A Graph-Theoretic Approach to Model Genomic Data and Identify Biological Modules Asscociated with Cancer Outcomes

Deanna Petrochilos

A dissertation presented
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

University of Washington
2013

Reading Committee:
Neil Abernethy, Chair
John Gennari,
Ali Shojaie

Program Authorized to Offer Degree:
Biomedical Informatics and Health Education

University of Washington

Abstract

Using Graph-Based Methods to Integrate and Analyze Cancer Genomic Data

Deanna Petrochilos

Chair of the Supervisory Committee:
Assistant Professor Neil Abernethy
Biomedical Informatics and Health Education

Studies of the genetic basis of complex disease present statistical and methodological challenges in the discovery of reliable and high-confidence genes that reveal biological phenomena underlying the etiology of disease or gene signatures prognostic of disease outcomes. This dissertation examines the capacity of graph-theoretical methods to model and analyze genomic information and thus facilitate using prior knowledge to create a more discrete and functionally relevant feature space. To assess the statistical and computational value of graph-based algorithms in genomic studies of cancer onset and progression, I apply a random walk graph algorithm in a weighted interaction network. I merge high-throughput co-expression and curated interaction data to search for biological modules associated with key cancer processes and evaluate significant modules in terms of both their predictive value and functional relevance. This approach identifies interactions among genes involved in proliferation, apoptosis, angiogenesis, immune evasion, metastasis, and energy metabolism pathways, and generates hypotheses for future cancer biology studies. Based on the results of this work, I conclude that graph-based approaches are powerful tools for the integration and analysis of complex molecular relationships that reveal significant coordinated activity of genomic features where previous statistical and analytical methods have been limited.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

Glossary

Betweenness Centrality: A network measure of the extent to which a node in a graph lies on the shortest path between all other nodes in the graph.

Candidate Gene: A gene prioritized with particular potential as a disease-causing gene or therapeutic target in a genomic study of many possible disease-linked genes.

Cancer Gene: In the present study, a cancer gene is one associated with cancer based on evidence in the Online Inheritance in Man database. Genes were queried for cancer terms and manually verified for involvement in cancer.

Closeness Centrality: A network measure of how close a node is to all other nodes in the graph.

Clustering Coefficient: A network measure of the density of the local connectivity of a node in a graph, calculated by the fraction of edges in triads.

Degree Centrality: A network measure of the total number or weight of the edges connected to a node. The value may reflect incoming edges, outgoing edges, or both.

Gene: A hereditary unit of DNA located at a fixed position on a chromosome. The gene may encode a protein, an RNA sequence, or regulatory sequence that controls the expression or activity of other genes.

Metabolic Interaction: An interaction between two enzymes which share a common metabolite in a metabolic pathway.

Module: A complex of genes or proteins that "interact with preferred partners weakly, transiently, or conditionally, forming a biological module serving a specific collective function" (Hartwell[147]).

MicroRNA (miRNA): An RNA sequence transcribed in the nucleus and exported to the cytoplasm that targets mRNAs from other genes to post-transcriptionally repress the expression of those mRNAs.

mRNA: Messenger RNA, a sequence resulting from the transcription of a gene that encodes a protein.

Network Community: A group of closely related and connected nodes within a graph.

Pathway: A group of molecules, including genes, proteins, metabolites and/or environmental factors that interact in a series of steps to perform a certain function in the cell. These pathways are often involved in signaling cascades, metabolic processes, or gene regulation.

Protein-Protein Interaction: The binding of two proteins to perform a biological function. These interactions are usually identified experimentally by yeast two-hybrid experiments or mass spectometry.

Random Walk (in Graphs): An algorithm that begins at a random node i in a graph and takes a random step to an adjacent node j to determine the probability that one will transverse from i to j in a walk of length t. A transition matrix $\boldsymbol{P}$ summarizes these probabilities at t=0 and $\boldsymbol{P}^t$ at time t.

Signaling Interaction: An interaction between two proteins or genes that participate in a signaling cascade where both share a reaction event.

Acknowledgements

I would like to thank the faculty and administration of the Department of Biomedical and Health Informatics for creating an inspiring academic environment and giving me the resources to learn about the field of biomedical and health informatics, to grow as a scholar, to explore, to collaborate and pursue this research. In particular, I thank Peter Tarczy-Hornoch for his leadership in our department and for his advice during my first years in the program, and I had many helpful discussions with George Demeris, Anne Turner, Brian Brown, Sandy Turner and Joan San.

Thank you to my advisor, Neil Abernethy for his dedicated and thoughtful mentorship and for his input on my work. His support has helped me succeed in my studies and improve my research. I thank my committee members, Ira Kalet, Ali Shojaie, Barbara Endicott-Popovsky and John Genarri who have been so generous with their guidance and feeback on my research. They have helped me develop my methods and my work has become much stronger because of their excellent advice.

I would like to thank Aaron Chang for his mentorship and input on microRNA analysis, Cornelia Ulrich for her support and advice on methods in epidemiology and cancer biology, and Alexander Tsiatis for his helpful input on random walks and their applications. I thank Matthew Brauer, Harris Shapiro, and George Kan for their mentorship during internships.

I would also like to acknowledge financial support from the National Library of Medicine (grant T15LM07442), the ARCS Foundation, and GO-MAP at the University of Washington that helped make this research possible.

Finally, I am so thankful for wonderful family and friends. I extend my love and gratitude to my parents and siblings who have always been there to encourage and support me, to my husband Rama, who has been there to motivate, inspire and cheer me on, and to our two munchkins, Ben and Sophia, who bring us brightness and smiles on a daily basis.

Chapter 1: Introduction

An increasingly large influx of biological data from high-throughput experimental methods is available to biologists who seek to understand the influence of genes in the etiology and progression of complex diseases including cancer, diabetes and cardiovascular disease. The availability of this data presents an unprecedented opportunity to biologists to translate this information into knowledge about pathological processes and their interventions; however it remains a challenge to extract functionally relevant genes from data sets containing tens of thousands of measurements of gene expression, proteomic, epigenetic or environmental factors that may contribute to various disorders. Typically, studies analyzing genotype-phenotype associations utilize statistical methods that search for the the most significant individual genes associated with putative outcomes. However, there is a growing interest in methods that examine how multiple genes interact in biological systems to account for the interdependent physiological processes underlying complex diseases. This has led to investigation of methods that can incorporate prior knowledge of biological systems, including evidence of genetic interactions and common pathways, with empirical approaches suited to identifying multiple predictors in large data sets. The goal of this research is to address this challenge using graph-theoretical frameworks to improve the modeling and analysis of such large-scale genomic data.

1.1: Challenges in Large Scale Genomic Studies

A primary challenge in the statistical analysis of the genetic basis of disease is the effective exploration of high-dimensional genetic predictors to isolate a small but relevant sample for further study. To estimate such effects, genetic studies may focus on an initial set of candidate genes with functional relevance or explore tens of thousands of genes for significant associations with outcomes of interest. Typically, χ-squared, *t*-statistics, or multivariate regression analysis are used to measure significant up-regulation or down-regulation of genes. However, while these methods may be able to distinguish significant statistical interactions with substantial effect sizes, they are generally not tailored to assess more subtle interactions among multiple genes and environmental factors. Genes yielding

small or modest signals or those with non-additive effects may not be detected by these methods, and such approaches do not adequately account for the genetic complexity of non-Medelian phenotypes. These limitations motivate a search for better methods to model and identify key interactions associated with dysregulated pathways. Here, I propose the use of graph theory and network models to address challenges associated with genomic analysis of complex disease.

Statistical, data integration and computational issues need to be addressed to improve current analytical approaches in the genetic analysis of massive data sets and complex diseases. Statistical issues that arise when testing high-dimensional genomic data include: 1) genes involved in complex diseases tend to only have small individual effects and univariate studies may be underpowered to detect these effects; 2) testing the statistical significance of a high order of genes corresponding to numerous hypotheses leads to multiple testing, where a fraction of the significant results will occur by chance; 3) testing a large number of genes and their interactions may overfitting estimates, and consequently results are often not generalizable or reproducible, and; 4) commonly applied statistical methods, including χ-squared, *t*-tests and regression analyses lack the power to measure the joint effects of gene interactions, where the combined effect of the genes significantly exceeds the sum of their individual effects. Data mining approaches (i.e. decision trees, random forests, SVM, and Bayesian analyses) have also been applied to more accurately model genetic predictors and their interactions[1]; however, more efficient methods are needed to increase the power to detect coordinated behavior of genes and integrate prior knowledge in the search for genetic associations, thereby narrowing the parameter space and decreasing uncertainty.

Ample evidence of biological interactions and masive amounts experimental data are available from the literature, or are curated in online databases housing expression, gene regulatory, protein interaction and other multi-scale genomic information. These data sources can be leveraged as prior knowledge in computational models of genes in the context of biological systems, including molecular pathways and condition-specific interactions. Incorporating evidence of multi-scale interactions and information regarding known disease genes focuses the search for genes that are more likely to have functional relevance and improves interpretability by presenting genes in the context of their various biological processes. Several mathematical and computational methods have been applied to link prior evidence with empirical approaches in the analysis of the molecular pathology of disease. Examples include mathematical models, Bayesian networks, Boolean networks, and Petri

Nets. These approaches are often oriented toward discrete, well-described datasets where reaction events are quantified in detail and/or annotation of interaction data is available. They are not generally applicable or efficient in the context of exploratory studies of high-throughput data or genome-wide feature selection of genetic predictors.

Graph-based analyses are widely used to model social and information networks and have been shown to be powerful analytical tools when used to study relationships in large-scale data sets. Network approaches in genome studies build on the common hypothesis that disease genes share observable patterns in biological networks. In this context, genes that are critical in terms of disease etiology also play a central role in the topology of the network and tend to interact closely with other disease genes. By providing a framework to integrate diverse biological interactions and analytical tools to study these relationships, graph-theoretical approaches provide powerful models to investigate genomic activity in complex systems. These models can be applied to address statistical, modeling and data integration issues associated with high-throughput biological data analyses in genotype-phenotype studies. To explore the use of network algorithms to improve the investigation of genes in complex disease, this dissertation implements and evaluates a graph-theoretical framework to integrate and analyze the coordinated behavior of genes in cancer onset and progression.

1.2: Research Objectives

The broad aim of this work is to implement and evaluate a graph-based platform to model interaction information, use prior evidence and analyze high-throughput data in the context of biological networks. Such an approach can be applied to mine large-scale data sets to generate hypotheses for further cancer-based research. Previous studies have shown that network characteristics and community detection can be useful in genotype-phenotype studies[2-4]. This study builds on earlier work to demonstrate the strength of graph-based analysis of biological networks to address statistical and computational issues in high-dimensional genetic analysis, by the: 1) integration and modeling of genomic evidence including interaction, experimental, and regulatory data, to narrow the feature space and to facilitate the discovery of reliable candidate genes supported by prior evidence of their biological relevance, 2) identification of biological modules associated with cancer that, in contrast to gene set enrichment approaches, can focus on relevant regions of activity in large pathways and can include genes that are span multiple pathways, 3) focus on the discovery

and interpretability of biological modules by using an graph-based random walk algorithm optimized for community finding that maximizes module scores and controls for module size.

The principle hypothesis of this dissertation is that **using a graph-theoretical approach to study large-scale genomic data, focusing on network characteristics and module generation in biological networks, provides a powerful framework for data integration and improves performance and interpretation in analyses of the coordinated behavior of genes in complex disease**. Specific objectives of this study are to assess the predictive value of network features to detect cancer genes; and to use biological networks as a framework to integrate genetic interaction and regulatory information to better understand the genetic basis of cancer. My specific objectives are listed and summarized below:

I. *Assessing the Network Characteristics of Cancer-Associated Genes in Metabolic and Signaling Networks*
II. *Using Weighted Random Walks to Identify Cancer-Associated Modules in Expression Data*
III. *Analysis of microRNA Data in Random Walk-Generated Gene Expression Modules*

1.2.1: Assessing the Network Characteristics of Cancer-Associated Genes in Metabolic and Signaling Networks

Graph-theoretic methods have been broadly applied to study properties of interactions in metabolic, regulatory, and signal transduction pathways. The first objective is to study the relationship between network features and cancer genes within a biological network and to measure the power of network characteristics to predict cancer genes. The hypothesis associated with this objective is that **cancer genes demonstrate higher centrality and tend to interact closely and cluster with other cancer genes in the network.** Using a generalized linear model, I evaluate the predictive power of centrality measures and clustering coefficients associated with cancer genes in the interactome. Further, I assess the modularity of cancer genes by applying a random-walk algorithm to discover communities enriched with cancer-associated genes. The results show that cancer genes in metabolic and signaling networks exhibit significant topological differences in terms of degree, clustering coefficient, and modularity; and these features demonstrate greater predictive ability in signaling networks. These findings give an empirical basis for the use of algorithms

employing similar network-based measures to prioritize disease genes or to predict disease states[5].

To address this objective, I develop custom parser in R to extract pairwise relation and reaction interactions from the Kyoto Encyclopedia of Genes and Genomes (KEGG)[6] used to create global signaling and metabolic networks. Network features are calculated for individual nodes (genes) in the metabolic and signaling networks, including: degree, closeness and betweenness centrality and clustering coefficient. The ability of network characteristics of individual genes to predict cancer status is assessed using linear regression and a gold standard list derived from OMIM entries. To evaluate the cohesiveness of cancer genes in the network, I apply a random walk algorithm to discover dense cancer-associated modules within an integrated network of protein-protein, metabolic, signaling and regulatory interactions. Modules significantly enriched with cancer genes are identified by comparison against a hypergeometric distribution. The biological relevance of modules discovered in the community search is evaluated based on evidence in the literature and comparison with curated pathways.

1.2.2: Using Weighted Random Walks to Identify Cancer-Associated Modules in Expression Data

The etiology of cancer involves a complex series of genetic and environmental influences. The second objective is to better represent and study the intricate genetics of carcinogenesis by applying a weighted random walk and modularity-driven clustering algorithm to search for modules of interacting genes that are significantly associated with cancer onset and progression. The hypothesis associated with this objective is that **the graph-based random-walk and community finding algorithm can be used to integrate prior evidence, to model interaction data and to yield interpretable, biologically-relevant modules.** A network of biological interactions is constructed to search for groups of genes composing cancer-associated modules. I implement *Walktrap*[7], a random-walk-based community detection algorithm to identify significant modules in the weighted interactome that predispose to tumor development in hepatocellular carcinoma (HCC), adenoma development in colorectal cancer (CCA), and prognosis in breast cancer (BC). Results are compared with those generated by several other recent tools developed to discover cancer-related disease modules in gene interaction networks. The findings show that significant modules include interactions among transcription factors (*SPIB*, *RPS6KA2* and *RPS6KA6*) and cell-cycle regulatory genes (*BRSK1*, *WEE1* and *CDC25C*), inflammation and

proliferation (*SOCS2, IL20RA and CBLC*) and growth factors (*IRS2, FGF7*) that are highly connected with known cancer genes, are functionally related to cancer, and show potential value as therapeutic targets.

Specifically, to address this objective, an interactome is created from metabolic, signaling and protein-protein interactions derived from KEGG and HPRD. I downloaded three cancer expression datasets from GEO, a study of hepatocellular carcinoma by Roessler[8] (GSE14520), a breast cancer prognosis study by Desmedt[9] (GSE7390), and colorectal cancer data from Sabates-Bellver[10] (GSE8671). Corresponding fold change data are transformed to create a vector of edge-weights for the interaction network. The *Walktrap* algorithm is used to calculate distances between nodes used to cluster communities in the network and identify dense modules associated with cancer. The community detection algorithm is an agglomerative clustering function; the merge stop is assessed using multiple criteria, including a threshold to limit module size (200), and maximization of module score and of modularity. Modules are scored based on cumulative expression values and are assessed for significance by comparing these scores against estimates from a random distribution.

1.2.3: Evaluation of the Use of Weighted Random Walks and Expression Data to Identify Cancer-Associated Modules

Results from section 1.2.2 are evaluated to measure the efficiency and performance of the random walk and module search. I assess biological relevance of significant modules using functional annotation derived from ConsensusPathDB and supporting evidence from the literature. To further evaluate relevant functionally enriched pathways from the top scoring modules, these annotations are compared against significant pathways identified by Gene Set Enrichment Analysis (GSEA). Next, I evaluate performance of the top scoring modules against modules generated by other tools used to detect significant modules in weighted interaction networks, *jActiveModules* and *Matisse.* An OMIM-derived list of cancer genes is used as a gold standard to assess detection and significant enrichment of cancer genes in these modules. The results show that the *Walktrap* algorithm performs well in comparison to related tools and can identify modules significantly enriched with cancer genes, their joint effects and promising candidate genes. Smaller overall module size allows for more specific functional annotation and facilitates the interpretation of significant modules

1.2.4:Analysis of microRNA Data in Random Walk-Generated Expression Modules

The final objective is to leverage the ability of a molecular interaction network to integrate interaction data, mRNA expression, and microRNA (miRNA) expression data. The hypothesis here is that **modules enriched with targets associated with miRNA hits will identify high-confidence candidate cancer genes based on correlation between mRNA and miRNA expression data**. I aim to identify miRNA-mRNA pairs involved in cancer onset and progression by using *Walktrap* to discover modules predisposing to cancer and enriched with miRNA-mRNA activity in expression data. Several methods to integrate and score miRNA data are evaluated.

To carry out this analysis, the initial interactome for this study was created from from metabolic, signaling and protein-protein interactions derived from KEGG and HPRD. I downloaded two cancer expression datasets from GEO that include miRNA-mRNA correlation data, a study of hepatocellular carcinoma from Burchard[11] (GSE22058), and a study of breast cancer survival by Buffa[12] (GSE22220). I examined methods to integrate correlated miRNA-mRNA pais into the network analysis using several matching and weighting approaches. Matching methods assessed include: optimal matching, retaining the top three or five matches, and including all matches. Corresponding mRNA-miRNA fold change data are used to create a vector of edge-weights for the interaction network applying weights to edges using fold change of the adjacent nodes, or a linear transformation of fold change values. I implement the *Walktrap* algorithm to calculate distances between nodes and to to identify modules associated with cancer. Significant over-representation of miRNAs in these modules is evaluated by enrichment analysis using a hypergeometric distribution. The results highlight miRNAs including *miR-22, miR-151, miR-93* and *miR-33b* and targets *LIFR, CYPA4A11, SH3GL2* and *MYBL2* for their potential role in cancer.

1.2.5:Evaluation of Analyzing miRNA Data in Random Walk-Generated Expression Modules

Here I evaluate integrating miRNAs in a weighted interaction network to reliably identify cancer-associated modules that are enriched with miRNAs and mRNA targets. The value of using miRNAs and their target correlations is compared to approaches that 1) use only mRNA data, and 2) use both miRNA data and mRNA data in the absence of a network

model. I review the functional annotation of the modules, and evidence from the literature. The following steps summarize the methodology used to address this objective. To evaluate matching and weighting strategies used to integrate miRNAs in the network, I calculate cancer-gene enrichment scores of significant modules found by using these methods. The performance of each approach is assessed by measuring Precision, Recall and Matthews Correlation Coefficient (a function of overall Precision and Recall). I review functional annotation of significant modules using ConsensusPathDB, and compare my findings with original studies from which the data was attained, the HCC study by Burchard 2010[11], and the BC study by Buffa 2011[12]. Results are also validated using supported evidence in the literature for top mRNA and miRNAs associated with these outcomes. Finally, to determine the effect of using miRNA data, I compare significant modules discovered using miRNA-mRNA findings against those found using mRNA-only approach on comparable HCC and BC data.

1.3: Contributions

This research contributes to the fields of biomedical informatics, genomics, and cancer biology. Overall contributions include: the implementation and evaluation of methods for high-dimensional data analysis, applications of graph-based and module-discovery algorithms in biology, approaches to data integration in biology, and a set of hypothesis for further cancer-based gene interaction studies. These contributions are relevant to the fields of biomedical informatics and cancer biology, and generally applicable to large-scale data analysis integrating diverse information to identify significant interactions among entities and their association with outcomes of interest. The contributions of this dissertation are summarized below.

- Development of a graph-theoretical approach to study the coordinated behavior of genes in complex disease. This method improves upon standard statistical analyses of high-throughput genomic data by focusing on the discovery of significant gene interactions rather than single candidate genes, and by using prior evidence of biological interactions to narrow the feature space.

- Expansion beyond predefined gene sets to allow investigation of gene interactions within a region of a pathway or that overlap across pathways. This research leverages

known pathway information, yet uses these definitions loosely as it allows for identification of modules that cross *a priori*-defined pathways and gene sets.

- Demonstration of a basis for the predictive value of centrality features and community finding in a biological network seeded with cancer genes.

- Demonstration of the ability of the network to integrate prior information in the investigation of the genomic basis of cancer. The graph-based framework serves as platform for the integration and computational analysis of prior evidence of biological interactions and regulatory information in the study of the genetic basis of complex disease.

- Creation of a gold standard list of cancer genes used in evaluation of modules. Each gene in the network is queried for cancer-associated terms and each match is manually verified. These methods improve upon previous approaches to summarize cancer gene data, based on the specificity and coverage of queries and manual verification.

- Contribution to research applying graph-based frameworks in genome analyses and module-finding by using a random walk algorithm optimized to discover communities in large biological network. Most previous work using random walks has focused on gene prioritization rather than module discovery and this approach shows strong performance when compared to similar tools and results in smaller, more interpretable modules.

- Presentation of a scoring metric to score module significance and enrichment using a bootstrap distribution. Where many previous studies have used correlation or p-values to apply weights to network eges, here, fold change values are used to build more robust weight distributions, to focus the study on outcomes rather than the strength of correlation between genes, and to increase sensitivity to differential expression measurements.

- Hypothesis generation to guide future studies investigating candidate genes and their interactions in cancer studies. This research identifies potential therapeutic targets that are implicated in breast cancer, hepatocellular carcinoma and colorectal cancer.

- A generalizable example of how to integrate diverse information and find communities of closely related entities to guide other applications of graph-based research.

1.4: Dissertation Overview

This dissertation is organized as follows. Chapter 1 describes the research challenges inherent in the study of the genetic basis of complex disease and presents a summary of research objectives, methods and contributions. Chapter 2 reviews background work in network and graph analysis in genomic studies and sets the context for the present work in the context of these studies.

