T Lushnikova, O Monni, H Pere, J Tapper, M Tarkkanen, A Varis, V M Wasenius, M Wolf, and Y Zhu. Dna copy number losses in human neoplasms. *Am J Pathol*, 155(3):683–94, Sep 1999.

[43] F Kschischang, B Frey, and H Loeliger. Factor graphs and the sum-product algorithm. *Information Theory, IEEE Transactions on*, 47(2):498 – 519, 2001.

[44] C T Kuan, C J Wikstrand, and D D Bigner. Egf mutant receptor viii as a molecular target in cancer therapy. *Endocr Relat Cancer*, 8(2):83–96, Jun 2001.

[45] Su-In Lee, Dana Pe'er, Aimée M Dudley, George M Church, and Daphne Koller. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc Natl Acad Sci USA*, 103(38):14062–7, Sep 2006.

[46] S Y Lin, W Xia, J C Wang, K Y Kwong, B Spohn, Y Wen, R G Pestell, and M C Hung. Beta-catenin, a novel prognostic marker for breast cancer: its roles in cyclin D1 expression and cancer progression. *Proceedings of the National Academy of Sciences of the United States of America*, 97(8):4262–4266, April 2000.

[47] William M Lin, Alissa C Baker, Rameen Beroukhim, Wendy Winckler, Whei Feng, Jennifer M Marmion, Elisabeth Laine, Heidi Greulich, Hsiuyi Tseng, Casey Gates, F Stephen Hodi, Glenn Dranoff, William R Sellers, Roman K Thomas, Matthew Meyerson, Todd R Golub, Reinhard Dummer, Meenhard Herlyn, Gad Getz, and Levi A Garraway. Modeling genomic diversity and tumor dependency in malignant melanoma. *Cancer Research*, 68(3):664–73, Feb 2008.

[48] Richard Y M Ma, Tommy H K Tong, Alice M S Cheung, Anthony C C Tsang, Wai Ying Leung, and Kwok-Ming Yao. Raf/mek/mapk signaling stimulates the nuclear translocation and transactivating activity of foxm1c. *J Cell Sci*, 118(Pt 4):795–806, Feb 2005.

[49] Rutika J Mehta, Rohit K Jain, Samuel Leung, Jennifer Choo, Torsten Nielsen, David Huntsman, Harikrishna Nakshatri, and Sunil Badve. FOXA1 is an independent prognostic marker for ER-positive breast cancer. *Breast cancer research and treatment*, 131(3):881–890, February 2012.

[50] J S Michaelson and P Leder. beta-catenin is a downstream effector of Wnt-mediated tumorigenesis in the mammary gland. *Oncogene*, 20(37):5093–5099, August 2001.

[51] Kevin P Murphy, Yair Weiss, and Michael I Jordan. Loopy belief propagation for approximate inference: An empirical study. *In Proceedings of Uncertainty in AI*, pages 467–475, May 1999.

[52] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–8, Oct 2008.

[53] H Ogata, S Goto, K Sato, W Fujibuchi, H Bono, and M Kanehisa. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 27(1):29–34, Jan 1999.

[54] Philipp Pagel, Stefan Kovac, Matthias Oesterheld, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Pekka Mark, Volker Stümpflen, Hans-Werner Mewes, Andreas Ruepp, and Dmitrij Frishman. The mips mammalian protein-protein interaction database. *Bioinformatics*, 21(6):832–4, Mar 2005.

[55] John W Park, Richard M Neve, Janos Szollosi, and Christopher C Benz. Unraveling the biologic and clinical complexities of her2. *Clin Breast Cancer*, 8(5):392–401, Oct 2008.

[56] D Williams Parsons, Siân Jones, Xiaosong Zhang, Jimmy Cheng-Ho Lin, Rebecca J Leary, Philipp Angenendt, Parminder Mankoo, Hannah Carter, I-Mei Siu, Gary L Gallia, Alessandro Olivi, Roger McLendon, B Ahmed Rasheed, Stephen Keir, Tatiana Nikolskaya, Yuri Nikolsky, Dana A Busam, Hanna Tekleab, Luis A Diaz, James Hartigan, Doug R Smith, Robert L Strausberg, Suely Kazue Nagahashi Marie, Sueli Mieko Oba Shinjo, Hai Yan, Gregory J Riggins, Darell D Bigner, Rachel Karchin, Nick Papadopoulos, Giovanni Parmigiani, Bert Vogelstein, Victor E Velculescu, and Kenneth W Kinzler. An integrated genomic analysis of human glioblastoma multiforme. *Science*, 321(5897):1807–12, Sep 2008.

[57] Pradip Raychaudhuri and Hyun Jung Park. FoxM1: a master regulator of tumor metastasis. *Cancer research*, 71(13):4329–4333, July 2011.

[58] P J Roberts and C J Der. Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. *Oncogene*, 26(22):3291–3310, May 2007.

[59] Kwanghee Ryu, Choa Park, and Youngjoo Lee. Hypoxia-inducible factor 1 alpha represses the transcription of the estrogen receptor alpha gene in human breast cancer cells. *Biochemical and biophysical research communications*, 407(4):831–836, April 2011.

[60] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–9, Apr 2005.

[61] Kristopher A Sarosiek, Lucas E Cavallin, Shruti Bhatt, Ngoc L Toomey, Yasodha Natkunam, Wilfredo Blasini, Andrew J Gentles, Juan Carlos Ramos, Enrique A Mesri, and Izidore S Lossos. Efficacy of bortezomib in a direct xenograft model of primary effusion lymphoma. *Proceedings of the National Academy of Sciences of the United States of America*, 107(29):13069–13074, July 2010.