Chapter 3 is an assessment of the power of using network characteristics as predictive features in studying metabolic and signaling networks. This chapter examines the hypothesis that cancer genes demonstrate higher centrality measures and act as hub genes in the network. It also presents an evaluation of the cohesiveness of cancer-labeled genes and their affinity to form modules with other cancer genes and cluster in network communities. Chapter 4 describes implementation of a random walk algorithm optimized for community-finding and weighted with mRNA expression values, to search for communities of interacting genes that are significantly associated with cancer onset and progression. Findings from this analysis are evaluated by reviewing functional annotation, evidence from the literature and comparing performance to similar tools. Chapter 5 presents an expanded assessment of the ability of the molecular interaction network to integrate diverse types of interaction data, mRNA expression, and miRNA expression data. The chapter examines the hypothesis that modules enriched with targets associated with miRNA hits will consist of more reliable candidate genes based on the correlation between two significant data sources. Several methods are evaluated to integrate and analyze miRNA in the network and for their ability to improve the search for modules associated with cancer.

Finally, Chapter 6 concludes by reviewing the methods, results, contributions, conclusions and avenues for future work based on this research.

Chapter 2: Network Biology and the Cancer Genome

2.1: Introduction

Network models serve both as a framework to represent and visualize complex interactions, and as a platform for further analysis, compatible with extensive libraries of graph algorithms. A network consists of entities and their relationships, the entities are nodes in the graph, and edges describe the relationships between nodes. Graph-based models have been widely used to describe and investigate relationships and information exchange in social and information networks; from the transmission of disease in communities[13], to the flow of data in telecommunication networks[14,15], and the search for relevant links in the Google search algorithm[16]. Graph-based approaches differ from conventional statistical analyses by shifting the focus on observations and their associations with the outcome, to relationships between observations and their attributes under specific conditions.

Graph-theoretical concepts of node centrality, the shortest path between nodes, and clustering, have been applied to study gene interactoins to annotate genes, discover their functional or putative relevance, or study their joint influence in disease outcomes[17]. As putative processes underlying complex disease result from multiple genetic or environmental factors acting in concert, network analysis is well-suited to the study of the causes of these complex phenotypes. In genomic applications, such analyses have been applied to model multi-scale interactions including protein-protein interactions (PPI), metabolic, and signaling networks, as well as expression, sequencing, copy number, and genomic mapping data[18].

2.2: Overview of Biological Pathways and Interaction Networks

The following is a summary of signaling, regulatory, metabolic, and PPI networks as presented by Junker and Schrieber[19] and Pavlopoulos[20]. *Signal transduction* networks model cascades of directed, reversible events that begin via actions of molecules outside of the cell. These molecules, including hormones, cytokines, and growth factors, bind to the surface of the cell and initiate a cascade of signal transduction events (reactions) that ultimately target

transcription factors or metabolic enzymes within the cell. The cascade of reactions involves modification of a downstream molecule by a preceding molecule, often via phosphorylation or ubiquitination. These reactions may be mediated by protein-protein interactions, however many reactions may involve other chemical factors including steroid hormones, second messengers such as cAMP, or environmental factors such as calcium stress, UV, or irradiation. These cascades can regulate cell functions such as glucose uptake, cell growth, sensory signals and regulation of gene expression. Regulatory networks, for example, are specific types of signal transduction networks that involve binding of transcription factors to their target genes to modulate gene expression. An example of a signaling network involved in regulation of cellular growth and differentiation, the *TGF-β* signaling pathway, is presented in Figure 1. Gathering signal transduction data is challenging and costly, and the most of this information is derived from yeast experiments or from orthologous data collected across species.

Metabolic networks model biochemical reactions or interconversions (usually catalyzed by proteins termed enzymes) between metabolites. Metabolites can be small molecules such as glucose or amino acids or macromolecules such as polysaccharides and glycans. Metabolic networks consist of directed, irreversible reactions, and are bipartite networks, wherein one class of nodes represents an enzyme and another class represents a metabolite. Metabolic interactions generally include relationships between enzymes that share a common metabolite. Data are primarily derived from laboratory experiments modeling specific chemical reactions and these networks are often described in great detail using mathematical models of the underlying reactions. An example of a metabolic pathway, Prostaglandin and Leukotriene Metabolism, is presented in Figure 2.

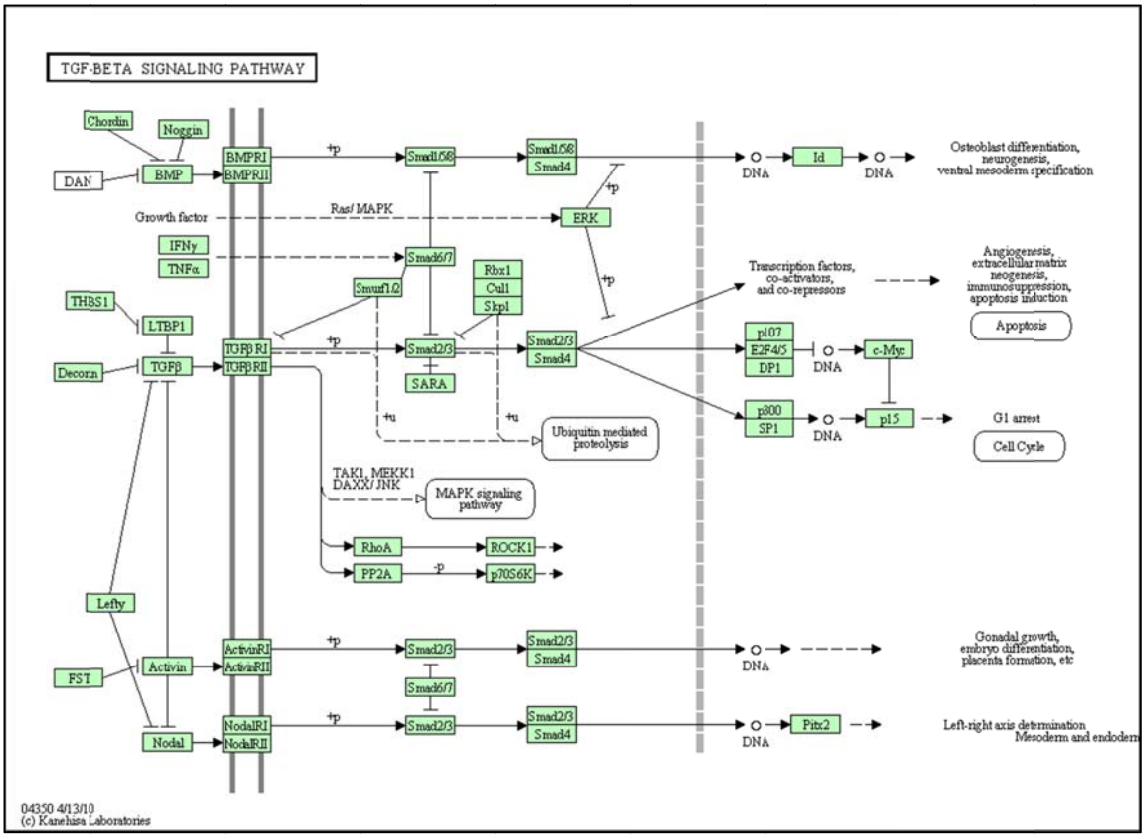


Figure 1: The TGFβ Signaling Pathway. The TGF-β pathway is described in detail by Alberts[21] and Weinberg,[22]. In the KEGG[6] diagram, arrows refer to interactions where the preceding molecule is a precursor to and promotes the formation of the succeeding molecule. Lines with a crossed end refer to interactions where the preceding enzyme inhibits the succeeding enzyme. Transforming Growth Factor-β (TGF-β) is a protein that initiates a signaling pathway that results in downstream activity including angiogenesis, extracellular matrix development, immune regulation, tissue repair, apoptosis, and cell-cycle arrest; and these pathways play an important role in cancer onset and progression. The diagram shows TGF-β can be directly inhibited by proteins LTBP1, Decorin, or Lefty, or Activin. When TGF-β reaches the cell surface, TGF-β type I (TGF-β RI) and type II (TGF-β RII) trans-membrane receptors create a dimer of two kinases where TGF-β RII can phosphorylate TGF-β RI. Smad Ubiquitin Regulating Factors 1 and 2 (Smurf1/2) can inhibit formation of the TGF-β RI / TGF-β RI II dimer by phosphorylation of TGF-β RI. If TGF-β RII successfully phosphorylates TGF-β RI I, TGF-β RI binds and phosphorylates Smad2/3. SARA, an achoring protein can promote this phosphorylation by recruiting Smad2/3 to activate TGF-β RI. Smad2/3 then creates an oligomer with Smad 4. Rbx1, Cul1 and SKP1 may compete to bind with Smad2/3, and thereby disrupt the formation of the complex. The Smad2/3 and Smad4 oligomer translocates to the nucleus to activate various genes, conditional on cell type and state, including P107, EdF/5, DP1, p300, and SP1. Activation or degradation of these proteins (ie. by ubiquitin-mediated proteolysis), can block expression of c-myc oncogene or increase expression of cell-cycle control genes p15 or p21. TGF-β with MAPK signaling pathways also interact via TAK1, MEKK1, RHOA, and DAXX/JNK to dephosphorylate PP2A and signal p70S6K, initiating cell proliferation.

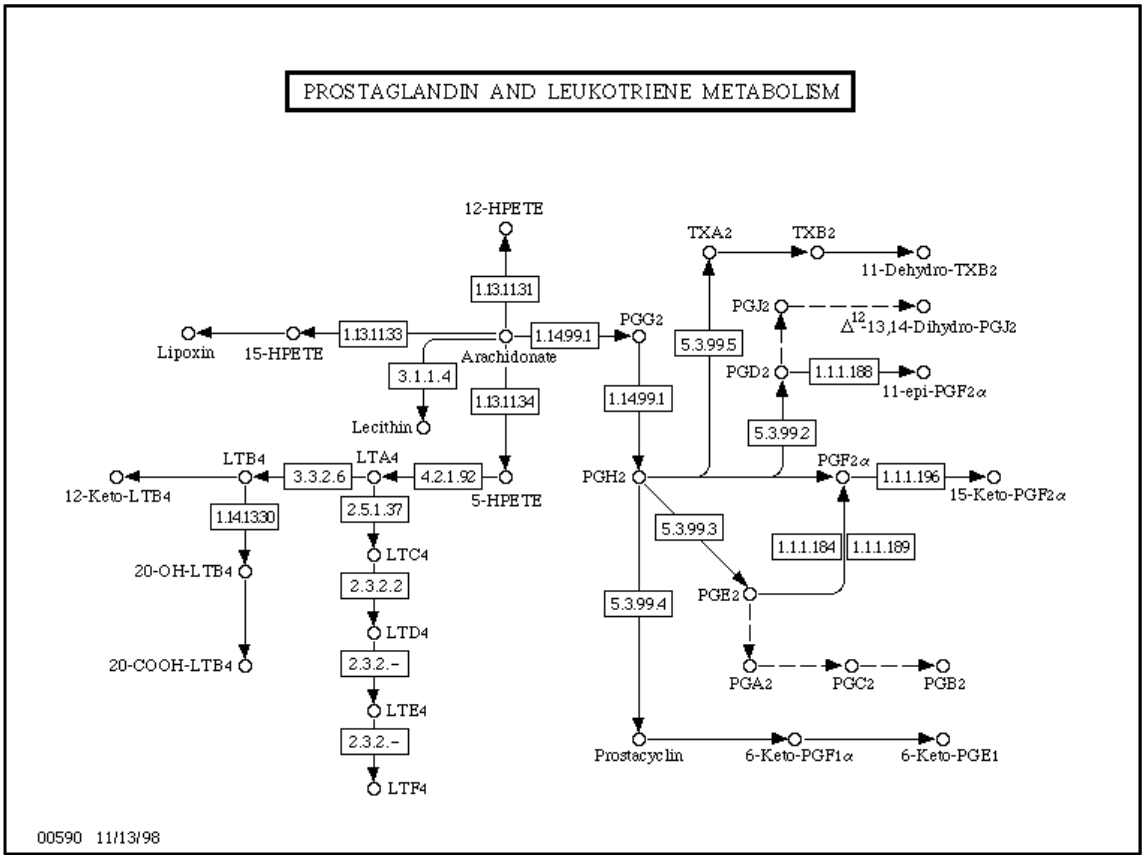


Figure 2: Prostaglandin and Leukotriene Metabolism. The Arachidonate metabolism pathway is described in detail by Alberts[21] and Weinberg[22]. In this KEGG[6] diagram, rachidonate is a polyunsaturated fatty acid that can be transformed into prostaglandins or leukotriene. For illustrative purposes, prostaglandin synthesis is described in specific detail. Prostaglandins are important regulators of immune response and inflammation. The COX1/2 enzyme, prostaglandin endoperoxidase synthase (also known as prostaglandin endoperoxide synthetase-PES or prostaglandin synthase - PGH1/2) is a fatty acid cyclooxygenase that adds two oxygen molecules to arachidonate and interacts with arachidonate in the initial steps of the prostaglandin synthesis pathway. It converts an arachidonate radical via cyclogenase activity to form prostaglandin-G2 (PGG_2). PGG_2 then reacts again with arachidonic acid and is reduced by peroxidase activity to prostaglandin-H2 (PGH_2). Depending on cell type, PGH_2 may react with other enzymes to convert arachidonic acid to prostaglandins or thromboxane. Thromboxane-A-synthetase converts the enzyme PGH_2 to enzyme thromboxane A2 (TXA_2), typically in blood platelets, and TXA_2 spontaneously hydrolyzes to produce thromboxane B2 (TXB_2) which is involved in the contraction of vascular smooth muscle and induces platelet aggregation and selectin release. TXB_2 is finally metabolized to 11-Dehydro TXB_2. In vascular endothelial cells or smooth muscle cells, PGH_2 may react with the metabolite prostaglandin-1 synthetase to convert PGH_2 to prostaglandin I (PGI_2 or Prostacyclin), whose activity can induce coronary vasodilation, bronchoconstriction, inhibit aggregation of platelets, and decrease norepinephrine release. PGI_2 spontaneously hydrolizes to form urinary metabolite 6-Keto-PGJ_{1a}, and perfuses in liver to compound 6-Keto-PGE_1. In serum albumin, PGH_2 reacts with prostaglandin-D-synthetase to produce prostaglandin-D2 (PGD_2), prostaglandin-F-synthetase may then reduce PGD_2 and PGH2 to 11-epiPGF_{2a}, which inhibits platelet aggregation. Alternatively, PGD_2 may undergo non-enzymatic dehydration to produce metabolite PGJ_2 which then non-enzymatically becomes metabolites delta12-PGJ_2 and 15-Deoy-delta 12,14-PGJ_2. In macrophages, or the kidney, prostaglandin-E-synthase converts PGH_2 to Prostaglandin-E2, which can interact with enzymes carbonyl reductase (NADPH, or adelhyde reductase) and prostaglandin-E2-9-reductase to become PGF_{2a} or convert non-enzymatically to prostaglandin-A2, prostaglandin-C2 or prostaglandin-B2. PGH_2 may converts to PGE_2 via interaction of PGE-synthase and spontaneously hydrolyzes with PGF_{2a} and enzyme 15-hydroxyprostaglandin-D dehydrogenase (NADP+) to become 15-Keto-PGF2a.

PPI networks are undirected networks that model interactions between protein molecules and may overlap with signaling and metabolic networks. Proteins can interact transiently, as in signal transduction networks, can be modified by enzymes, or can act as scaffolds, sequestering interactions to specific locations within the cell. Forces that mediate such interactions include hydrophobic effects, van der Waals interactions, electrostatic interactions, and hydrogen bonds. Most but not all protein-protein interactions are followed by chemical reactions. PPI network data are typically collected from microarray or yeast two-hybrid (Y2H) experiments.

Signaling, PPI, and metabolic networks share several common characteristics. One is that these networks exhibit a small-world property where most nodes are not well-connected with other network nodes, but each node is reachable from any other node in relatively few steps[23]. These networks have been shown to be scale-free where more high-degree nodes (hubs) exist than would be expected by chance, and they follow a power-law rule whereby hubs connect dense communities of low-degree nodes[24] . Direct links between high degree proteins are suppressed whereas links between high degree and low degree proteins are favored. These networks are robust and complex, they are vulnerable to specific attacks against highly centralized and connected hub nodes; but random attacks are unlikely to disturb the network, as most nodes are not critical hubs[25]. Low-degree vertices may play important roles in maintaining network integrity and hubs may facilitate brokering communication between distant parts of the network.

Generally, signal transduction networks exhibit more specific cellular activities than PPI networks and metabolic networks are more stable and well described than signaling or PPI networks. Signaling and metabolic pathways have been described in quantitative detail in model organisms and in humans[26,27]. Both signaling and metabolic networks are *anisotropic*; they have specific inputs and outputs, compared to PPI's which are *isotopic* and lack specific inputs and outputs. Given that complete information is lacking with signaling interactions, and since these interactions show considerable overlap with the more discrete metabolic and PPI networks, relatively little network analysis work has been done with signaling networks versus metabolic and PPI networks[19].

Functional networks typically merge PPI data with a wide variety of genomic information, for example: known disease genes, co-expression data, proteomic data, regulatory interactions, genetic variation and genome mapping data. These data are extracted from curated data sources or experimentally derived from original studies. Relationships between genes or protiens are indicated by network edges, while discrete, categorical or

continuous variable measurements can be incorporated by annotation and/or weights applied to graph nodes or edges. Annotation of edge variables often includes binary relationships or correlation values; and node annotations may include node classes, or quantitative or qualitative descriptors of experimental measurements or significance values.

Along with massive generation of sequencing and expression data, extensive functional, interaction, and pathway information are available from the literature and online resources to augment and annotate biological networks. The Gene Ontology[28] is a widely-used resource for retrieval of functional gene annotations. Interaction data are available from a growing number of databases that may focus on metabolic, signaling or protein interactions, these include: BIND[29], Reactome[30] , HPRD[31], STRING[32], INTACT[33], and the National Cancer Institute's Pathway Interaction Database [34]. Prominent databases containing detailed information on metabolic, transcription and signaling pathways include KEGG [35], BioCarta [36] and BioCyc [37]. Phenotypic information is available from OMIM[38] and MiMiner[39]. Online datasources with compilations of epigenetic and gene regulatory information include data on: predicted transcription factor targets (TRANSFAC[40]), signal-transduction (TRANSPATH[40]), predicted miRNA-targets (TargetScan[41], miRBase[42]), DNA methylation profiles (MethDB[43], MehtylomeDB[44]), phosphorylation sites (Phosida[45], NetPhorest[46]), ChIP-chip/Chip-Seq (Uniprobe[47], JASPAR[48]), and B-Cell signaling networks (HBCI[49]). Information from most of these databases and ontologies can be exported in computable forms such as SBML[50], BioPAX[51] or other XML formats, for further analysis; and a correspondingly wide array of tools has been implemented (for example, in Cytoscape [52] and Bioconductor[53]), to visualize and analyze information from external sites. Pathway, protein complex, and GO data have also been compiled in several online resources and are to generate integrated networks or provide functional annotation for gene lists (DAVID[54], ConsensusPathDB[55], PINA[56], ToppGene[57]).

2.3: Network Features and Definitions

The following are basic concepts of network theory summarized from work by Newman[58] and Pavlopoulos[20]. A network is defined as a pair of edges and vertices $G=(V,E)$, where edges are connections between the vertex nodes V, and $E(i, j)$ refers to an edge between vertices i and j. The network may be undirected, or directed, in the latter case

$G=(V,E,f)$, where f refers to a function that assigns directionality to an edge. Edges may be associated with weights where a weight $w(i,j)$ refers to the weight of the edge between vertices i and j. For undirected graphs, the adjacency matrix representation for graph G consists of an nxn matrix A such that $a_{ij} = 1$ if (i, j) is a subset of V or $a_{ij} = 0$ otherwise. In weighted graphs $A_{ij} = w_{ij}$ if $(i,j) \in V$ or A_{ij} otherwise. The adjacency list corresponding to graph G is an array of $|E|$ of pairwise elements that includes an entry for each edge in the network.

Centrality is the number of links leading in or out of a node; it reflects how well-connected the node is among other nodes in the network. The degree centrality $C_D(v)$ for a single vertex $d(v)$ in graph G with n nodes is[59]:

$$C_D(v) = \frac{d(v)}{n-1} \tag{1}$$

Specific in-degree and out-degree values are also calculated, estimating the degree of incoming edges and out-going edges, respectively.

Betweenness measures the extent to which a node lies on the shortest path between other nodes in the network. The equation divides the number of shortest paths from s to t that travel through vertex v divided by the total number of shortest paths σ from s to t[60]:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2}$$

Closeness is the mean geodesic within a network, where a lower closeness centrality corresponds to a shorter mean network distance between nodes. Within graph G, closeness measures the mean shortest path distance d from v to all other reachable nodes[59]:

$$\frac{\sum_{t \in \frac{V}{v}} d_G(v,t)}{n-1} \tag{3}$$

Finally, the local **clustering coefficient**, C_i, is a measure of how densely connected a node is within its subnetwork, and is estimated by comparing the fraction of nodes in triads to the total number of nodes[61]:.

$$C_i = \frac{2e_i}{d_i(d_i - 1)} \tag{4}$$

This value calculated by taking the actual number of edges e_i of triangles in a subgraph G_i and dividing this figure by the total number of possible edges where d_i = degree of node i. The clustering coefficient has been extended by Barrat et al.[62] to account for the topology of weighted networks.

2.4: Graph and Pathway-Based Approaches Using Prior Evidence in Genome Studies

As evidence supporting gene and other biological interactions is readily available in the literature and from online repositories, it is essential to find methods to link this information with quantitative analytical tools in studies of the genetic basis of disease- to reduce the feature space and provide a functional basis for interpretation of these analyses. Important aspects of such tools include the ability to model prior information (i.e., of pathways and interactions), and provide quantitative approaches to assess the significance of interactions with phenotypic variables. Mathematical, logical and probabilistic network models have been used for in-depth studies of reaction events of genes and proteins in specific, well-described pathways. However, large-scale data analysis demands more scalable methods to explore the influence of genome-wide interactions. Graph-based analyses provide a foundation for such methods with frameworks to model biological data and study gene interactions. These approaches typically search for significant activity disease genes in networks, analyze network properties of disease genes, search for candidate genes in the vicinity of known disease genes, or search for communities of densely connected interactions.

Mathematical, logical and probabilistic network models of biological networks have been explored to build, represent and analyze biological pathways[63,64]. Mathematical models describe reaction events in pathways using ordinary differential equations and have been used to quantify the activity of specific pathways under experimental conditions, from prototypic models of respiratory metabolism in eukaryotes[65] to the construction of synthetic models of

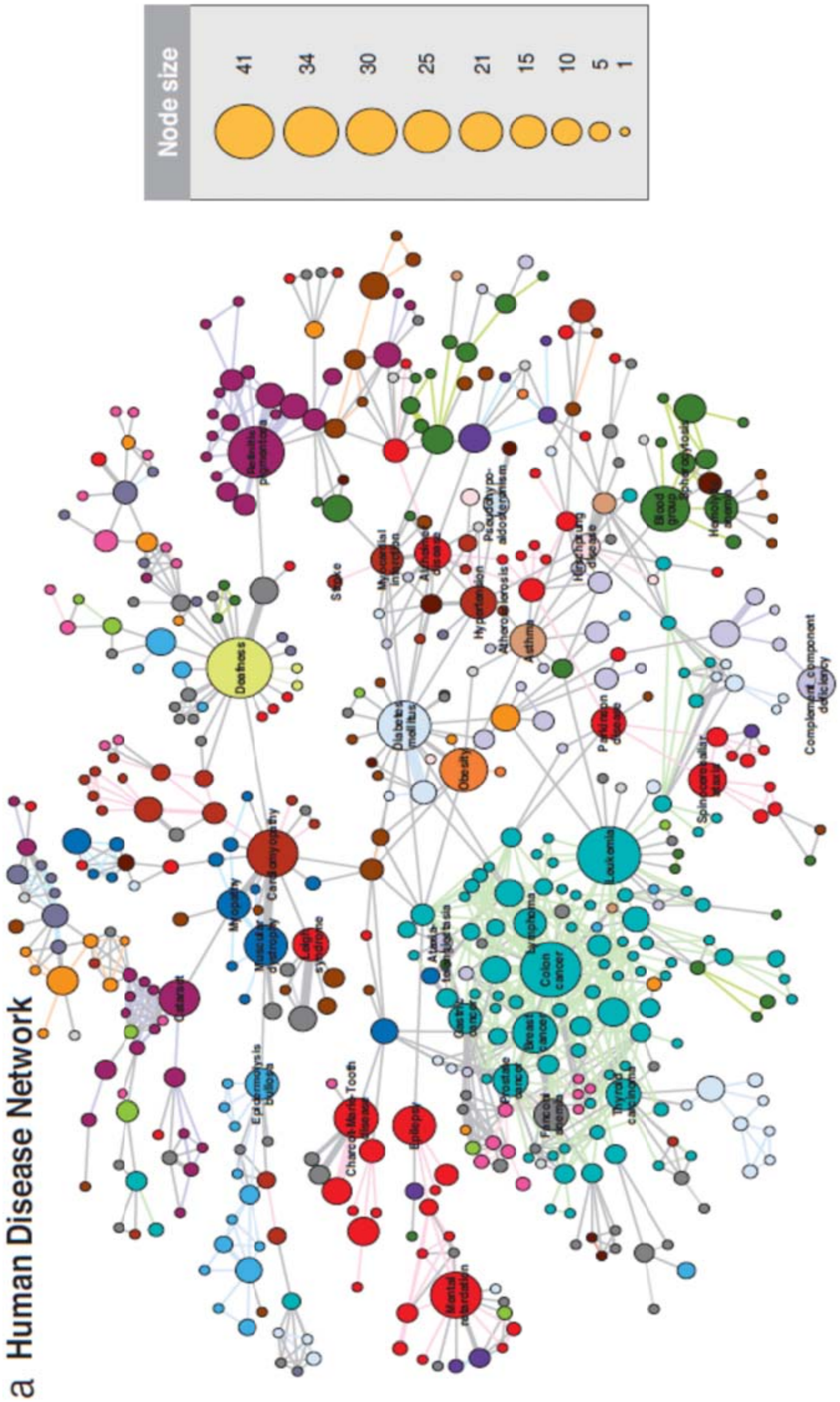
human metabolic and signaling pathways[26]. These models are characterized by high precision, but are limited to well-described biological pathways for which detailed kinetic information or reaction annotation is availalbe; and as these models are deterministic, they do not allow for the uncertainty that can be accounted for in logical and probabilistic networks[63,66]. Bayesian networks[67-69], dynamic Bayesian networks[70] Markov models, and Boolean networks[71,72] have been used to reconstruct biological pathways from gene expression data, protein interaction data and the literature, and have been used in sensitivity analysis to isolate critical genes in a network[71,72]. Petri nets originally developed to test concurrency in computer networks[73,74], have been used to model biological networks as concurrent processes. For small experiments, the use of logical or probabilistic models is a powerful approach; however, for large data sets with tens of thousands of interactions and greater uncertainty in terms of the underlying data, the use of mathematical or computational models may not be scalable or pragmatic[63].

Gene set analysis approaches leverage pathway and interaction information to determine enrichment of differentially expressed genes in gene sets representing canonical pathways, protein complexes, functional GO categories or network modules[75-78]. Gene Set Expression Analysis (GSEA)[79] uses predefined gene sets to investigate expression data to find significantly enriched sets of genes. Mootha [80] and Subramanian[79] identify enriched pathways in gene expression data by ranking significantly up-regulated and down-regulated genes, labeling the results with GO annotations and assessing the statistical significance of enriched functional categories. Later developments include using KEGG to augment functional categories [81], including topological features to define interactions[82-84], considering correlation and connectivity among genes[85,86] and covering data associated with multiple outcomes[76]. Further, a number of studies, including work by Keller[87], Efroni[66], and Ben-Hamo[88] implement algorithms to discover significantly dysregulated subpathways of curated gene sets in expression data. However, these approaches are limited in that they may not detect enrichment of specific regions of large pathways and they do not search for enriched genes that interact across multiple pathways.

Network-based studies shift from searching for enrichment of expression profiles in curated pathways and gene-sets to exploring network features of disease genes in "interactomes" comprised of integrated functional, pathway or interaction information. Network-theory concepts describing node centrality and measuring distances between nodes have been applied to interaction networks to better understand disease. Several studies conclude that disease genes tend to act as hubs in interaction networks while their

intermediate nodes add to the robustness of the network, and centrality and distance measures can help identify critical disease genes[89,90,91]. Guimera et al.[92] classify interactions in a biological network and identify intra- and inter-modular motifs associated with disease. They find that disease-related genes share similar link properties and emphasize that the most informative aspects of a network are not global, but rather local properties. Hovrath et al. [93] apply graph theory to extract geometric characteristics related to microarray coexpression data and use intra-modular network concepts and eigengene statistics to infer gene significance. They derive measures of network adjacency, density, hub genes and connectivity and translate these to geometric interpretations to build hypotheses about gene significance. Their findings suggest modular membership of putative genes and identify disease genes corresponding to hubs in co-expression networks.

A number of studies focus specifically on cancer genes and their topological characteristics in interaction networks. Jonsson et al.[94,95] map known cancer genes in an orthologous PPI network and perform cluster analysis to show that cancer genes act as hubs in metastatic subnetworks. They conclude that cancer genes are more well-connected, belong to communities with a higher degree of connectivity and are more likely to lie at community interfaces, or act as global central cores. Sun[96] and Cai[97] draw similar conclusions but add that there is an inverse relationship between cancer gene labels and clustering coefficient. Rahmani[98] and Wang[99] use topological properties to predict cancer genes in protein interaction (PPI) networks and find that these methods help identify candidate disease genes and gene signatures. Further, using expression data mapped to a PPI network, Wachi et al.[100] show that lung cancer genes have a higher extent of connectivity than do normal genes. These studies support the theory that essential genes and cancer genes are more likely to act as highly connected hubs in biological networks; although, in general, genes related to other complex diseases do not tend to exhibit high centrality[97,101].


a Human Disease Network
Node size
41
34
30
25
21
15
10
5
1
Deafness
Cardiomyopathy
Diabetes mellitus
Obesity
Leukemia
Colon cancer
Breast cancer
Lymphoma
Epilepsy
Mental retardation
Cataract
Muscular dystrophy
Myopathy
Thyroid carcinoma
Asthma
Hypertension
Parkinson disease
Complement_component deficiency
Blood group
Spherocytosis
Retinitis pigmentosa
Charcot-Marie-Tooth disease
Epidermolysis bullosa
Stroke
Myocardial infarction
Alzheimer disease
Atherosclerosis
Hirschsprung disease
Spinocerebellar ataxia
Gastric cancer
Prostate cancer
Fanconi anemia
Ataxia-telangiectasia

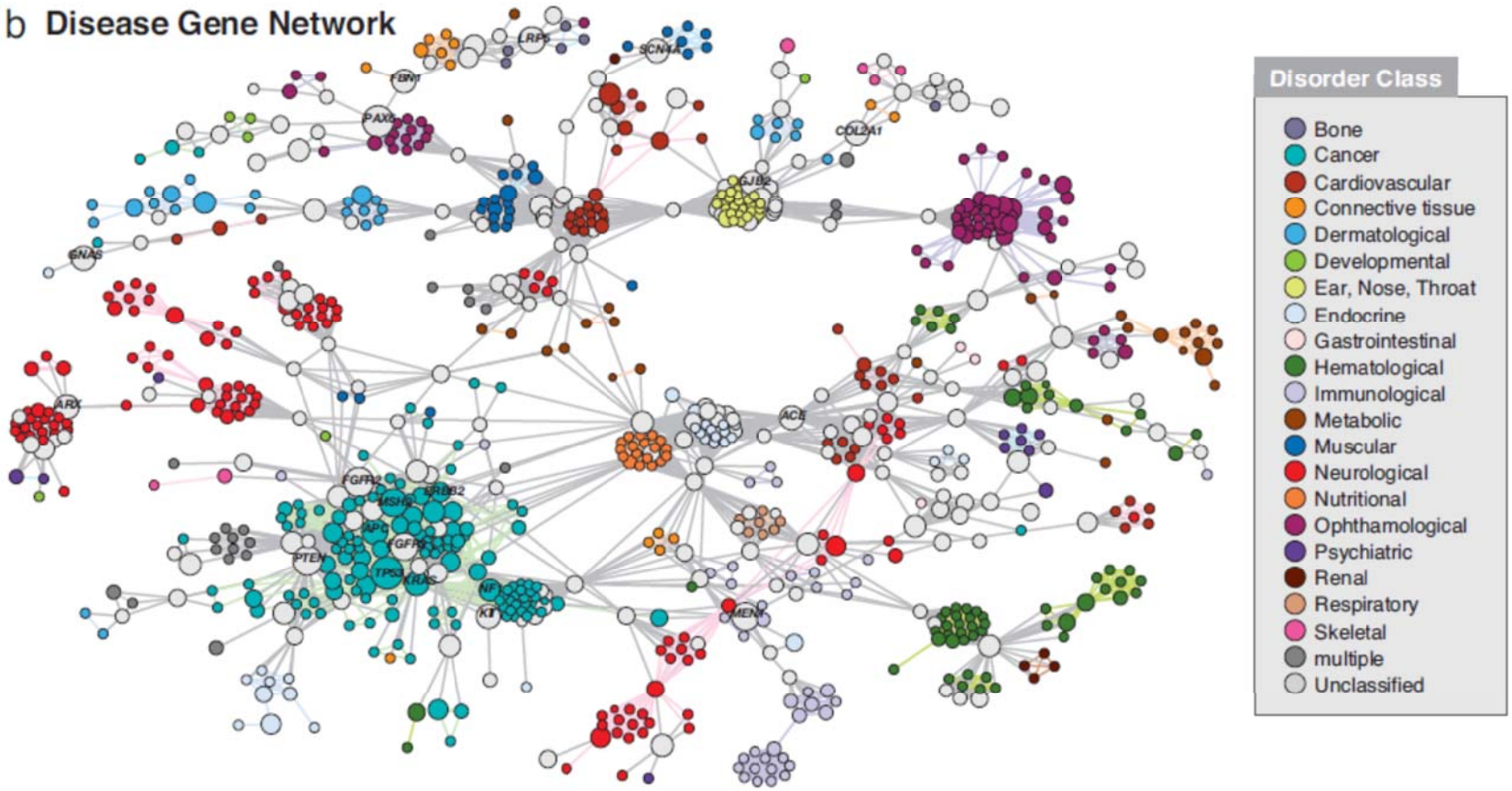


Figure 3: The HDN and the DGN constructed by Goh. From GOh et al. [101] (a) In the HDN, each node corresponds to a distinct disorder, colored based on the disorder class to which it belongs, the name of the 22 disorder classes being shown on the right. A link between disorders in the same disorder class is colored with the corresponding dimmer color and links connecting different disorder classes are gray. The size of each node is proportional to the number of genes participating in the corresponding disorder (see key), and the link thickness is proportional to the number of genes shared by the disorders it connects. The name of disorders with >10 associated genes are indicated, as well as those mentioned in the text. For a complete set of names, see Goh SI Fig. 13. (b) In the DGN, each node is a gene, with two genes being connected if they are implicated in the same disorder. The size of each node is proportional to the number of disorders in which the gene is implicated (see key). Nodes are light gray if the corresponding genes are associated with more than one disorder class. Genes associated with more than five disorders, and those mentioned in the text, are indicated with the gene symbol. Only nodes with at least one link are shown.

Curated disease phenotype and gene information has been combined in "*diseaseome*" networks to study the relationship between network similarity measured by distance and phenotype similarity. Goh and colleagues present a gene-disease network extracted from OMIM to describe the topology of disease genes and highlight genes that overlap multiple disease types[101]. The "Human Disease Network" and "Disease Gene Network" constructed by Goh are shown in Figure 3. Bauer-Mehren[102] present a comprehensive database of gene and disease associations and find that the corresponding network displays a modularity of disease genes, and furthermore, there are a core set of biological pathways underlying most human diseases. A core set of disease related processes in the disease interactome was also suggested by Janvic and Pruzjili[103]. Phenotype-genotype relations, GO similarity[104], microRNA[105,106], text–mining[107,108], coexpression data[109,110], and SNP/eQTL/mutation[111-114] data have also been used in functional disease networks to identify genes with similar disease phenotypes. Xu et al.[107] use toplogical features to describe similarities between disease genes in a PPI network and apply a linear classifier to identify diseases with similar genetic signatures and prioritize novel disease genes. Lavi et al.[115] find that coexpressed genes tend to be closer in the network of interactions and use an SVM classifier to define specific signatures for expression phenotypes. Wu[116] and Li[117] merge PPI, genotype-phenotype information and known gene-disease relationship to search for candidate genes in an integrated network and find that using network similarity and distance measures to model genotype-phenotype evidence improves the search for candidate disease genes.

Based on the hypothesis that disease genes will be closer to and share topological features with other disease genes in an interaction network, several studies use seed disease genes to search for neighboring putative genes. Such approaches map query genes onto a biological network to prioritize closely related genes for further research. Wu and colleagues[118] seed a PPI-glioblastoma network with known cancer genes to search for neighboring genes associated with glioblastoma. Shi et al.[119] label an interaction network with published colorectal cancer signatures and use a network-derived signature to train a SVM classifier, resulting in highly predictive cancer signature. Other studies seed the interaction network with query genes based on disease-specific experimental data, for example siRNA[120], copy number[121,122] variation or proteomic data[123,124], to find disease-related modules. Such methods have been applied to identify disease-related genes in Alzheimer's Disease[125], Parkinson's[126], cardiovascular disease[127-130] and type-1 diabetes[131].

Applications of this type have proved to be particularly useful in identifying key interactions in regulatory networks[132]. Several studies link transcription factors and their

known targets in PPI networks to isolate dysregulated regulatory interactions in disease[133-135]. MiRNA-mRNA interactions have also been studied to identify important regulatory relationships in putative phenotypes[136-139]. Liu and colleagues model DNA-methylation interactions in a PPI network to detect cancer-related genes with aberrant methylation[133]. Other regulatory and epigenetic information added to biological networks to improve the understanding of disease processes include adding splicing factor information[140,141], mutation analyses[142], B-cell signaling interactions[143], copy number variation/somatic mutations[144,145], GWAS[122], and protein adduct information[146].

Complex diseases are characterized by the coordinated dysregulated activity of multiple genes and biological processes. This feature of complex phenotypes has motivated the search for groups of interacting genes or modules, associated with disease. Hartwell discusses the importance of modular structures of biological pathways and defines biological modules as proteins or protein complexes that "interact with preferred partners weakly, transiently, or conditionally forming a biological module serving a specific collective function"[147]. Corresponding to this definition, network modules are subnetworks of a biological network that comprise a set of highly interconnected nodes within a larger network, and their definition is not restricted to, but may overlap with pathway interactions or functional complexes. While pathway and data conforms to strict (yet somewhat arbitrary) pathway boundaries, module membership is based on coordinated activity across multiple biological processes and interaction types. Module data has been used in genomic studies to annotate genes[148] and study complex genetic interactions associated with experimental outcomes[149-151]. Prominent community finding algorithms are based on cliques, edge-betweenness, label-propagation, spectral approaches, and clustering algorithms using distances generated by random walks[152,153].

Module discovery approaches are agnostic to abstract definitions of gene sets, and can therefore account for interactions among members of gene sets, circumventing the generally arbitrary boundaries associated with curated pathways. Using interaction information generated from coexpression data, Segal et al.[87,154] identify graph modules and analyze module expression to detect regulatory genes related to cancer. Dittrich and colleagues [155]apply a Steiner Tree to identify subnetworks of cancer-related genes microarray studies. The algorithm finds an optimally connected subgraph spanning a network weighted by expression data. Variations of similar linear programming approaches have been applied by Zhao[156] and Backes[157]. Ideker and Chuang et al.[149,150,158] developed *jActiveModules*, a Cytoscape plugin, to search for significant modules in expression data. They apply a

simulated annealing algorithm to seed and construct the modules. Later work by Ulitsky [151] uses a seed-based clustering algorithm to discover significant cancer modules and to find minimally connected subnetworks to describe gene signatures in case control studies. Alcaraz et al.[159] develop a variation of this approach to find maximally connected subnetworks. Maximal cliques[119,160], diffusion processes[161], geometric clustering[162], SVM[163], and mutual exclusivity[144] methods have also been applied to define local neighborhoods of disease-related genes. In summary, these studies conclude that graph-based algorithms can successfully identify functionally relevant, cancer-associated modules in expression data.

2.5: Graph-based Random Walks in Gene Prioritization and Module Discovery

Among graph-based approaches, the random walk on graphs has many applications and is an effective approach to define distances between nodes. The walk begins at a random node and at each time point takes a step to an adjacent node. A transition matrix determines the probability that the walker will visit node *i* by time *t*, and this matrix is used to calculate distances between nodes. These distances can be applied to measure similarity between nodes, and as input to clustering methods to discover densely-connected communities in large graphs. The random walk algorithm is a powerful method among other community detection algorithms, based on studies evaluating the performance of comparable network clustering methods. As compared to other community discovery approaches, studies by Navlakha and Kingsford [164] and Orman and Labatut [153] find that random walk approaches individually outperform clustering and neighborhood approaches in PPI networks.

Random walks on networks have been applied in the context of genomic studies to search for disease-related genes. Kholer et al.[112] apply a random walk algorithm in a functional interaction network to identify novel disease genes by their proximity to known disease genes based on genome mapping and interaction data. They conclude that random walks outperform other distance-based methods in prioritizing related disease genes. Li et al.[165] and Yao et al.[166] use a random walk to find disease genes in a genome-phenome network, with specific applications to cancer data. Tu et al. [167] implement a heuristic random walk in an integrated network to find regulatory modules in gene expression data, identifying the most likely path from quantitative trait loci to a candidate gene. More recently, Komurov et al. [168,169] use a random walk algorithm with gene expression data to prioritize candidate genes and discover clusters associated with cancer and other outcomes. These studies show

the random walk has been well-adapted to genome studies, and can be used to efficiently describe similarity measures between entities in biological networks.

Implementations of the random walk differ based on the distance metrics used and optimization or heuristic strategies. The focus of this study will be a specific implementation of a random walk algorithm, *Walktrap*[7] which merges an optimized distance metric and modularity-based calculation for community-finding. The algorithm is developed by Pons and Latapy [7] and implemented in *igraph*[170]. As the algorithm typically becomes trapped within a local community, it is referred to as *Walktrap*. The random walk, compared to other popular hierarchical clustering approaches, or seed clustering methods, utilizes the structure of the network to build distance metrics, and the *Walktrap* random walk algorithm optimizes the community search using the graph theoretic concept of modularity. Using distance metrics defined by transition probabilities, the algorithm searches the network for communities of closely related nodes to find cancer-associated modules. *Walktrap* has performed optimally in terms of accurate and efficient community finding in large networks[153,171]. Further, in benchmark testing, I found the random walk to be computationally more efficient than the using edge-betweenness, spectral methods, or spanning trees to detect communities.