[62] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. Pid: the pathway interaction database. *Nucleic acids research*, 37(Database issue):D674–9, Jan 2009.

[63] Eran Segal, Nir Friedman, Naftali Kaminski, Aviv Regev, and Daphne Koller. From signatures to models: understanding cancer using microarrays. *Nature Genetics, Published online: 02 November 2008; — doi:10.1038/ng.259*, 37 Suppl:S38–45, Jun 2005.

[64] Kazushi Shigemasa, Lijun Gu, Timothy J O'Brien, and Koso Ohama. Skp2 overexpression is a prognostic factor in patients with ovarian adenocarcinoma. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 9(5):1756–1763, May 2003.

[65] Zahid H Siddik. Cisplatin: mode of cytotoxic action and molecular basis of resistance. *Oncogene*, 22(47):7265–7279, October 2003.

[66] Michael E Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics (Oxford, England)*, 27(3):431–432, February 2011.

[67] T Sørlie, C M Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, M B Eisen, M van de Rijn, S S Jeffrey, T Thorsen, H Quist, J C Matese, P O Brown, D Botstein, P Eystein Lønning, and A L Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19):10869–10874, September 2001.

[68] R C Stein. Prospects for phosphoinositide 3-kinase inhibition as a cancer treatment. *Endocr Relat Cancer*, 8(3):237–48, Sep 2001.

[69] John D Storey and Robert Tibshirani. Statistical methods for identifying differentially expressed genes in dna microarrays. *Methods Mol Biol*, 224:149–57, Jan 2003.

[70] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, 102(43):15545–50, Oct 2005.

[71] P Tamayo, D Slonim, J Mesirov, Q Zhu, S Kitareewan, E Dmitrovsky, E S Lander, and T R Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA*, 96(6):2907–12, Mar 1999.

[72] Yongjun Tan, Pradip Raychaudhuri, and Robert H Costa. Chk2 mediates stabilization of the foxm1 transcription factor to stimulate expression of dna repair genes. *Mol Cell Biol*, 27(3):1007–16, Feb 2007.

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

[74] TCGA. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

[75] Richard Tothill, Anna Tinker, Joshy George, Robert Brown, Stephen Fox, Stephen Lade, Daryl Johnson, Melanie Trivett, Dariush Etemadmoghadam, Bianca Locandro, Nadia Traficante, Sian Fereday, Jillian Hung, Yoke-Eng Chiew, Izhak Haviv, Australian Ovarian Cancer Study Group, Australian Ovarian Cancer Study Group, Dorota Gertig, Anna Defazio, and David Bowtell. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical Cancer Research*, 14(16):5198, Aug 2008.

[76] X B Trinh, W A A Tjalma, P B Vermeulen, G Van den Eynden, I Van der Auwera, S J Van Laere, J Helleman, E M J J Berns, L Y Dirix, and P A van Dam. The VEGF pathway and the AKT/mTOR/p70S6K1 signalling pathway in human epithelial ovarian cancer. *British journal of cancer*, 100(6):971–978, March 2009.

[77] Olga G Troyanskaya, Mitchell E Garber, Patrick O Brown, David Botstein, and Russ B Altman. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18(11):1454–61, Nov 2002.

[78] V G Tusher, R Tibshirani, and G Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, 98(9):5116–21, Apr 2001.

[79] Marc J van de Vijver, Yudong D He, Laura J van't Veer, Hongyue Dai, Augustinus A M Hart, Dorien W Voskuil, George J Schreiber, Johannes L Peterse, Chris Roberts, Matthew J Marton, Mark Parrish, Douwe Atsma, Anke Witteveen, Annuska Glas, Leonie Delahaye, Tony van der Velde, Harry Bartelink, Sjoerd Rodenhuis, Emiel T Rutgers, Stephen H Friend, and René Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 347(25):1999–2009, Dec 2002.

[80] V Varadan, P Mittal, C J Vaske, and S C Benz. The Integration of Biological Pathway Knowledge in Cancer Genomics: A review of existing computational approaches. *Signal Processing Magazine, IEEE*, 29(1):35–50, 2012.

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

[82] C Vogel, M A Cobleigh, D Tripathy, J C Gutheil, L N Harris, L Fehrenbacher, D J Slamon, M Murphy, W F Novotny, M Burchmore, S Shak, and S J Stewart.

First-line, single-agent herceptin(r) (trastuzumab) in metastatic breast cancer. a preliminary report. *Eur J Cancer*, 37 Suppl 1:25–29, Jan 2001.

[83] K A Voutilainen, M A Anttila, S M Sillanpää, K M Ropponen, S V Saarikoski, M T Juhola, and V-M Kosma. Prognostic significance of E-cadherin-catenin complex in epithelial ovarian cancer. *Journal of clinical pathology*, 59(5):460–467, May 2006.

[84] WHO. Fact sheets: Cancer. http://www.who.int/mediacentre/factsheets/fs297/en/, February 2009.

[85] Li Zhao and Peter K Vogt. Helical domain and kinase domain mutations in p110alpha of phosphatidylinositol 3-kinase induce gain of function by different mechanisms. *Proceedings of the National Academy of Sciences of the United States of America*, 105(7):2652–2657, February 2008.

[86] Jingchun Zhu, J Zachary Sanborn, Stephen Benz, Christopher Szeto, Fan Hsu, Robert M Kuhn, Donna Karolchik, John Archie, Marc E Lenburg, Laura J Esserman, W James Kent, David Haussler, and Ting Wang. The ucsc cancer genomics browser. *Nat Methods*, 6(4):239–40, Apr 2009.