We begin with graph G and its associated adjacency matrix A. When graph G is unweighted $A_{ij} = 1$ if i and j are connected in G, and $A_{ij} = 0$ otherwise. In the weighted network, $A_{ij} \in \mathbb{R}^+$. The random walk process starts at a vertex i and at each time point in the walk of length t, a random step is taken to an adjacent node j. Here t is set to 3. The transition probability at each step is $P_{ij} = \frac{A_{ij}}{d(i)}$ where $d_{(i)}$ is the degree of vertex i, $d(i) = \sum_j A_{ij}$. Transition probabilities define the transition matrix $\boldsymbol{P}$ of the random walk, and powers of $\boldsymbol{P}$ determine the probability P^t_{ij} that the walker will traverse from i to j over time t. As t tends towards infinity, this probability tends towards the degree or weighted degree (strength) of vertex j:

$$\forall_i \lim_{t \to +\infty} P^t_{ij} = \frac{d(j)}{\sum_k d(k)} \tag{5}$$

where k is an index of all nodes n in graph G. Pons and Latapy calculate structural similarity between vertices and communities using probabilities P^t_{ij} to measure the distance between two nodes. This calculation has an important advantage compared to other distance metrics

in that it can be computed efficiently and can be used in hierarchical clustering. The distance between the two vertices *i* and *j*, r_{ij} is given by:

$$r_{ij} = \sqrt{\sum_{k=1}^{n} \frac{\left(P_{ik}^{t} - P_{jk}^{t}\right)^2}{d(k)}} \tag{6}$$

Similarly, the distance between two communities C_1 and C_2 is:

$$r_{C_1C_2} = \sqrt{\sum_{k=1}^{n} \frac{(P_{c_1k}^{t} - P_{c_2k}^{t})^2}{d(k)}} \tag{7}$$

where $P_{c_jk}^{t}$ measures the probability of traversing from a node in C_j to node k (j=1,2). At each step in the merge algorithm, two communities in partition R are selected to be merged if the merge minimizes the mean σ_k of the squared distances between each vertex and its community:

$$\sigma_k = \frac{1}{n} \sum_{C \in R_k} \sum_{i \in C} r_{iC}^2 \tag{8}$$

After the merge step, the decrease in squared distances $\Delta\sigma$ between the communities is found by:

$$\Delta_\sigma(C_1 - C_2) = \frac{1}{n} \frac{|C_1||C_2|}{|C_1| + |C_2|} r_{C_1C_2}^2 \tag{9}$$

And the distance, $\Delta\sigma(C_3, C)$, between a community C_3 (resulting from the merge of C_1 and C_2), and any other community C is:

$$\Delta\sigma(C_3, C) = \frac{(|C_1| + |C|)\Delta\sigma(C_1, C) + (|C_2| + |C|)\Delta\sigma(C_2, C) - |C|\Delta\sigma(C_1, C_2)}{|C_1| + |C_2| + |C|} \tag{10}$$

Modularity Q is maximized when the fraction of edges e_C inside the community C is compared to the fraction of edges bound to community C, a_C in partition R:

$$Q(R) = \sum_{C \in P} e_C - a_C^2 \tag{11}$$

Further background and details of the *Walktrap* implementation are provided in the original work [7].

Chapter 3: Assessing Network Characteristics of Cancer Associated Genes in Metabolic and Signaling Networks

3.1: Introduction

Cancer and other complex diseases have intricate roots in the molecular pathways of the cell. There is growing interest in the application of graph theoretic methods integrating metabolic, regulatory and signaling interactions to study cancer in the context of complex and conditional biological phenomena associated with the disease. Many recent pathway or network-based approaches merge network interaction information with high-throughput experimental methods to search for putative genes of clinical interest. In the context of cancer, these methods are applied to discover novel cancer genes that contribute to the complex phenomena of cancer onset or progression, or play a key role in cancer pathways.

A body of work leveraging biological network information finds enrichment of differentially expressed genes in gene sets representing canonical pathways, protein complexes, functional GO categories or network modules[75,77,172,173]. Network degree measures have been used to identify important hub genes and their interacting partners in cancer [174,175] and as features in a support vector machine classifier to detect cancer genes[176]. Graph algorithms have also been applied to discover dense modules of cancer genes in protein-protein interaction (PPI) networks weighted by expression data[150,151,155]. Other approaches map query genes onto a disease network of closely related known disease genes to prioritize novel candidate genes for further research[112,116]; or use prior knowledge of genes associated with cancer (from the literature review, curated sources, or experimental data) to define seed nodes to search for interacting disease genes in PPI networks[123,124,177]

These approaches build on a common hypothesis that disease genes share observable patterns in biological networks. Namely, that genes that are critical in cancer etiology also play central role in the network topology and tend to cluster with other cancer genes. Therefore, the extent to which these genes act as network hubs and the likelihood that they form communities with other cancer genes have implications in the design of experiments that search for novel cancer genes and their interactions in biological networks.

Previous studies have shown correlation between cancer genes and the topological features of biological networks. Jonsson and colleagues[94] conclude that cancer genes are more well-connected and more likely to lie at community interfaces than non-cancer genes in an orthologous PPI network. Further, using lung cancer expression data mapped to a PPI network, Wachi et al.[100] show that cancer genes have a higher degree connectivity compared to normal genes. Sun and Cai[96,97] draw similar conclusions but note an inverse relationship between cancer genes and network clustering coefficients. These studies support the theory that essential genes and cancer genes are more likely to act as highly connected hubs in biological networks; though, in general, genes related to other complex disease do not tend to exhibit high centrality[97,101].

This study investigates the relationship between network features and cancer genes in signaling and metabolic interaction networks and quantifies the predictive value of these features using a generalized linear model[5]. Metabolic networks are distinguished from signaling networks in feature selection to investigate relative differences in cancer gene topology. I assess the modularity of cancer communities in each network to identify cancer gene-enriched modules. Previous methods using curated cancer gene information focus on nearest neighbor and seed approaches to rank nearby interactions as potential disease genes. Further, I apply a community detection algorithm based on a random walk to find cohesive communities of cancer-related genes in cancer-enriched modules. This algorithm improves upon previous network-naïve clustering approaches, as it considers the network structure when calculating distances between nodes and community-finding; and, in contrast to approaches based on pathway enrichment, the modules focus on activity of specific interactions within pathways and span multiple pathways.

3.2: Methods

3.2.1:Overview

I develop a KEGG parser to extract data from the Kyoto Encyclopedia of Genes and Genomes (KEGG)[6] to generate signaling and metabolic networks. To analyze network statistics, I calculate and compare means and employ a generalized linear model to assess

predictive network characteristics. I then implement a random walk algorithm to search the network for dense communities of cancer genes and evaluate results using functional annotation and evidence from the literature.

3.2.2: Network Construction

KEGG interaction data was extracted from KGML files using a custom parser (Appendix A). This interface facilitated the retrieval of pairwise interaction data and metadata from XML formatted files to create a comprehensive list of interactions corresponding to metabolic and signaling pathways. KEGG enzyme IDs are resolved as lists of KEGG gene IDs and each gene is translated to HUGO gene symbols. Metabolic interactions were defined as a relation between two neighboring enzymes that share a common metabolite; signaling interactions were defined as two genes that participate in a signaling cascade, where both genes share a reaction event.

Pairwise interactions were used to construct global metabolic and signaling networks. To build these networks, I processed 141 signaling pathways and 83 metabolic pathways, representing 95% of KEGG pathways. The resulting directed metabolic and signaling networks consist of 1302 vertices and 15923 interactions (Figure 4), and 2989 vertices and 16772 interactions (Figure 5), respectively. Networks were created and analyzed using the *igraph* package in R[53,170].

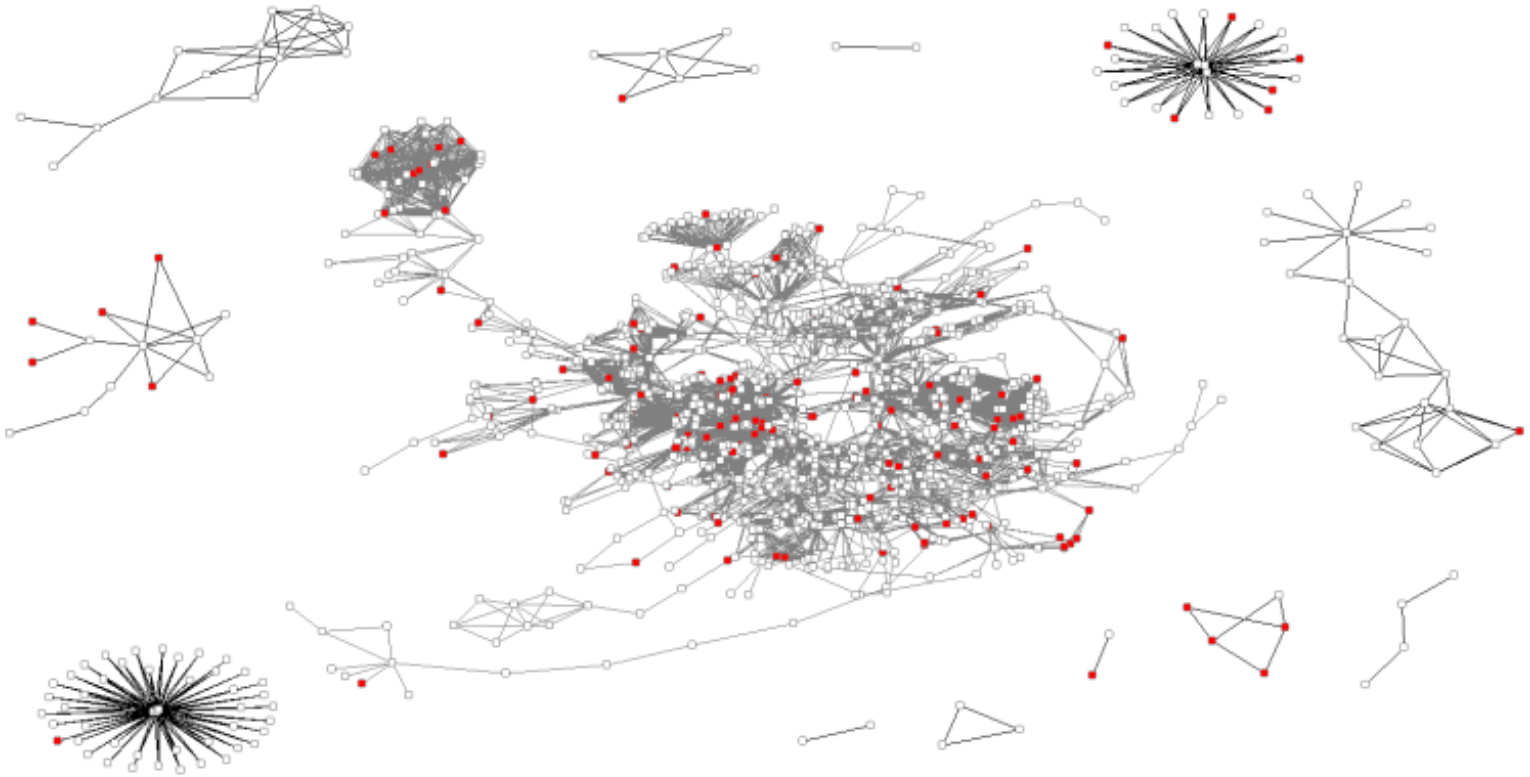


Figure 4: The Global Metabolic Network: Figure 4 shows the metabolic network derived from KEGG consisting of 1302 vertices and 15923 edges (interactions). Red nodes designate genes implicated in cancer risk or etiology. The network shows that the majority of KEGG genes are connected in a large component. Cancer genes appear more widely dispersed in the metabolic network relative to the signaling network

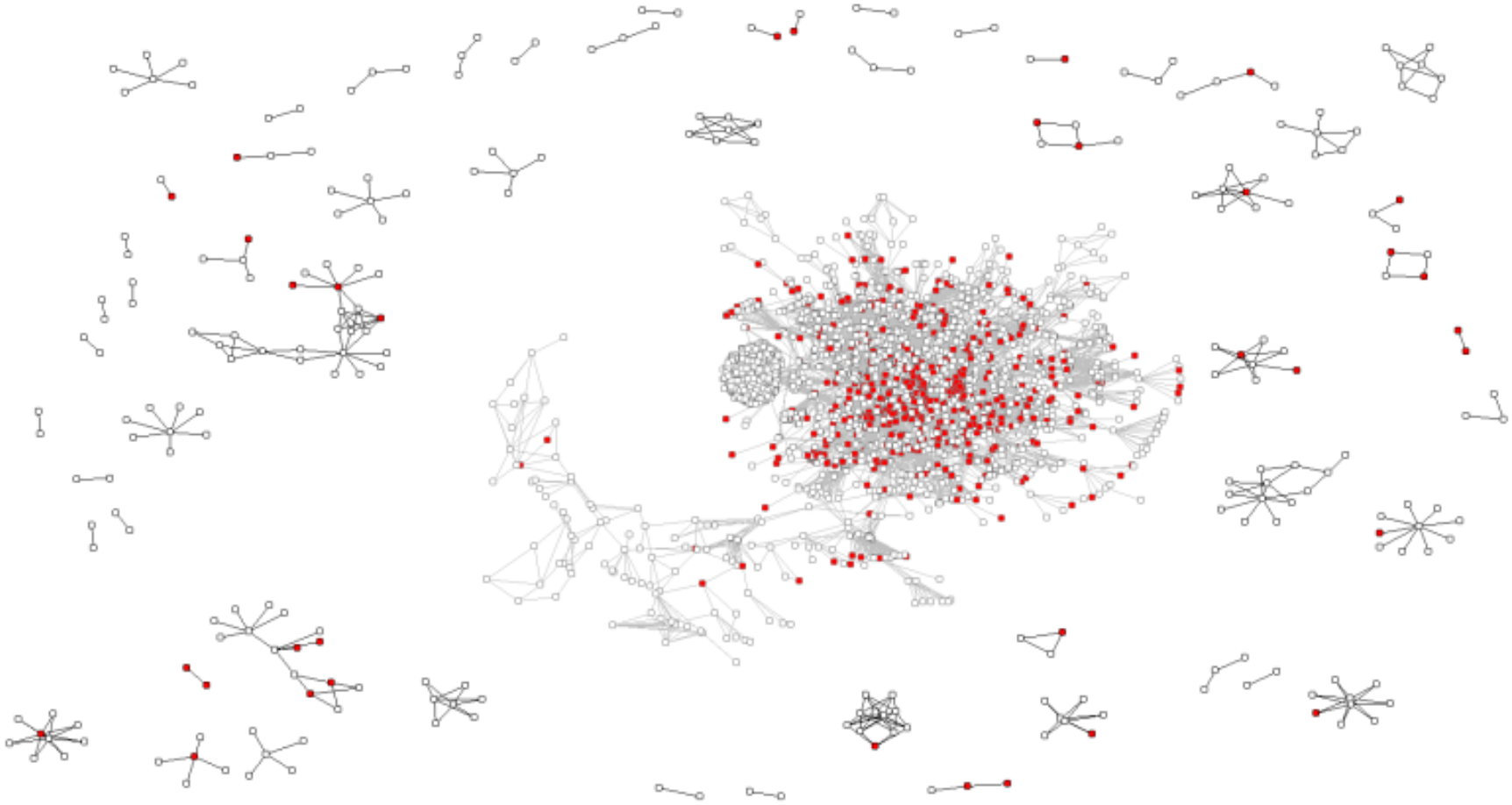


Figure 5: The Global Signaling Network. Figure 5 shows the signaling network derived from KEGG consisting of 2989 vertices and 16772 edges (interactions). Red nodes designate genes implicated in cancer risk or etiology. The majority of KEGG genes are connected in a large component. Cancer genes appear more cohesive in the signaling network relative to the metabolic network.

3.2.3:Definition of Cancer Genes

The sample set of genes in the feature analysis include 4291 genes from the metabolic network and signaling networks (Supplemental Files). To find a reliable subset of cancer associated genes, I investigated these genes in OMIM[38] for evidence that the gene might be involved in promoting or inhibiting cancer progression. All genes in the study were queried to determine if the gene is related to cancer (ie. "cancer", "carci-", "onco-", "leukemia", "tumor). Each matching record was reviewed to confirm that the gene was in fact a tumor suppressor, oncogene, or otherwise shown to be significantly associated with cancer status (ie. by differential expression, functional pathway analysis or SNP studies). Of the genes in the metabolic network, 164 were found to be associated with cancer; while 634 were associated with cancer in the signaling network. Approximately 5% of genes did not have corresponding record in OMIM. In such cases, the gene was labeled as non-cancer because of lack of data and was included with non-cancer gene class in subsequent analyses.

3.2.4: Network Features

Network features representing centrality measurements and the clustering coefficient were evaluated for their predictive ability. These features were selected to compare the network characteristics of cancer genes and non-cancer genes and to assess the relative importance of cancer genes in the toplogy of metabolic and signaling networks. Centrality features measuring degree, betweenness, and closeness, as well metric for node-level clustering coefficients are included in the analysis (described in Section 2.3).

3.2.5: Statistical Analysis

Network characteristics were examined using thresholds T of the top 15% and 20% of genes in each category. Cutoff values were chosen based on the performance of estimates of T, ranging from 5-30, in a test sample of six pathways. Genes were coded as 1 if they were greater than or equal to the threshold T and 0 otherwise; and these variables were used to fit the subsequent predictive model. To determine if there was a significant difference in means between cancer and non-cancer genes, I compared the mean of each feature using the

Wilcoxon rank sum test and applied logistic regression to assess the predictive value of these features.

Logistic regression is a type of generalized linear model suited to testing a discrete outcome and allows for the inclusion of multiple covariates, or interactions, in the model. It is described by:

$$log(\frac{\theta}{1-\theta})_j = \alpha + \beta_j x_j + W_g. \tag{12}$$

Here, for the *j*th predictor, the log odds represent the probability of case status, where θ is the average effect on cancer status given a positive predictor x_j. α is baseline risk, x_j is the exposure of interest, β is the coefficient, and W is the vector of covariates. The null hypothesis $H_0 = 0$, or no association between the feature and cancer status is modeled by,

$$H_0: \beta i = \beta 2 = \ldots = \beta m = 0. \tag{13}$$

The significance of the fit of the logistic regression model with Wald statistics, p-values and odds ratios are reported. The Z-score associated with the Wald estimate is based on a χ-squared distribution. The odds ratios describe the log odds of the status of the dependent variable *y* based on a unit change in *x*. Here *x* is a binary variable, 1 if *x* >= the threshold of the network feature, 0 otherwise. The ratio is calculated by dividing the odds of *y*=1 status given a positive predictor, by the odds of *y*=1 given a negative predictor. Examples of code used to query network statistics and calculate logistic regression values are presented in Appendix A.

3.2.6: Community Analysis

Random walks have been shown to be valuable when applied to study genomic data in biological networks[112,167]. The random walk algorithm implemented here was chosen because it incorporates the topology of the network to calculate distance metrics, and optimizes the community search component using the graph theoretic concept of modularity. Details of the random walk are described in section 2.5.

3.3: Results and Discussion

3.3.1: Global Network Statistics

Summary statistics were calculated for metabolic and signaling networks. The signaling network consists of 2989 vertices and 16772 interactions, with a diameter of 21 and a global clustering coefficient of 0.0943, compared to a clustering coefficient of 0.0037 in a randomly generated network with the same number of nodes and edges. The metabolic network consists of 1302 vertices and 15923 interactions with a diameter of 23; the clustering coefficient is 0.3910 compared to 0.0184 in a random network. Higher clustering coefficients in the biological networks reveal regions of dense, interconnected nodes. Modularity in the signaling network is 0.6431 and 0.6789 in the metabolic network versus 0.0028 and 0.0048 in corresponding random networks, suggesting that the topology of these biological networks exhibit high underlying modularity. All Wilcoxon p-values comparing centrality and clustering coefficient means between cancer and non-cancer genes in the signaling network were significant (betweenness p=1.404e-06, close p=2.2e-16, degree p=1.019 e-10, in-degree p=0.001124, out-degree p=1.093 e-10, CC p=3.489 e-06). In the metabolic network, differences among in-degree and out-degree means between cancer genes and non-cancer genes were significant (in-degree p=0.02301, out-degree p=0.01393).

3.3.2: Feature Prediction

Logistic regression is used to assess the predictive power of centrality features and clustering coefficient in the signaling and metabolic networks. Signaling networks exhibit highly significant centrality measures using thresholds $T = 20$ and $T = 15$. Closeness centrality was the most significant predictor, with highly significant p-values and significant odds ratios at the following thresholds: $T = 20$ ($p = 6.04^{e-14}$, OR= 2.77, SE= 1.15), and, $T = 15$ ($p = 1.83^{e-10}$, 6.04^{e-14}, OR= 2.66, SE= 1.17). P-values for all centrality features remained significant after Bonferroni correction, and were associated with odds ratios above 1 as summarized in Table 1. Notably, along with closeness, betweenness and overall degree were the strongest predictors of cancer gene status. The clustering coefficient was significant after correction for $T = 15$ (p=.0025, OR= .51, SE= 1.25), but not $T = 20$, and the estimates showed a negative association with cancer status. A negative association with clustering

coefficient is consistent with prior observations that while cancer genes tend to act as network hubs, their immediate neighboring nodes do not tend to be highly-connected[96,97].

Analysis of metabolic networks resulted in significant p-values after multiple testing correction for the following features and thresholds: out-degree $T = 20$ (p= 0.0001, OR= 2.01, SE= 1.20), in-degree $T = 15$ (p= 0.001, OR= 1.95, SE= 1.22), and closeness at threshold $T = 20$, (p= 0.0016, OR= 1.81, SE= 1.21), and $T = 15$ (p= 0.0002, OR= 2.10, SE= 1.22). However, odds ratios for these tests were offset by variability of the standard error and were non-significant for all features in the metabolic network. As with estimates in the signaling network, the clustering coefficient $T = 20$ is negatively correlated with cancer, though in the metabolic network, these results are non-significant.

Table 1 Logistic Regression Estimates for Network Features

Feature	Metabolic Networks				Signaling Networks			
	Z-Score	P-value	OR	OR SE	Z-Score	P-value	OR	OR SE
Top 20 Percent								
In Degree	2.01	0.0443	1.48	1.21	4.72	2.3200E-06	1.92	1.15
Out Degree	3.86	0.0001	2.01	1.20	4.75	2.0800E-06	1.94	1.15
Degree	3.21	0.0013	1.83	1.21	5.56	2.6700E-08	2.39	1.17
Between	1.47	0.1418	1.37	1.24	6.48	9.2100E-11	2.43	1.15
Close	3.15	0.0016	1.81	1.21	7.51	6.0400E-14	2.77	1.15
Clustering Coef	-0.11	0.9094	0.98	1.24	-2.60	0.0092	0.63	1.20
Top 15 Percent								
In Degree	3.30	0.0010	1.95	1.22	5.51	3.5600E-08	2.27	1.16
Out Degree	2.56	0.0105	1.71	1.23	5.17	2.3200E-07	2.15	1.16
Degree	2.44	0.0146	1.65	1.23	6.48	8.8800E-11	2.41	1.15
Between	0.50	0.6192	1.11	1.23	5.71	1.1400E-08	2.42	1.17
Close	3.71	0.0002	2.10	1.22	6.38	1.8300E-10	2.66	1.17
Clustering Coef	0.49	0.6261	1.12	1.26	-3.03	0.0025	0.51	1.25

Further, I examined interactions between multiple centrality predictors and their joint effects. None of the tests of interactions using additive or multiplicative models improved the significance of univariate estimates. There is no evidence for interaction of network features to predict cancer genes. These results do not agree with those of other studies suggesting that a combination of two centrality features show significantly stronger association with cancer status than one feature[97,178]. Such interactions may not be apparent in this study due to collinearity of the statistics.

3.3.3: Community Analysis

The largest community in the metabolic network has 204 nodes; in the signaling network the largest community has 763 nodes. There are 27 singletons and 20 pairs in the metabolic network and 26 singletons and 55 pairs in the signaling network. The random walk algorithm yielded 74 total clusters with 3 or more nodes in the metabolic network and 169 clusters in the signaling network. Average community size for a non-cancer gene in the metabolic network is 71, compared to 84 for cancer genes. In the signaling network, average community size for a non-cancer gene is 241, and 314 for a cancer gene. Distributions of cluster sizes are shown in Figures 6 and 7.

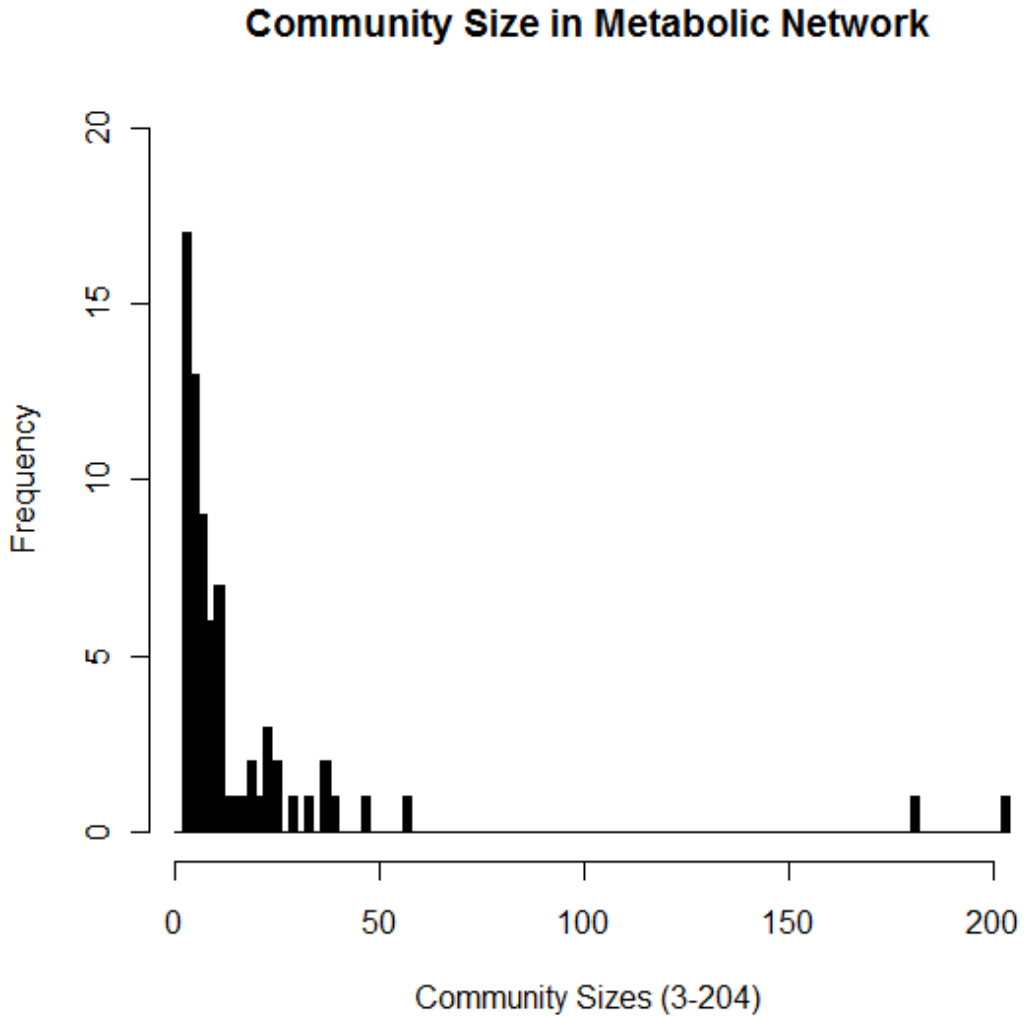


Figure 6: Distriibution of Community Sizes in the Metabolic Network. Community sizes for clusters with 3-204 nodes are shown. Most communities have fewer than 50 members and only a few communities have greater than 100 members. The largest cluster has 204 nodes.

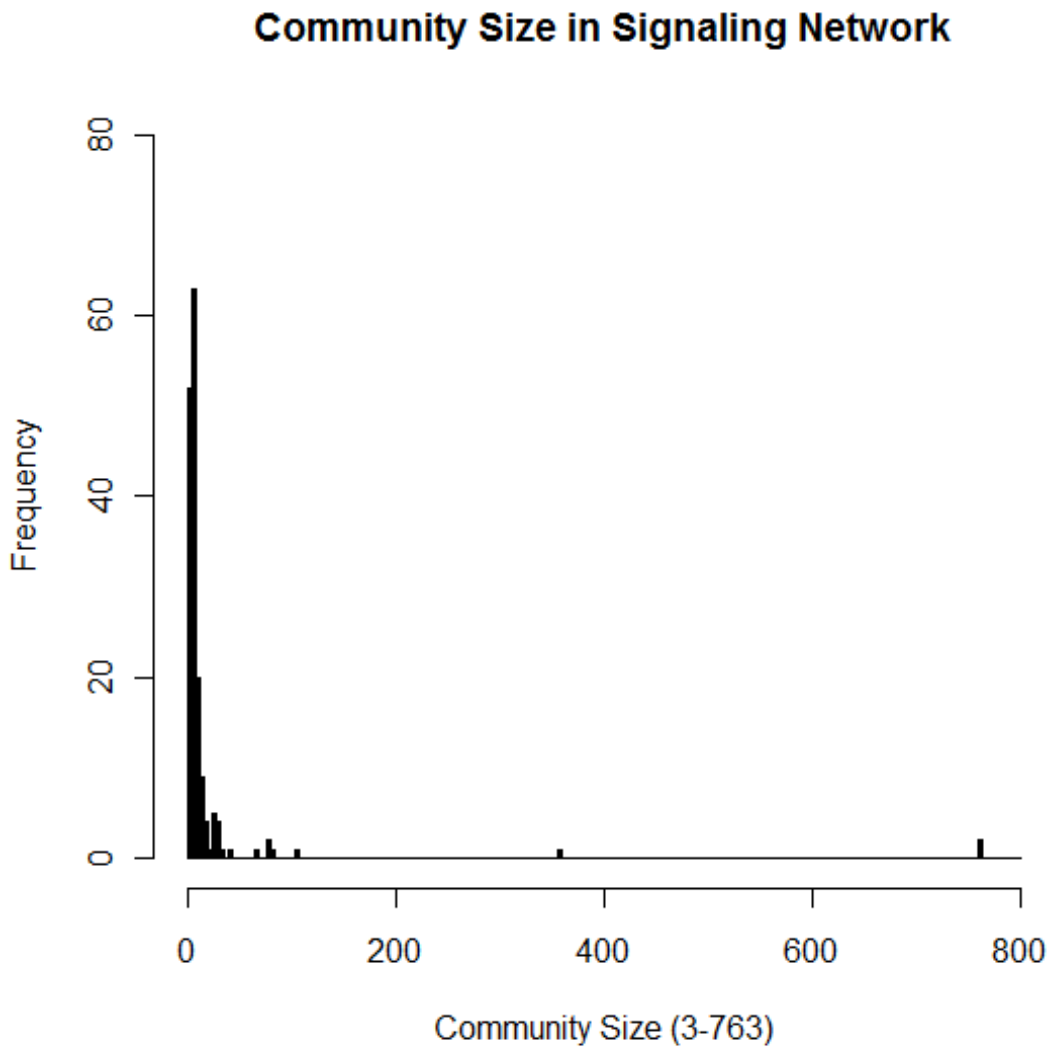


Figure 7: Distribution of Community Sizes in the Signaling Network. Community sizes for clusters with 3-763 nodes are shown. Most communities have fewer than 50 members and only a few communities have greater than 100 members. The largest cluster has 763 nodes.

Seven communities in the metabolic network were significantly enriched with cancer genes at α=.05. I exploredd the top five results. The most significant community is an exostosin gene family ($p=1.10^{e-3}$) consisting of five genes involved in glycosyltransferase activities and synthesis of heparan sulfate and heparin. This family plays a tumor-suppressor role and regulates cartilage and bone differentiation, ossification and apoptosis. The group is also involved in metabolism and differentiation signaling cascades. Among other top communities is a group of nine genes including *DNMT*, *AHCY* and *MAT* families, involved

in cysteine, adenosine, tyrosine and methionine metabolism functions, which help to stabilize cell replication ($p=.1.13^{e-02}$). A large family of 204 detoxifying genes, including *CYP*, *GST*, *HSD* and *UTG* genes, was significant at p=0.0045. These gene families play a key role in detoxification of carcinogens, and related mutations may predispose cells to cancer phenotypes. A group of nine genes, including tumor suppressors fumarate hydratase (*FH*) and succinate dehydrogenases (*SDHB, SDHC, SDHD*) involved in the citric acid cycle ,was also significant (p=0.01835). Finally, a group of thirty-six phosphate metabolism genes connected to *PIK3* and *PTEN* signaling cascades that control the cell cycle and differentiation (p=0.01844) (Figure 8) was identified. Most genes in this module are associated with cancer; however, potential genes that merit further investigation based on neighboring interactions are *SOX17, TYRP7* and *TCF7L*.

In general, top communities in the metabolic network largely regulate methylation, amino acid synthesis and metabolism, or are connected to differentiation, proliferation and growth signaling cascades. Cancer communities in metabolic pathways tend to be less cohesive than in signaling pathways as suggested by fewer, less significant communities. This may be attibuted to the dense background clustering coefficient in the metabolic community which makes identifying significant cohesion more challenging.

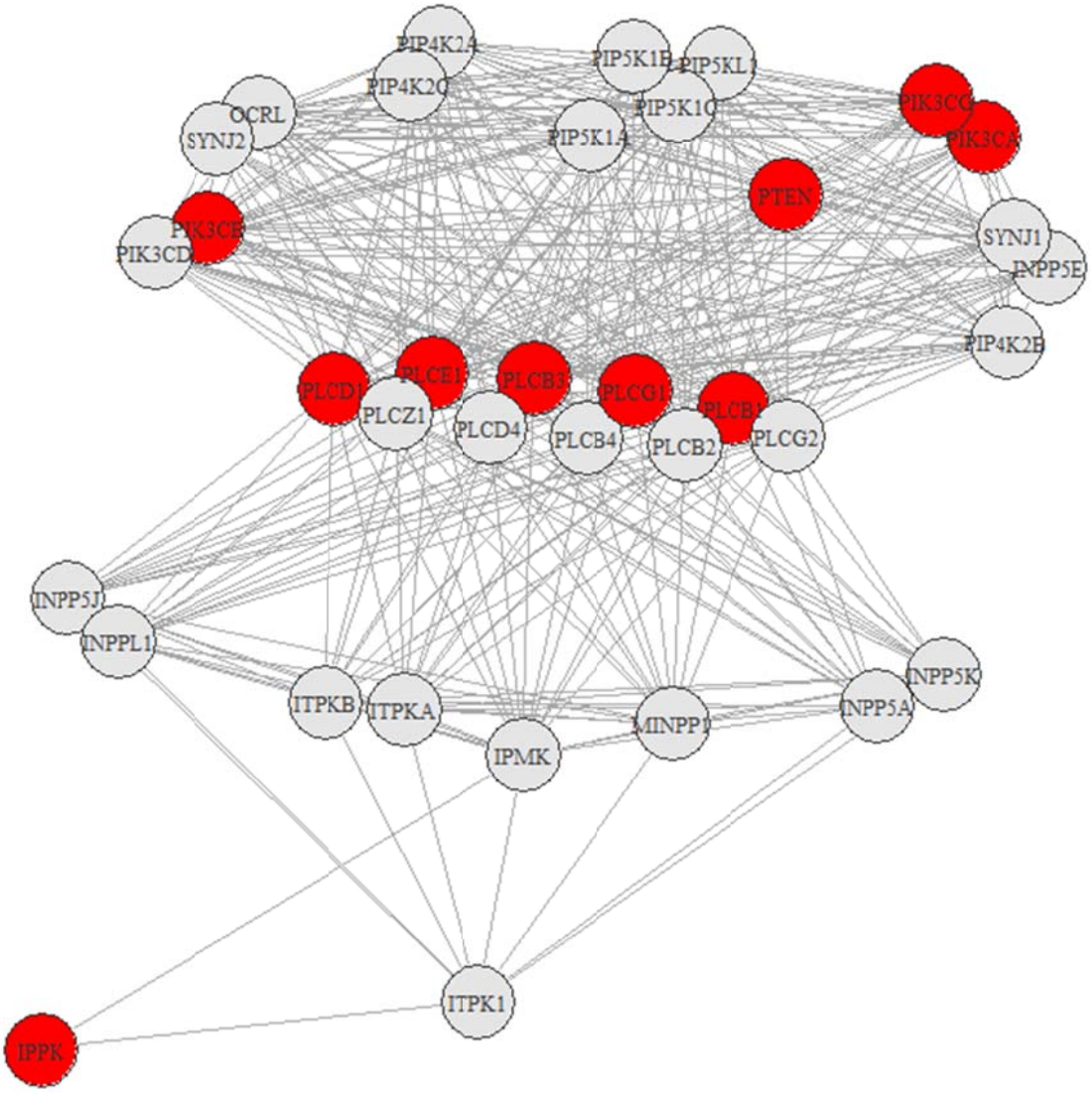


Figure 8: Metabolic Module Significantly Enriched with Cancer Genes. This module shows phosphate metabolism, *PIK3* signaling and *PTEN* signaling interactions in a community of thirty-six genes. Genes highlighted in red have known associations with cancer. The nework shows a number of phosphate metabolism genes act as cancer-related hubs in the network and others with a similar toplogy, such as *PLCZ1, PLCD4, PLCB4, PLCB2, and PLCG2*, can be interesting candidate cancer genes for further study.

Nineteen communities were significant in the signaling network at α=.05. Of the top five, one community of eighteen genes ($p=2.99^{e-08}$) is composed predominantly of genes in the *Wnt*-signaling pathway and interactions with cell adhesion, cell cycle and *TGFβ*-signaling, and arachidonic acid metabolism (Figure 9). Cancer-associated genes in this module include phospholipase-c genes; several genes in this family are associated with cancer and others that share a similar toplogy could be interesting candidate genes for further study, including *PLCZ1*, *PLCD4*, *PLCB4*, *PLCB2*, and *PLCG2*, as well as interactions with

PIP5 and *PIP4* signaling cascades. Another community of 107 nodes with a majority of *JAK-STAT* signaling genes that regulate growth and proliferation, shows interactions with cytokine receptor activity and tyrosine metabolism (p=0.0008). Cell cycle controls genes interacting with *p53*-signaling functions comprise another top community with eleven genes (p=0.000354). Genes in the *p53* pathway are also members of a community of thirty genes (p=0.000427).

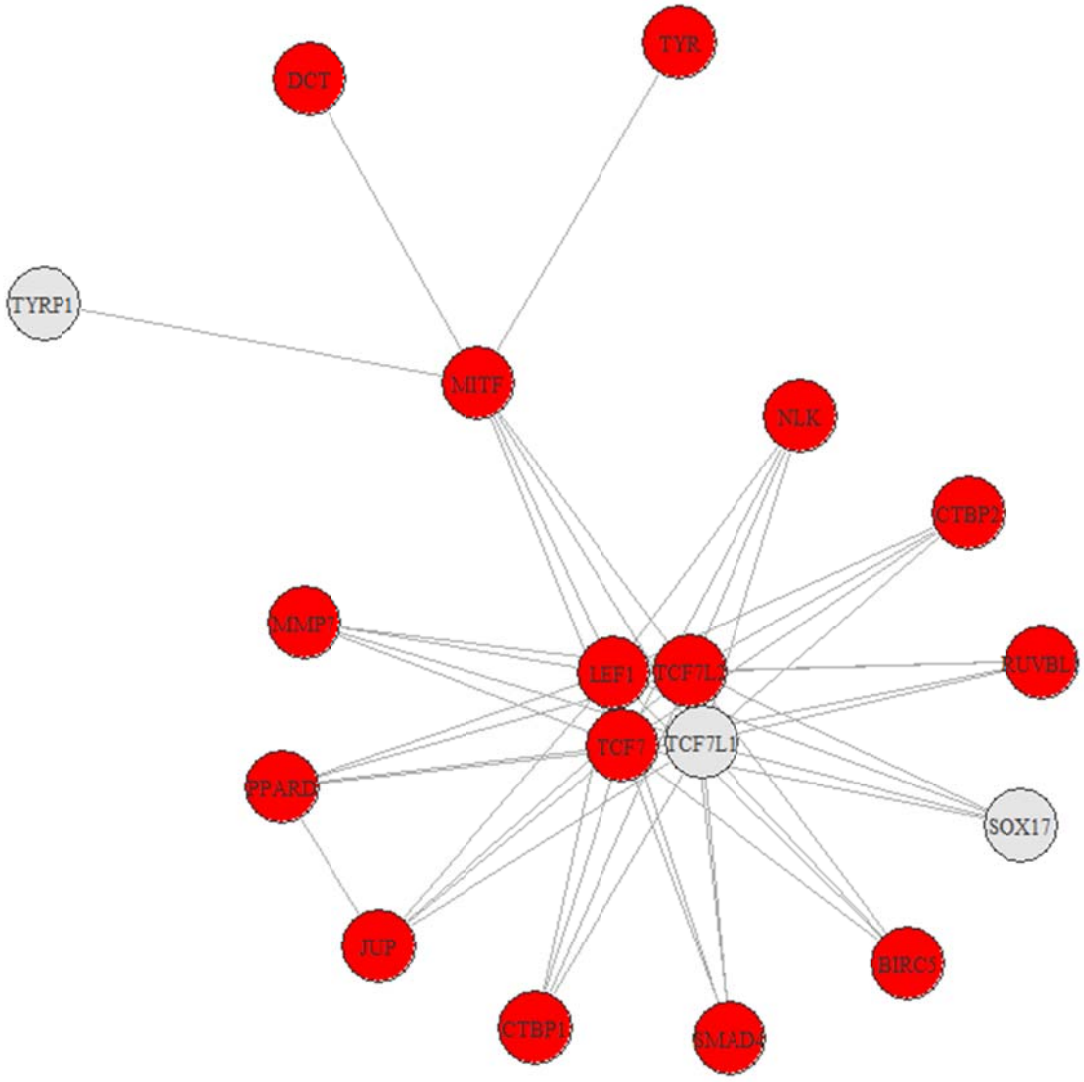


Figure 9: Signaling Subnetwork Significantly Enriched with Cancer Genes. The signaling subnetwork shows *Wnt*-signaling, *TGFβ*-signaling, cell adhesion and arachidonic acid metabolism interactions in a module of 18 genes. Genes are shown in increasing degree from left to right. Genes highlighted in green have known associations with cancer.

Of the top five modules, the most significant was a group of 763 genes ($p=1.80^{e-12}$). Approximately one in three genes in this community is associated with cancer. To investigate this large cluster of genes with greater resolution, I executed another iteration of the random walk using a threshold size of 200 for each community. This analysis identified a number of modules representeing interconnected signaling pathways, including *ErbB, mTOR, JAK-STAT, VEGF* and T-cell and B-cell signaling. The majority of cancer genes in these communities are also classified as oncogenes.

In summary of the analysis of metabolic networks, I found significant metabolic communities related to amino acid synthesis and metabolism, methylation regulation and signaling pathway interactions. The signaling modules represent a number of common pathways, consisting of *Wnt*-signaling, *JAK-STAT*, cell-cycle, *p-53* signaling communities and a very large community highly populated with genes involved in oncogenic signaling pathways such as *ErbB, mTOR*, and *VEGF*.

3.4: Conclusion

These results demonstrate that topological features in global metabolic and signaling networks exhibit predictive value in identifying known cancer genes, particularly in signaling networks. Wilcoxon rank sum comparisons between the mean values of network centrality and clustering coefficient are highly significant in signaling networks and moderately significant for measures of in-degree and out-degree in metabolic networks. Logistic regression estimates further quantify the predictive ability of centrality and clustering coefficient and show more predictive power in signaling networks compared to metabolic networks. Clustering coefficient is also significant in signaling networks, but shows an inverse correlation with cancer status, suggesting, in agreement with previous work, that although cancer nodes are highly connected, their neighbors are typically not well-connected [96,97].

Cancer genes in signaling communities tend to be more cohesive than those in metabolic communities and represent cell cycle, adhesion *Wnt*-signaling and *TGFβ*-signaling pathways among other cancer-related processes. When investigating the metabolic network, communities of cancer genes are frequently associated with methylation activity, amino acid synthesis and metabolism, and are characterized by interactions with signaling pathways. Many significant communities in both networks include interactions between signaling and

metabolic pathways. Thus, whereas treating metabolic and signaling pathways as distinct networks may increase power and accuracy; dense cancer modules often include genes that participate in both metabolic and signaling pathways.

Study bias or the large proportion of cancer genes in signaling pathways may influence the statistical evaluation of cancer genes in this study. However, the consistency and strong statistical results across topological measures and functional validation support an underlying association between the network centrality features and cancer.

These results have implications for future work mining for cancer genes using network proximity and degree, prioritizing gene targets and searching for disease-related metabolic and regulatory pathways. Network features can be of predictive value in identifying novel cancer genes, and examination of modules enriched with cancer genes can help elucidate complex interactions influencing cancer onset and progression. This evaluation integrates known cancer data with pathway interaction data and shows that key cancer genes group with other cancer genes in modular communities via complex intra- and inter-pathway interactions. In comparison to single gene and pathway analysis, a modular approach also allows for the discovery of new gene targets based on their relationships with more prominent cancer genes, and identification of complex genetic interactions across pathway definitions. Within such subnetworks, one can investigate the intersection of pathway activity and identify novel cancer genes by their interactions with known cancer genes. In Chapter 4, I expand the biological network to integrate protein-protein interaction and experimental data to search for modules associated with cancer phenotypes.

Chapter 4: Using Random Walks to Identify Cancer-Associated Modules in Expression Data

4.1: Introduction

Cancer biology involves an intricate series of genetic and environmental interactions that act in concert to influence the onset and progression of disease. The complex nature of this information motivates the search for analytical tools that can model these interactions to examine associations between gene interactions and cancer. Graph analyses facilitate these genotype-phenotype investigations by integrating evidence of biological interactions from high throughput experiments, the literature, and a growing number of online databases. Such networks provide a useful framework to study genes in the context of protein complexes, molecular processes, or biological modules.

Network and pathway-based approaches have been developed to search for enrichment of groups of genes, rather than individual genes, associated with clinical outcomes. Gene Set Enrichment Analysis (GSEA) [79] is a computational method that considers *a priori* defined gene sets to investigate expression data for significantly enriched sets of genes or pathways. GSEA focuses on the significance of groups of interacting genes rather than single-gene analyses; and variations of gene set analysis have been developed to improve statistical validity[75-78] and to use more granular methods to study pathway activity [66,85,86,154]. However, these approaches are limited in their ability to search for enriched genes that form a small component of large pathways, or genes than span multiple pathways.

Network analyses show promise in expanding the search for disease genes by investigating genes in the context of integrated curated and experimental interactions. Several studies have evaluated the topology of disease genes in these networks and found that disease genes tend to cluster with other disease genes[101], and that cancer genes are characterized by high centrality and cohesiveness in interaction networks [94,179]. Building on the hypothesis that nearby genes in an interaction network share a common biological function, other network studies seed disease genes in functional networks combining evidence of known disease genes from the literature, with eQTL or GWAS data to search for putative neighboring genes[107,112,176]. Similar applications integrate experimental data in the

interaction network, for example, significant genes from siRNA or proteomic experiments, to discover candidate genes given their proximity to query genes [123,124,177].

In Chapter 3 I establish a basis for using network centrality features and module-finding to identify cancer genes. Related studies apply graph-based approaches to construct cancer-associated modules using clinical data. Dittrich and colleagues[155] implement a Steiner Tree to find parsimonious subnetworks of cancer-related genes in microarray studies. The algorithm finds an optimally connected subgraph spanning an interactome weighted by expression data. Ideker et al. and Chuang et al.[149,150] apply a simulated annealing algorithm to find significant subgraphs associated with cancer in a protein interaction network. They initiate subgraph generation with seed genes and add nearby proteins to the subgraph until a maximum score is reached reflecting significant activity of the module in the expression data. Ulitsky and Shamir[151] use a seed clustering algorithm to discover significant modules in yeast and human cell cycle data. They use multiple heuristics to generate seeds in the network and similarity between genes to build clusters. These studies conclude that searching for modules in graphs can successfully identify functionally relevant modules in expression data.

Random walks have demonstrated strong performance in genomic studies, and when evaluated against other graph clustering algorithms used to partition complex networks [153,168]. Distances determined by the random walk are drawn upon to prioritize genes, or to cluster genes into modules. Kholer et al.[112] apply a random walk algorithm in a functional interaction network using known disease genes, interaction information and eQTL data. They identify novel disease genes determined by their proximity to known putative genes. Tu et al.[167] employ a heuristic random walk in an integrated network to find regulatory modules in gene expression data, identifying the most likely path from quantitative trait loci to a candidate gene. Komurov et al.[168,169] implement a random walk to search for cancer-related genes and their interactions in an integrated network. Their methods account for differential expression across experimental conditions and local network connectivity to prioritize candidate genes and hierarchically cluster genes into cancer-related subnetworks.

The performance of random walks in large, complex networks vary based on their distance metrics and greedy-search heuristics; and few random walk algorithms are tailored to community-finding. I implement a random-walk and community search algorithm, *Walktrap*[7], which is optimized for large networks and integrates a community search driven by distance metrics that are determined by transition probabilities. This algorithm has shown high efficiency and accuracy in revealing community structure in large networks[180]. *Walktrap* is applied in an expression-weighted interaction network consisting of metabolic,

signaling and protein interactions to discover, score and evaluate modules that are significantly associated with cancer outcomes. I employ stopping criteria in the clustering process using modularity, module size or maximum module score, to improve the search for informative modules. This approach demonstrates strong performance when compared with similar tools developed to discover subnetworks of disease genes in interaction networks and to identify functionally relevant cancer-associated modules that highlight candidate cancer genes and their interactions.

4.2: Methods

4.2.1:Overview

I employed a graph-based random walk algorithm in an integrated interaction network to mine expression data for modules of genes associated with cancer outcomes. First, metabolic, signaling and protein interactions from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [35] and the Human Protein Reaction Database (HPRD) [31] are used to construct a network of biological interactions. I then calculate edge weights based on expression data from three public datasets with multiple cancer outcomes: breast cancer, hepatocellular carcinoma and colorectal adenoma. The *Walktrap* random walk algorithm is applied in this network to discover modules of closely interconnected genes and build communities using distances derived from random walk process. Finally, each community is evaluated for significance by its module score. These methods are summarized in Figure 10.

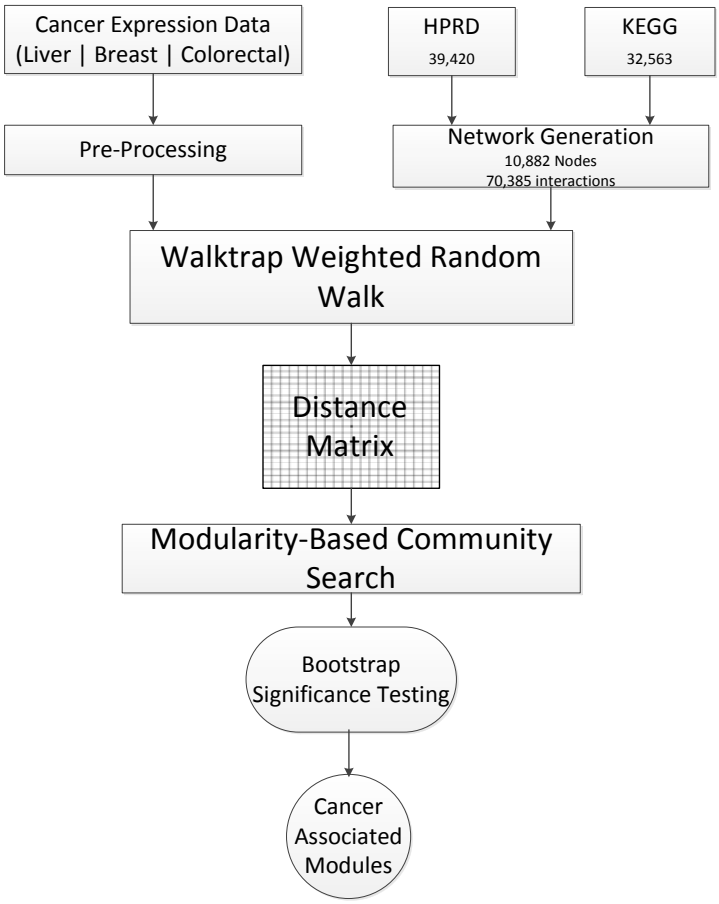


Figure 10: Flow Diagram of Network-Based Expression Analysis. Three cancer datasets from GEO and interactions from HPRD and KEGG are integrated in a weighted interaction network. The Walktrap random walk builds modules based on transition probabilities generated from the random walk process. The modules are assessed for their significance compared to a random distribution of expression values per module.

4.2.2: Gene Expression Data

Three cancer datasets were downloaded from the Gene Expression Omnibus (GEO)[181] covering onset of breast cancer (BC) prognosis, hepatocellular carcinoma (HCC), and adenoma development in colorectal cancer (CCA). GSE14520 is a hepatocellular cancer study from Roessler et al.[8], consisting of 22 paired tumor and non-tumor expression profiles using the Affymetrix HG-U133A 2.0 array. Desmedt et al.[9] published an expression dataset consisting of 198 samples to independently validate a 76-gene prognostic breast cancer signature as part of the TRANSBIG project (GSE7390). A total of 198 profiles from lymph node-negative patients (N-) were analyzed on the Affymetrix HG-U133A array, and each profile was associated with the Adjuvant!Online clinical risk index, identifying patients at

high risk for distant metastasis (good = 47, poor = 151). Sebates-Bellver[10] obtained tissue from sporadic colonic adenomas and normal mucosa of 32 colonoscopy patients and analyzed expression profiles using Affymetrix HG-U133A 2.0 arrays (GSE8671). Normal tissue was compared to colonic adenoma cancer precursor tissues. These data are summarized in Table 2. Normalized, log-transformed fold change values and p-values are calculated for each data set. P-values were corrected for multiple testing using the Benjamini and Hochberg false discovery rate[182] . All analyses were performed in R using Bioconductor [53].

Table 2: Description of Cancer Expression Data

GEO Accession	Reference	Clinical Outcome	Cases	Controls
GSE14520	Roessler 2010	Hepatocellular carcinoma tumors (HCC)	22 hepatocellular tumors	22 paired non-tumor
GSE7390	Desmedt 2007	Risk of early distant breast cancer metastasis (BC)	198 breast tumors from lymph-node negative patients	Prognosis scores for each sample
GSE8671	Sebates-Bellver 2007	Colorectal cancer adenomas (CCA)	32 paired sporadic adenoma	32 paired normal

4.2.3: Network Construction

The interactome for this study was built by extracting human interactions from KEGG and HPRD. KEGG relations were parsed from KGML files, representing 32,563 unique interactions. Metabolic reactions were defined as a relation between two neighboring enzymes that share a common metabolite; signaling reactions were defined as two genes that participate in a signaling cascade where both genes share a reaction event. A total of 39,240 protein-protein interactions were downloaded from HPRD. Duplicate nodes and edges were removed and the provenance of each interaction was saved as an edge attribute. The resulting global interaction network consisted of 10,882 nodes and 70,385 interactions. The largest connected cluster of unique pairwise interactions consisting of 10,642 nodes and 62,407 interactions was extracted for further analysis.

4.2.4: Weights and Significance Scoring

To determine edge weights in the interactome I used an average of the absolute fold change values of the two adjacent nodes. Compared to the use of p-values, I found fold change measures to be more robust weight factors as they had a more discrete range of values and a stable dispersion. This average weighting scheme was considered best suited to the random walk as it allows for more descriptive probabilities than weighting schemes using, for example, maximum or minimum values. Further, this scheme improves community cohesiveness in settings where indirect interactions may be correlated, but an intermediate interaction is not.

The magnitude of expression signal for each module was compared to a random distribution. Module weight was calculated by taking an average of the node weights; each node corresponds to a squared transformation of the maximum fold change for probes corresponding to each gene symbol. Higher-confidence modules with greater than three nodes are tested for significance. A module score is then calculated by comparing the significance of the module weight to a distribution of 5000 random samples of expression values for each module size. Code for scoring and significance testing of modules is described in Appendix A.

4.2.5: Definition of Cancer Genes

A gold standard reference list to label cancer genes is derived from evidence in OMIM. To evaluate the ability of these methods to identify cancer-related genes and interactions in significant modules, a list of cancer-related genes was created from OMIM, using text string matching and manual curation (Supplementary Files). I queried 6995 gene references including all genes in the clusters assessed, for cancer-related terms. Each matching record was reviewed to confirm that the gene was a tumor suppressor, oncogene, or otherwise shown to be significantly associated with cancer (i.e. by differential expression data, functional pathway analysis, genomic mapping, or SNP studies). The resulting list consisted of 1239 cancer-associated genes. Approximately 5% of genes did not have corresponding records in OMIM. In such cases, the gene was labeled as non-cancer because of lack of data and was included with non-cancer gene class in subsequent analyses

4.2.6: Community Analysis

Random walks have been shown to be valuable when applied to study genomic data in biological networks[112,167]. The random walk algorithm implemented here was chosen because it incorporates the topology of the network to calculate distance metrics, and optimizes the community search componenet by using the graph theoretic concept of modularity. Details of the random walk are described in section 2.5.

I implement stopping criteria to search for the optimal number of merge steps. The merge process is complete when one of the following conditions is met: 1) maximum size, 2) maximum modularity or, 3) maximum module score (Section 2.5). I tested a subset of larger maximum sizes between 250 and 500 which generally yielded in modules that were too general in terms of their functional annotation and therefore not as informative, and thus I chose a maximum size of 200 nodes as an upper bound to maintain interpretability. Community analysis code is presented in Appendix A.

4.3: Results and Discussion

4.3.1: Functional Annotation

Functional annotation of significant modules is assessed using ConsensusPathDB[55]. I queried genes in the top-scoring modules for over-representation analysis comparing against pathway gene sets (including: KEGG, WikiPathways [183], PID [34] and Reactome [30]), and a minimum overlap of two genes with the input gene list and the consensus pathway. Results were filtered using a default p-value of .01. Canonical cancer pathways and pathways associated with hallmarks of cancer are enriched in each cancer dataset (BC, HCC and CCA): cell-cycle control, DNA replication/repair, cellular adhesion/migration, apoptosis, angiogenesis, evasion of the immune response and immortality. A summary of statistics and a sample of representative pathways for the top scoring modules are presented in Table 3.

BC modules are highly enriched with cell cycle control, growth signaling, focal adhesion and angiogenesis control genes. A number of BC modules are also annotated with progesterone, estrogen and steroid hormone signaling; and levels of these hormones are

known to correlate with BC risk. In HCC, cytochrome P450, UBR, HSD detoxifying pathways and fatty acid metabolism are among the most enriched pathways. Inflammation and deregulation of liver-related detoxifying pathways are frequent markers of carcinogenic toxicity, oxidative stress and tumorogenesis. Chronic inflammation and the immune response are associated with adenoma formation in the colon; several related pathways are over-represented in CCA, including: chemokine, cytokine, T-cell receptor, fatty acid metabolism, and intestinal immunity. *Wnt* signaling is a key pathway in early stages of colorectal cancer and is enriched in CCA modules. Amino acid synthesis and metabolism pathways, associated with stability of DNA replication and repair, are over-represented across all three cancer types, although most notably in HCC. These pathways are also among the cancer-related processes highlighted in significant modules in Chapter 3.

Table 3: Functional Overview of Top Scoring Modules

Breast Cancer			
ID	**Score**	**Size**	**Key Functional Annotation**
134	40.20	16	DNA REPLICATION, ATR SIGNALING, CELL CYCLE, SYNTHESIS OF DNA, UNWINDING OF DNA
82	27.77	32	VEGF AND VEGFR SIGNALING, FOCAL ADHESION, CYTOKINE RECEPTOR INTERACTIONS, MTOR SIGNALING, PI3K CASCADE, ERBB SIGNALING, IRS SIGNALING, ANGIOGENESIS, FGFR SIGNALING, GLYPICAN1 NETWORK, SYNDECAN SIGNALING, IGF1 PATHWAY, ARF6 SIGNALING
226	21.26	16	NUCLEAR ESTROGEN RECEPTOR ALPHA NETWORK, REGULATION OF ANDROGEN RECEPTOR
224	19.08	27	METABOLISM OF NUCLEOTIDES, DNA REPLICATION, APOPTOSIS PATHWAY, ARF6 PATHWAY, CAM PATHWAY, TELOMERES EXTENSTION, PLC-G1 SIGNALING, GLUCAGON SIGNALING, C-MYC TRANSCRIPTION, GNRH SIGNALING, ERBB2 SIGNALING, EGFR SIGNALING IN CANCER
79	16.08	24	JAK-STAT SIGNALING, INTERFERON SIGNALING, CYTOKINE SIGNALING, GROWTH HORMONE RECEPTOR SIGNALING, LEPTIN SIGNALING, INSULIN SIGNALING, PROLACTIN SIGNALING, SIGNALING BY INTERLEUKINS, SHP2 SIGNALING, ERBB2 IN SIGNAL TRANSDUCTION AND ONCOLOGY, EPO SIGNALING, CD40/CD40L SIGNALING, EGFR SIGNALING, KIT SIGNALING
395	15.32	29	G ALPHA SIGNALING, GPCR SIGNALING, METABOLISM OF NUCLEOTIDES, CAM PATHWAY, SIGNALING BY ERBB2, SIGNALING BY EGFR IN CANCER, GROWTH FACTOR SIGNALING
182	14.59	12	FOXM1 TRANSCRIPTION, PROGESTERONE-MEDIATED OOCYTE MATURATION,
96	13.74	13	REELIN SIGNALING, GLYCOGEN METABOLISM, SIGNALING BY INTERLEUKINS, WNT SIGNALING, PHOSPHOINOSITIDE TARGETS, IFN-GAMMA PATHWAY, REGULATION OF MICROTUBULE CYTOSKELETON, TGF-BETA SIGNALING, KIT SIGNALING, SEMAPHORIN INTERACTIONS
321	10.99	5	VITAMIN A AND CAROTENOID METABOLISM, CYTOCHROME P450
145	10.97	11	CELL CYCLE, DNA DAMAGE RESPONSE, P53 SIGNALING, P38 MAPK SIGNALING, SONIC HEDGEHOG RECEPTOR, EFP CONTROLS CELL CYCLE AND BREAST TUMORS GROWTH, TGF BETA SIGNALING, INTEGRATED BREAST CANCER PATHWAY, MAPK SIGNALING, FOXM1 TRANSCRIPTION, AMPK SIGNALING
165	10.90	55	NUCLEAR ESTROGEN RECEPTOR NETWORK, ATF-2 TRANSCRIPTION, RETINOIC ACID RECEPTORS-MEDIATED SIGNALING, SIGNALING MEDIATED BY P38-ALPHA AND P38-BETA, FOXA1 TRANSCRIPTION
122	9.28	16	BCR SIGNALING, TCR SIGNALING, NATURAL KILLER CELL CYTOTOXICITY, FC EPSILON SIGNALING, PI3K SIGNALING, JNK SIGNALING, NF-KAPPA B SIGNALING, INTERLEUKIN SIGNALING, EPO SIGNALING, CDC42 REGULATION, EGF-EGFR SIGNALING, RAC1 REGULATION , REGULATION OF RHOA
143	8.97	11	SKP2 DEGRADATION OF P27/P21, FOXM1 TRANSCRIPTION, P73 TRANSCRIPTION, PRL SIGNALING, ATR SIGNALING, P53 PATHWAY, RB TUMOR SUPPRESSOR/CHECKPOINT, EFP CONTROLS CELL CYCLE/ BREAST TUMOR GROWTH, AKT SIGNALING, AHR PATHWAY, NOTCH SIGNALING, ERBB SIGNALING, PI3K CASCADE, AMPK SIGNALING, C-MYC TRANSCRIPTIONAL REPRESSION, SMAD2/3 SIGNALING
205	8.71	15	DNA DAMAGE RESPONSE, CELL CYCLE, INTEGRATED BREAST CANCER PATHWAY, WNT SIGNALING, AURORA A SIGNALING, LKB1 SIGNALING, C-MYC TRANSCRIPTION REGULATION, BARD1 SIGNALING, ATM PATHWAY, PLK3 SIGNALING, HEDGEHOG SIGNALING, ERBB SIGNALING, P53 PATHWAY, HTERT TRANSCRIPTIONAL REGULATION, VEGFR1/ VEGFR2 SIGNALING, AP-1 TRANSCRIPTION, E2F TRANSCRIPTION, BRCA1 BRCA2 AND ATR IN CANCER, ARF INHIBITS BIOGENESIS, NUCLEAR ESTROGEN RECEPTOR ALPHA NETWORK, AMPK SIGNALING
89	8.54	7	REGULATION OF IGF ACTIVITY BY INSULIN-LIKE GROWTH FACTOR BINDING PROTEINS
189	8.25	7	C-MYB TRANSCRIPTION, TRANSCRIPTIONAL MISREGULATION IN CANCER, AP-1 TRANSCRIPTION
348	8.20	29	REGULATION OF ACTIN CYTOSKELETON, SHC CASCADE, FGFR SIGNALING, MAPK SIGNALING, PHOSPHOLIPASE C CASCADE, PI3K CASCADE, IRS SIGNALING, INSULIN SIGNALING, SYNDECAN SIGNALING, ERBB SIGNALING, FOCAL ADHESION, ANGIOGENESIS
173	8.18	6	METABOLISM OF NUCLEOTIDES, DRUG METABOLISM, E2F TRANSCRIPTION
99	7.47	7	P38 SIGNALING MEDIATED BY MAPKAP KINASES, CELL CYCLE, INSULIN-MEDIATED GLUCOSE TRANSPORT, PI3K SIGNALING MEDIATED BY AKT, INTEGRIN SIGNALING, MTOR SIGNALING, BETA CATENIN SIGNALING, ERBB1 SIGNALING, PDGFR-BETA SIGNALING, SIGNALING BY HIPPO
12	7.25	23	MAPK SIGNALING, ATF-2 TRANSCRIPTION, REGULATION OF P38-ALPHA AND P38-BETA, TOLL LIKE RECEPTOR CASCADE, ERBB1 SIGNALING, NGF SIGNALING, RAS SIGNALING
Hepatocellular carcinoma			
408	72.64	24	DRUG METABOLISM - CYTOCHROME P450, METABOLISM OF AMINO ACIDS, FATTY ACID METABOLISM GLYCOLYSIS/GLUCONEOGENESIS, ETHANOL OXIDATION, ARACHIDONIC ACID METABOLISM, TAMOXIFEN METABOLISM, VITAMIN A/CAROTENOID METABOLISM, ESTROGEN METABOLISM, AHR PATHWAY
10	34.22	49	DRUG METABOLISM, STEROID HORMONE BIOSYNTHESIS, RETINOL METABOLISM, CYTOCHROME P450 METABOLISM, METABOLISM OF AMINO ACIDS, TAMOXIFEN METABOLISM, FATTY ACID OXIDATION, BENZO(A)PYRENE METABOLISM, AHR PATHWAY, AFLATOXIN B1 METABOLISM, IL-10 SIGNALING

513	22.92	4	ALTERNATIVE COMPLEMENT PATHWAY, COMPLEMENT AND COAGULATION CASCADES
579	19.55	13	METABOLISM OF STEROID HORMONES AND VITAMINS A AND D, METABOLISM OF LIPIDS AND LIPOPROTEINS, MINERALOCORTICOID BIOSYNTHESIS, GLUCOCORTICOID METABOLISM
603	17.58	6	METABOLISM OF AMINO ACIDS
31	14.24	14	PPAR SIGNALING, FATTY ACID, TRIACYLGLYCEROL, AND KETONE BODY METABOLISM, ADIPOCYTOKINE SIGNALING, METABOLISM OF LIPIDS AND LIPOPROTEINS, AMPK SIGNALING
97	13.93	5	ONE CARBON POOL BY FOLATE, METABOLISM OF AMINO ACIDS AND DERIVATIVES
361	9.55	16	DNA REPLICATION, CELL CYCLE, UNWINDING OF DNA, SYNTHESIS OF DNA
314	9.47	10	FATTY ACID METABOLISM, GLYCEROLIPID METABOLISM, METABOLISM OF AMINO ACIDS
34	9.08	14	TOLL-LIKE RECEPTOR SIGNALING, HTLV-I INFECTION, ACTIVATION OF AP-1 TRANSCRIPTION FACTORS, MAPK SIGNALING, TWEAK SIGNALING, TGF BETA SIGNALING, INTERLEUKIN SIGNALING, RIG-I-LIKE RECEPTOR SIGNALING, HEPATITIS B VIRUS, IGF-1 SIGNALING, HEPATOCYTE GROWTH FACTOR RECEPTOR SIGNALING, JAK-STAT SIGNALING, FAS PATHWAY
598	8.94	4	KEAP1-NRF2 PATHWAY, METABOLISM OF AMINO ACIDS AND DERIVATIVES
360	8.73	7	INSULIN SIGNALING, GLYCOGEN METABOLISM, GLUCOSE METABOLISM, CARBOHYDRATE METABOLISM
112	8.65	5	MRNA SPLICING, MRNA PROCESSING
515	8.46	10	ONE CARBON POOL BY FOLATE, FOLATE METABOLISM
257	8.23	5	UREA CYCLE AND METABOLISM OF AMINO GROUPS, METABOLISM OF AMINO ACIDS
153	7.29	5	GLUCOCORTICOID & MINERALCORTICOID METABOLISM, METABOLISM OF STEROID HORMONES & VITA/D, METABOLISM OF LIPIDS & LIPOPROTEINS, PROSTAGLANDIN SYNTHESIS/ REGULATION
123	7.22	7	ONE CARBON FOLATE METABOLISM, METHYLATION, METABOLISM OF AMINO ACIDS
254	7.03	6	METABOLISM OF NUCLEOTIDES, METABOLISM OF AMINO ACIDS AND DERIVATIVES
429	7.02	9	SIGNAL TRANSDUCTION BY L1, MTOR SIGNALING, RSK ACTIVATION, PROSTATE CANCER, L1CAM INTERACTIONS, CREB PHOSPHORYLATION THROUGH THE ACTIVATION OF RAS, MAPK SIGNALING
414	6.50	35	MAPK SIGNALING, ATF-2 TRANSCRIPTION, CELL SIGNALING IN H. PYLORI INFECTION, ACTIVATION OF AP-1 TRANSCRIPTION FACTORS, FC EPSILON RI SIGNALING, NOD1/2 SIGNALING, GNRH SIGNALING, JNK SIGNALING, CD40/CD40L SIGNALING, C RIG-I-LIKE RECEPTOR SIGNALING, TGF BETA SIGNALING, VEGF SIGNALING, EGF-EGFR SIGNALING, FOSB GENE EXPRESSION
Colorectal adenoma			
257	33.48	50	CHEMOKINE SIGNALING, GPCR SIGNALING, NF-KAPPA B SIGNALING, CXCR3 SIGNALING, TOLL-LIKE RECEPTOR SIGNALING, NOD-LIKE RECEPTOR SIGNALING, INTESTINAL IMMUNE NETWORK FOR IGA PRODUCTION, INTERLEUKIN SIGNALING, CELL SIGNALING IN H.PYLORI INFECTION
182	21.57	25	TIGHT JUNCTION INTERACTIONS, TRANSENDOTHELIAL MIGRATION, CELL-CELL COMMUNICATION, CAMS
158	18.94	9	P75(NTR) SIGNALING, DEGRADATION OF THE ECM, ECM ORGANIZATION, SYNDECAN SIGNALING
770	12.58	8	ETHANOL OXIDATION, METABOLISM BY CYTOCHROME P450, TYROSINE METABOLISM, FATTY ACID METABOLISM, GLYCOLYSIS/GLUCONEOGENESIS, VITAMIN A/CAROTENOID METABOLISM
14	11.51	5	C-MYC TRANSCRIPTIONAL REPRESSION, SMAD2/3 SIGNALING, CELL CYCLE, PATHWAYS IN CANCER
452	8.75	10	GLYCOSPHINGOLIPID BIOSYNTHESIS, GLYCOSAMINOGLYCAN BIOSYNTHESIS
487	7.16	28	MAPK SIGNALING, ATF-2 TRANSCRIPTION, ACTIVATION OF AP-1 TRANSCRIPTION FACTORS, NOD-LIKE RECEPTOR SIGNALING, FC EPSILON SIGNALING, GNRH SIGNALING, TOLL-LIKE RECEPTOR SIGNALING, INTERLEUKIN SIGNALING, TGF BETA SIGNALING, VEGF SIGNALING, EGF-EGFR SIGNALING, KIT SIGNALING, RANKL-RANK SIGNALING, COLORECTAL CANCER, S1P2 PATHWAY, NONCANONICAL WNT SIGNALING, ARF6 PATHWAY, ERBB SIGNALING, TBXA2R SIGNALING
301	7.06	7	TRANSCRIPTIONAL MISREGULATION IN CANCER, RB REGULATION, INTERLEUKIN SIGNALING, C-MYB TRANSCRIPTION, INTERFERON SIGNALING, FOXA2/FOXA3 TRANSCRIPTIONS, SMAD2/3 SIGNALING
758	6.91	5	METABOLISM OF AMINO ACIDS AND DERIVATIVES
762	6.59	12	WNT SIGNALING, SECRETIN FAMILY OF RECEPTORS, HTLV-I INFECTION, SIGNALING BY GPCR
240	6.59	28	G PROTEIN SIGNALING, CAM PATHWAY, PLC-GAMMA1 SIGNALING, NUCLEOTIDE METABOLISM, SIGNALING BY ERBB2, SIGNALING BY EGFR, SIGNALING BY FGFR, SIGNALING BY PDGF
757	6.53	12	METABOLISM OF STEROID HORMONES AND VITA/D, METABOLISM OF LIPIDS AND LIPOPROTEINS, GLUCOCORTICOID & MINERALCORTICOID METABOLISM, BILE ACID AND BILE SALT METABOLISM
410	6.49	6	JAK-STAT SIGNALING, CYTOKINE-CYTOKINE RECEPTOR INTERACTION, SHP2 SIGNALING, INTERLEUKIN SIGNALING, ROLE OF ERBB2 IN SIGNAL TRANSDUCTION AND ONCOLOGY
412	6.21	9	DNA REPLICATION, CELL CYCLE, UNWINDING OF DNA, ATR SIGNALING, E2F TRANSCRIPTION

345	6.13	14	NEUROTROPHIN SIGNALING, GNRH SIGNALING, CREB PHOSPHORYLATION, PKA ACTIVATION, CAM PATHWAY, INSULIN SIGNALING, PGC-1A REGULATION, RAS REGULATION, SMAD2/3 SIGNALING
267	6.06	6	METABOLISM OF PROTEINS
334	6.04	12	BETA-CATENIN PHOSPHORYLATION CASCADE, SIGNALING BY WNT, GLYCOGEN METABOLISM, PLATELET HOMEOSTASIS, DNA REPLICATION, CELL CYCLE, DNA DAMAGE RESPONSE
111	5.74	11	ECM-RECEPTOR INTERACTION, FOCAL ADHESION, INTEGRIN INTERACTIONS, NCAM INTERACTIONS, SYNDECAN SIGNALING, PROTHROMBIN ACTIVATION, PDGF SIGNALING, VEGFR3 SIGNALING
54	5.73	4	NONE
125	5.67	20	CHEMOKINE SIGNALING, G ALPHA SIGNALING, SIGNALING BY GPCR, ACTIVATION OF PKA, INTESTINAL IGA IMMUNE NETWORK, CELL SIGNALING IN HELICOBACTER PYLORI INFECTION
183	5.41	6	BETA-CATENIN PHOSPHORYLATION CASCADE, CTLA4 INHIBITORY SIGNALING, GLYCOGEN METABOLISM, WNT SIGNALING, DNA REPLICATION, CELL CYCLE, IMMUNE SYSTEM, DNA DAMAGE
156	5.38	4	HEMATOPOIETIC CELL LINEAGE
144	5.35	16	CELL CYCLE, P38/MAPKAP SIGNALING, LKB1 SIGNALING, INSULIN-MEDIATED GLUCOSE TRANSPORT, PI3K/AKT SIGNALING, INTEGRIN SIGNALING, FOXO FAMILY SIGNALING, MTOR SIGNALING, ERBB1 SIGNALING, PDGFR-BETA SIGNALING, ATR SIGNALING, PLK1 SIGNALING, RB TUMOR SUPPRESSOR/CHECKPOINT, RAP1 SIGNALING, INTEGRATED CANCER PATHWAY, ATM PATHWAY, SHC SIGNALING, ARMS-MEDIATED ACTIVATION, IGF1 PATHWAY, IRS SIGNALING

4.3.2: Breast Cancer

BC fold change measurements were filtered below an FDR-adjusted p-value of .01 and data associated with the remaining 2074 probes was used to weigh the network. The merge process reached a maximum size at step 2069, and the community search resulted in 8116 singletons, 206 pairs, 77 triplets, and 174 modules (module size $(3 > \text{size} \leq 200)$. The top-scoring modules are summarized in Table3 and presented in Appendix B and in high resolution as Supplementary Files. I examined the top-scoring modules in more detail by manually reviewing functional annotation and reviewing visualizations of the modules. These modules were investigated to identify target genes, interactions with known cancer gnes, and interactions between pathways.

Significant BC modules are annotated with relevant cancer-associated pathways (Table 3) and plausible cancer-related interactions. Module 143 (Figure 11) is composed of cyclins regulating the cell cycle and a link to telomere formation (*E2F5*). *SKP2* is a known oncogene and interacts with cyclins to promote cell proliferation and evade apoptosis [184]. *SKP2* and cyclin *CCNA2* both show significantly altered activity in the expression data. Both genes interact with *BRCA2* via *CDK2*. Module 79 (Figure 12) shows interactions between inflammatory markers and *JAK* which are involved in *JAK/STAT* transcription activity, cellular proliferation and differentiation. The *JAK/STAT* pathway is associated with B-cell growth and proliferation and genes in this pathway have been shown to be involved in cancer. *SOCS1*, *SOCS2*, *SOCS3* and *CBLC* mediate growth and are involved in the cytokine response. Differentially expressed genes include *SOC2*, *SOC3*, *CBLC* and *IL20RA*; and the coordinated interaction and altered expression of these genes suggest they play a concerted role in BC progression. Module 82 (Figure 13) shows interaction between a number of growth factors and receptors, including *VEGFA*, *FIGF*, *IGFIR*, *PDGFRA*, *EGFR* and the oncogene *MET* and the tumor regulator *ErbB4*. *IRS2* affects proliferation and regeneration of cells, its expression is critical during development and growth, and the gene may influence cancer survival [185,186]. Oncogene *MET* interacts with several growth factors, including *FGF7* which is involved in epithelial proliferation and may play a role in gastric cancer [186,187]. *VEGFA* is a known metastatic vascular growth marker and a therapeutic target for breast cancer survival. Both *IRS2* and *FGF7* represent interesting candidate disease genes given their key functions in cell proliferation and growth.

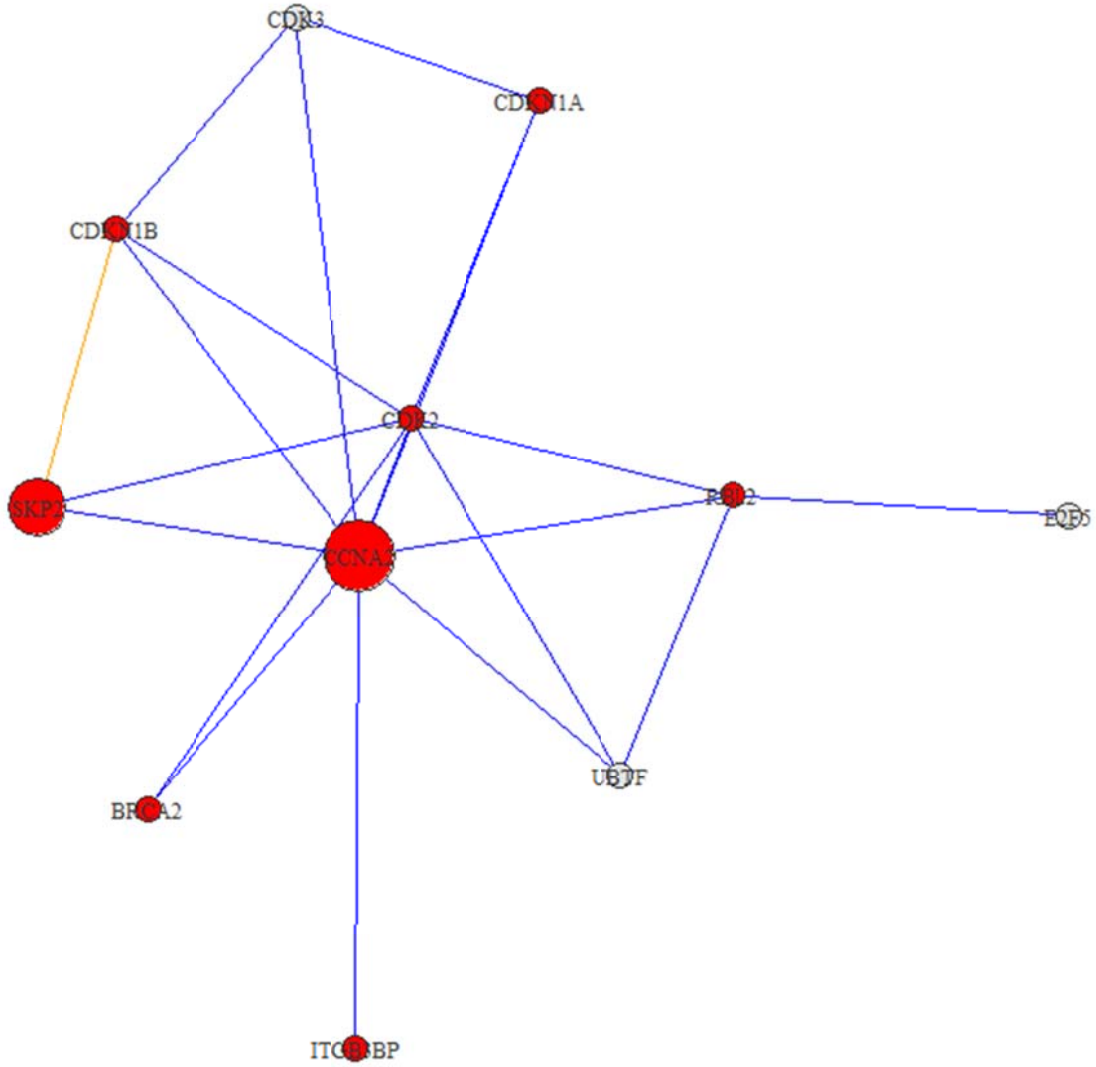


Figure 11: BC Network Module 143. Module 143 shows interactions among cyclins, *SKP2* and *BRCA2*. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, green from KEGG, and orange from both databases.

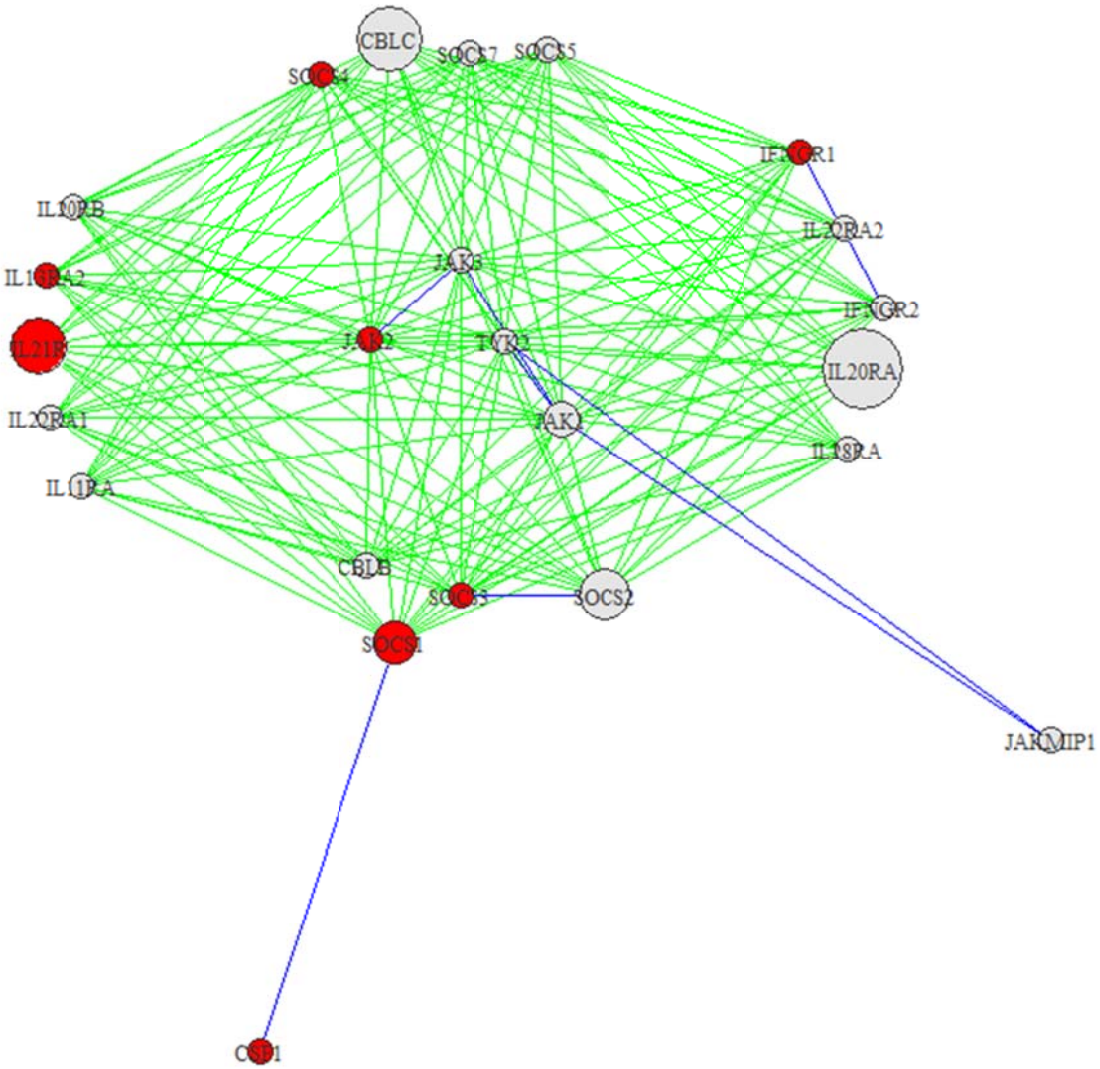


Figure 12: BC Network Module 79. The module shows interactions among cytokines, SOC genes and genes in the JAK-STAT pathway. The JAK-STAT pathway is associated with B-cell growth and proliferation and a number of genes in this pathway are related to cancer. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, and gree from KEGG.

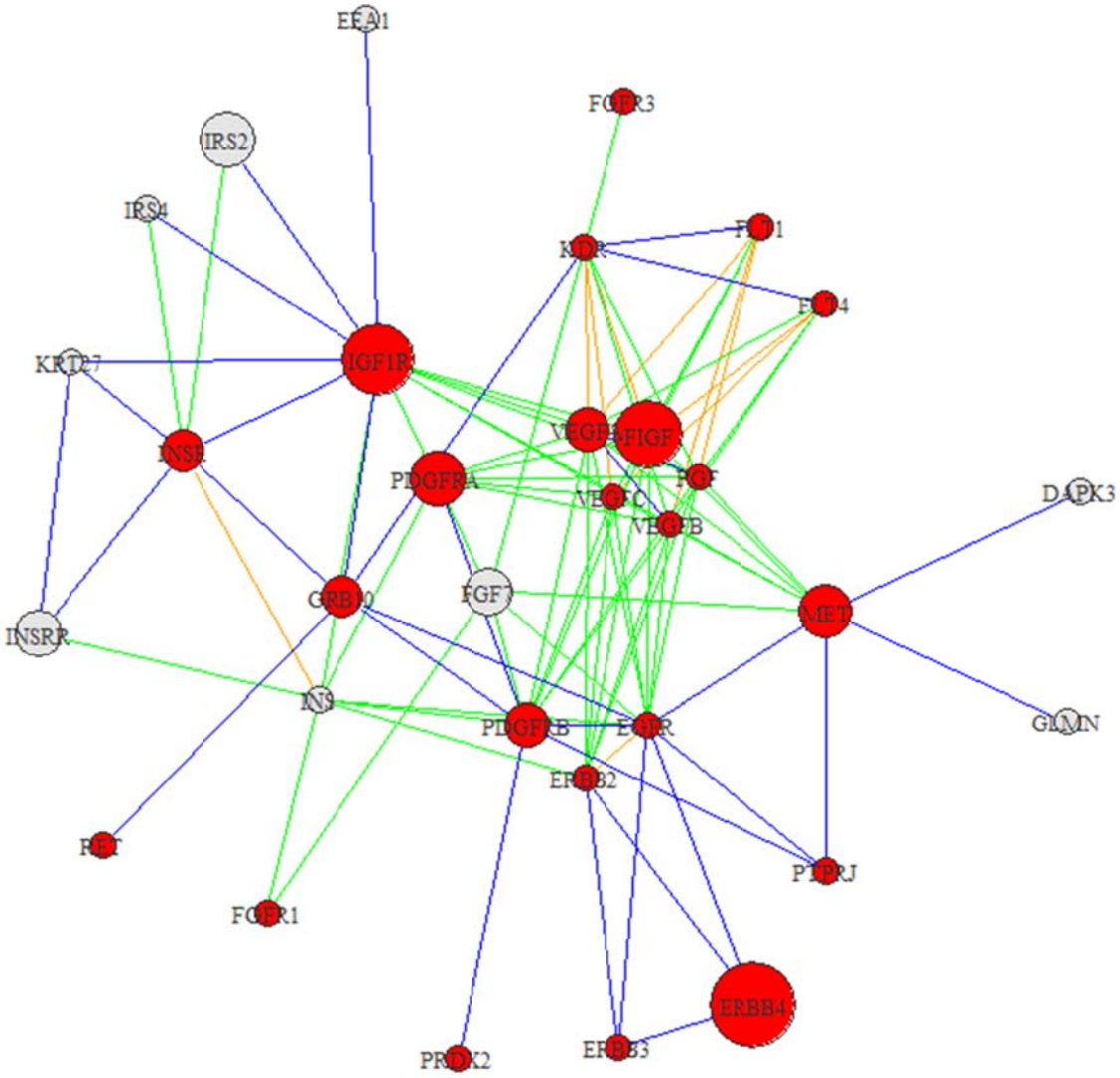


Figure 13: BC Network Module 82. Module 82 shows iinteractions among the *MET* oncogene and critical cancer-associated growth factors including *IGF1R, PDGFRA, VEGFA,* and *ERBB4.* Among genes in this module, *IRS2* and *FGF7* are differentially regulated and may be interesting targets for further research. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, green from KEGG, and orange from both databases.

Table 4: Key Genes described in BC Modules

Gene	Gene Description	Module	Function
SKP2	S-phase kinase-associated protein 2	143	Mediates the ubiquitination and subsequent proteasomal degradation of target proteins involved in cell cycle progression, signal transduction and transcription
CCNA2	cyclin A2	143	Essential for the control of the cell cycle at the G1/S (start) and the G2/M (mitosis) transitions
BRCA2	breast cancer 2, early onset	143	Involved in double-strand break repair and/or homologous recombination.
CDK2	cyclin-dependent kinase 2	143	Serine/threonine-protein kinase involved in the control of the cell cycle; essential for meiosis.
JAK1	Janus kinase 1	79	Tyrosine kinase, involved in the IFN-alpha/beta/gamma signal pathway. Kinase partner for the interleukin (IL)-2 receptor
SOCS1,-2,-3	suppressor of cytokine signaling1,-2,-3	79	SOCS family proteins form part of a classical negative feedback system that regulates cytokine signal transduction, involved in negative regulation of cytokines that signal through the JAK/STAT pathway.
IL21R	interleukin 21 receptor	79	Transduces the growth promoting signal of IL21, and is important for the proliferation and differentiation of T cells, B cells, and natural killer (NK) cells. The ligand binding of this receptor leads to the activation of multiple downstream signaling molecules, including JAK1, JAK3, STAT1, and STAT3.
CBLC	Cbl proto-oncogene	79	Regulator of EGFR mediated signal transduction
FIGF	c-fos induced growth factor (VEGF D)	82	Growth factor active in angiogenesis, lymphangiogenesis and endothelial cell growth, stimulating their proliferation and migration and also has effects on the permeability of blood vessels.
IFGIR	insulin-like growth factor 1 receptor	82	Receptor tyrosine kinase which mediates actions of insulin-like growth factor 1 (IGF1). The activated IGF1R is involved in cell growth and survival control. IGF1R is crucial for tumor transformation and survival of malignant cells.
PDGFRA	platelet-derived growth factor receptor, alpha polypeptide	82	Tyrosine-protein kinase that acts as a cell-surface receptor for PDGFA, PDGFB and PDGFC and plays an essential role in the regulation of embryonic development, cell proliferation, survival and chemotaxis.
EGFR	epidermal growth factor receptor	82	Receptor tyrosine kinase binding ligands of the EGF family and activating several signaling cascades. Binding of EFGR to a ligand induces receptor dimerization and tyrosine autophosphorylation leads to cell proliferation.
MET	met proto-oncogene (hepatocyte growth factor receptor)	82	Receptor tyrosine kinase that transduces signals from the extracellular matrix into the cytoplasm by binding to hepatocyte growth factor/HGF ligand. Regulates many physiological processes including proliferation, morphogenesis and survival.
ErbB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4	82	Tyrosine-protein kinase that plays an essential role as cell surface receptor for neuregulins and EGF family members and regulates organ development, gene transcription, cell proliferation, differentiation, migration and apoptosis.
IRS2	insulin receptor substrate 2	82	Mediates the control of various cellular processes by insulin
FGF7	fibroblast growth factor	82	Plays an important role in the regulation of embryonic development, cell proliferation and differentiation.
VEGFA	vascular endothelial growth factor A	82	This gene encodes a member of the PDGF (platelet-derived growth factor)/VEGF (vascular endothelial growth factor) family. Growth factor active in angiogenesis, vasculogenesis and endothelial cell growth. Induces endothelial cell proliferation, promotes cell migration, inhibits apoptosis and induces permeabilization of blood vessels. Binds to FLT1/VEGFR1 and KDR/VEGFR2 receptors, heparan sulfate and heparin.
INSRR	insulin receptor-related receptor	82	Receptor with tyrosine-protein kinase activity. Activates a signaling pathway that involves IRS1 and AKT1/PKB
INSR	insulin receptor	82	Binding of insulin to the insulin receptor (INSR) stimulates glucose uptake . Many tumors have altered expression of IGF1R and its ligands and this constitutes an early, possible initiating, event in tumorigenesis.
PDGFRB	platelet-derived growth factor receptor, beta polypeptide	82	Tyrosine-protein kinase that acts as cell-surface receptor for PDGFB, PDGFD and PDGFA. Plays an essential role in the regulation of embryonic development, cell proliferation, survival, differentiation, chemotaxis and migration.

4.3.3: Hepatocellular Carcinoma

HCC data included 16,360 probes after filtering by p-values. The maximal score was reached at 2393 steps, resulting in 7666 singletons, 352 pairs, 128 triplets, and 198 modules. At this step size, the maximum module size was 54 (module size $(3 > \text{size} \leq 54)$. Top scoring modules are summarize in Table 3, and presented in Appendix B and in high resolution as Supplementary Files. I reviewed modules 361, 429 and 414 (Figures 14-16) in greater detail. Module 361 consists of interactions between a family of cyclins, origin recognition complexes and minchromosome maintenance genes. These genes exhibit high differential expression and function in regulation of the cell-cycle and cellular proliferation. The series of interactions in this module have implications in cancer. Kinase activation of *CDC7,* a gene known to be highly expressed in cancer, is dependent on expression of *DBF4* [188]. *MCM5* forms a complex with *MCM2* [189], a candidate oncogene that is phosphorylated by *CDC7*. *ORC5L* associates with both *CDC7* and *MCM5* in the network and this group of genes display altered expression in HCC tissue. Module 429, includes upregulation of *IGFI* which is known to alter cancer risk [190], the oncogene *NOV*, and transcription factors *RPS6KA2* and *RPS6KA6*. These transcription factors are associated with the *RSK* family of genes, involved in activation of map kinase growth signaling, cell cycle control and differentiation and may be implicated in cancer development [191,192]. Given their importance in cell development and association with *IGFI* and *NOV*, these *RSK* transcription factors are compelling candidate genes. Module 414 shows the interaction between *MAPK* signaling genes, the *DUSP* family and well known *FOS* and *JUND* oncogenes. The *DUSP* genes are known to regulate *MAPK* signaling cascades, and a number of these *MAPK* genes are known to be involved in cancer. *RIPK2* is not well-described, but is believed to play an important role in apoptosis. *DUSP1*, *DUSP4*, *PTPRR* and *RIPK2* are also highly upregulated. By their association with known cancer genes and high differential expression, these genes are promising targets for therapeutic research.

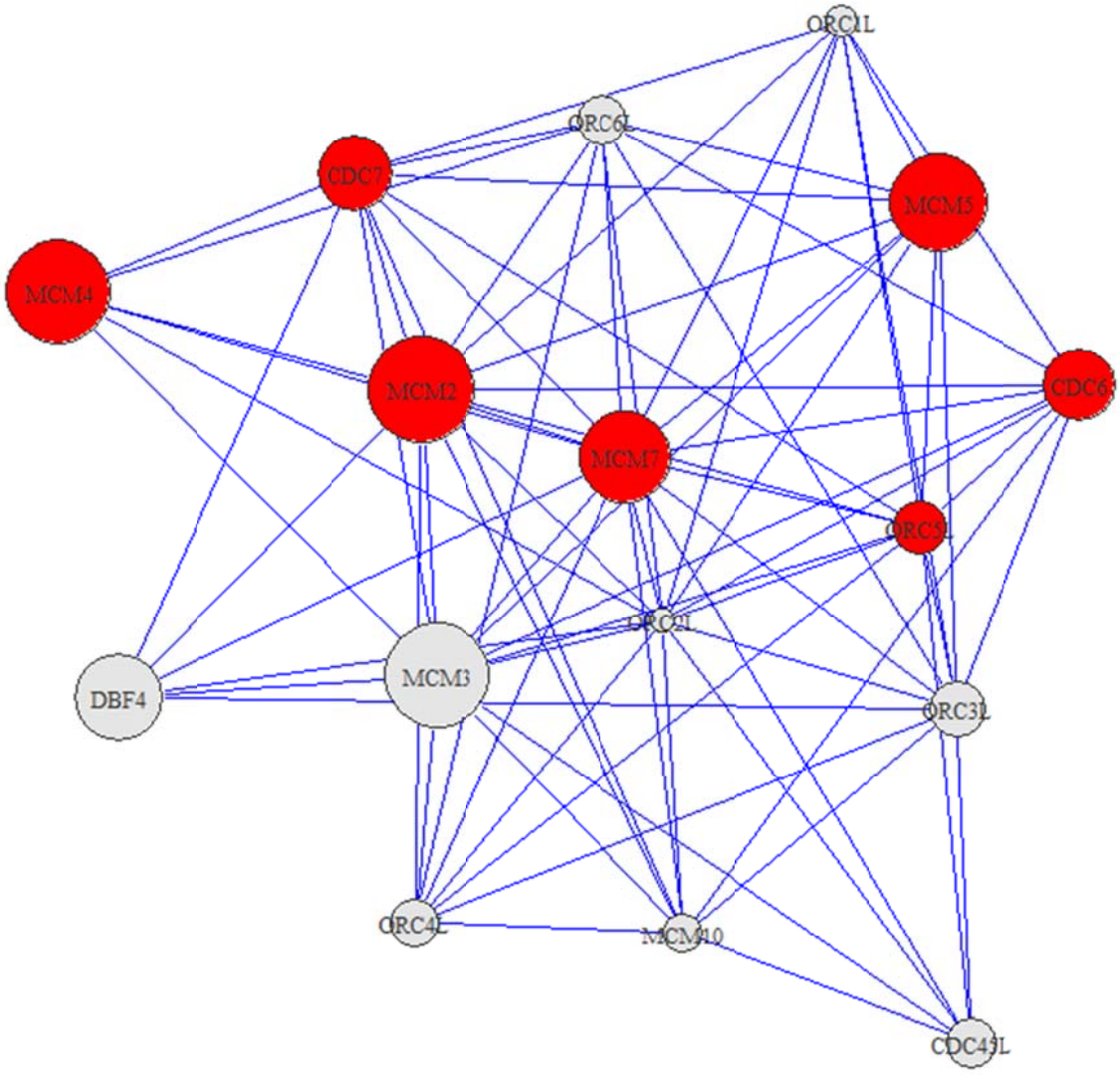


Figure 14: HCC Network Module 361. Module 361 shows interactions among *MCM*, *ORC* genes involved in cell-cycle control and *DBF4*. A number of *MC*M genes are known to be involved in cancer, and *DBF4* appears to play an interesting role in the cell cycle via interactions presented in this network and with other critical cell-cylce control genes. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, and gree from KEGG.

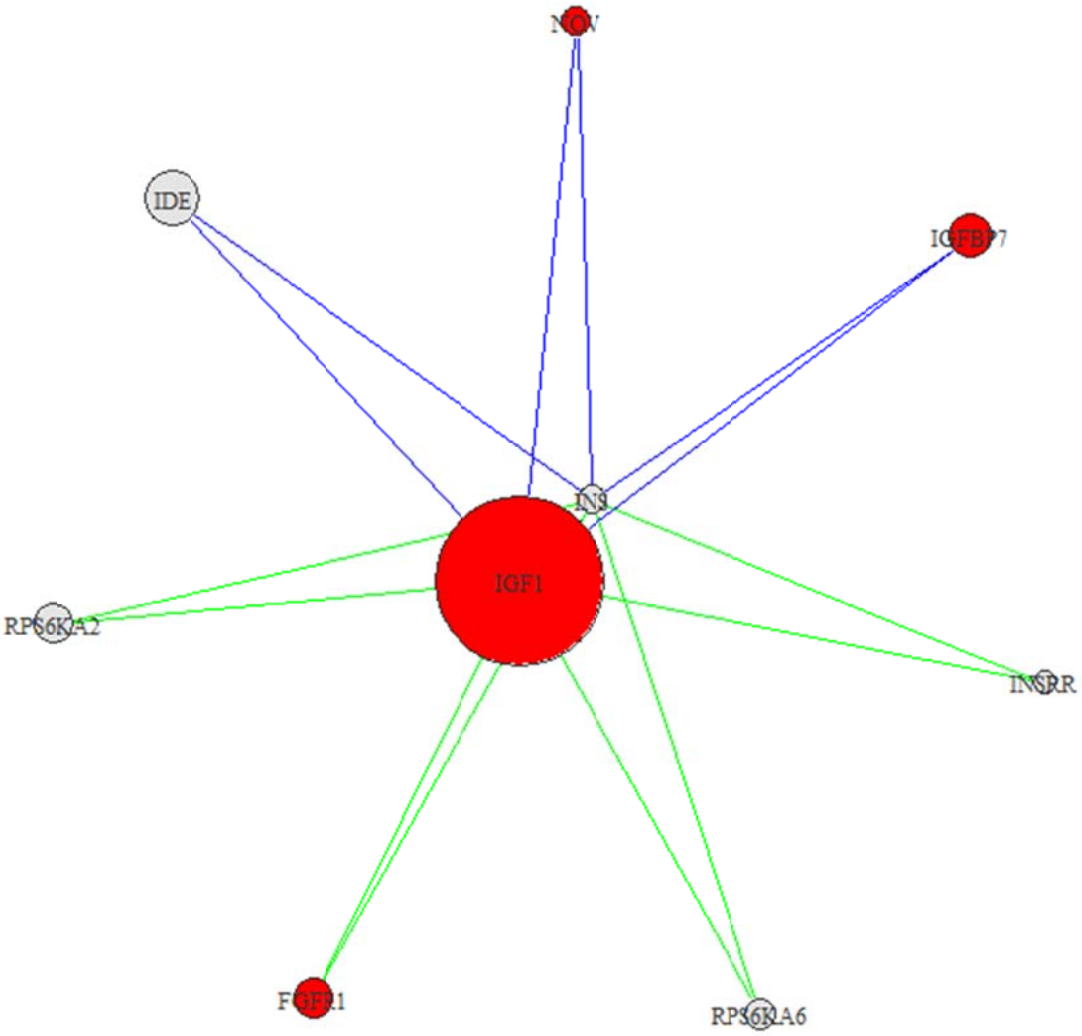


Figure 15: HCC Network Module 429. Module 429 shows interactions between *IGFI,* the *NOV* oncogene and *RPS6KA* transcription factors. The *RPS6KA* transcription factors are not well-described but seem to play in important role in cell growth and proliferation. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, and gree from KEGG.

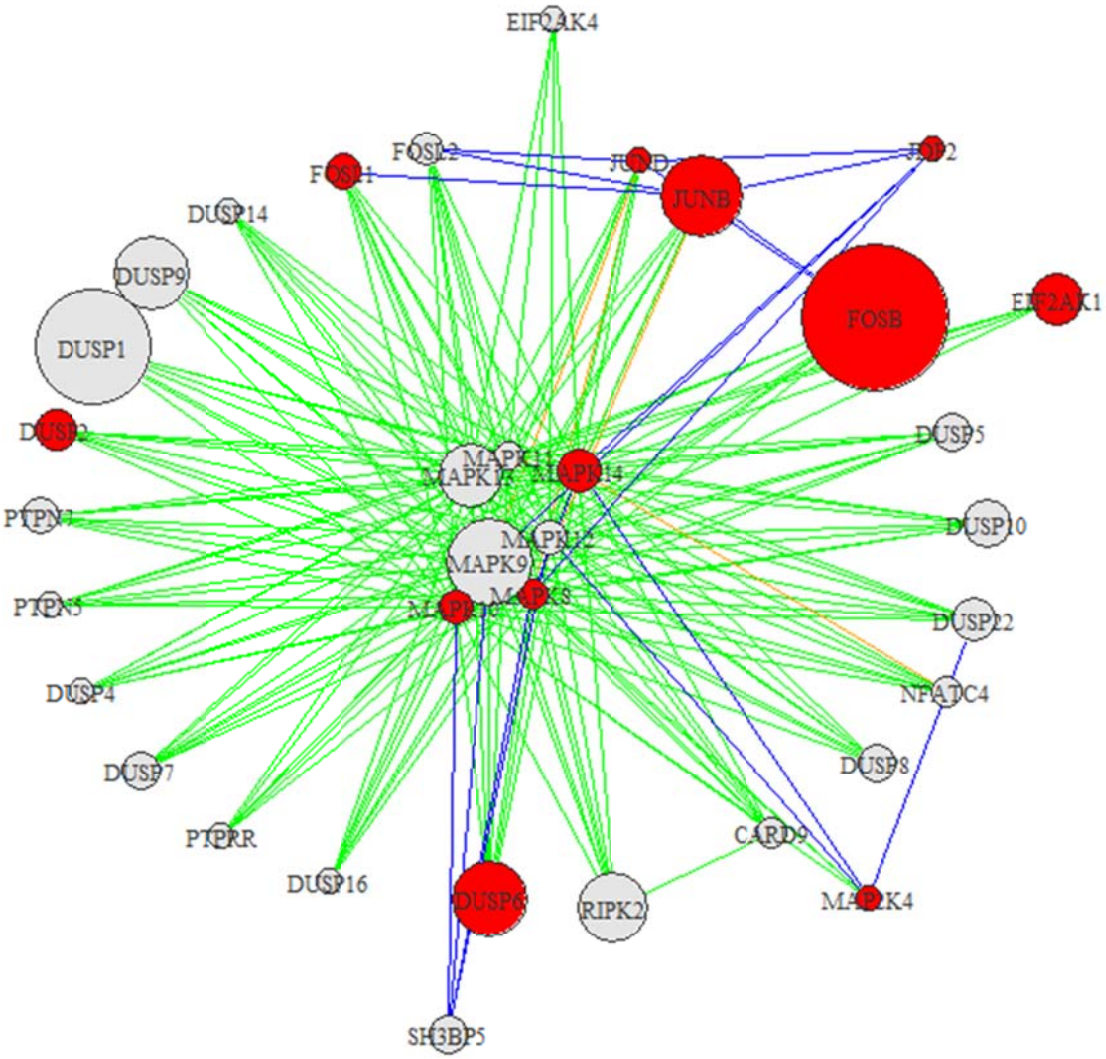


Figure 16: HCC Network Module 414. Module 414 shows interactions among *MAPK*, *DUSP* genes and *FOSB* and *JUNB* oncogenes. The DUSP family of genes is known to regulate the activity of *MAP* kinases, a number of which play a role in cancer. This module presents interactions among *MAPK* genes and the oncogene *JUNB*, protooncogene *FOSB*, and *RIPK2*. *RIPK2* is not well-described, but appears to play a role in apoptosis. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, green from KEGG, and orange from both databases.

Table 5: Key Genes described in HCC Modules

Gene	Gene Description	Module	Function
CDC7	cell division cycle 7 homolog	361	Phosphorylates substrates that regulate the G1/S phase transition and DNA replication, including MCM2 and MCM3.
DBF4	DBF4 homolog	361	Regulatory subunit for CDC7 which activates its kinase activity thereby playing a central role in DNA replication and cell proliferation. Required for progression of S phase. The complex CDC7-DBF4 selectively phosphorylates MCM2 and is then involved in regulating the initiation of DNA replication during cell cycle
ORC5L	origin recognition complex, subunit 5	361	The origin recognition complex (ORC) is a highly conserved protein complex essential for the initiation of the DNA replication in eukaryotic cells. Studies in yeast demonstrated that ORC binds specifically to origins of replication and serves as a platform for the assembly of additional initiation factors such as Cdc6 and Mcm proteins.
CDC6	cell division cycle 6 homolog	361	Involved in the initiation of DNA replication and s in checkpoint controls that ensure complete DNA replication before mitosis. Reported to be regulated in response to mitogenic signals and transcriptional control involving E2F proteins.
MCM2,-3,-4,-5,-7	minichromosome maintenance complex component 2,-3,-4,-5,-7	361	The MCM2-7 complex (MCM complex) is the putative replicative helicase essential for 'once per cell cycle' DNA replication initiation and elongation in eukaryotic cells. Required for DNA replication and cell proliferation
IGF1	insulin-like growth factor 1 (somatomedin C)	429	The insulin-like growth factors are structurally and functionally related to insulin but have a much higher growth-promoting activity.
IDE	insulin-degrading enzyme	429	Plays a role in the cellular breakdown of insulin, IAPP, glucagon, bradykinin, kallidin and other peptides, and thereby plays a role in intercellular peptide signaling.
NOV	nephroblastoma overexpressed	429	Immediate-early protein likely to play a role in cell growth regulation
IGFBP7	insulin-like growth factor binding 7	429	Binds IGF-I and IGF-II with low affinity. Stimulates prostacyclin (PGI2) production and cell adhesion.
RPS6KA2	ribosomal protein S6 kinase, 90kDa, polypeptide 2	429	Serine/threonine-protein kinase that acts downstream of ERK (MAPK1/ERK2 and MAPK3/ERK1) signaling and mediates mitogenic and stress-induced activation of transcription factors, regulates translation, and mediates cellular proliferation, survival, and differentiation. May function as tumor suppressor in epithelial ovarian cancer cells.
RPS6KA6	ribosomal protein S6 kinase, 90kDa, polypeptide 6	429	Constitutively active serine/threonine-protein kinase that exhibits growth-factor-independent kinase activity. Participates in p53/TP53-dependent cell growth arrest signaling and plays an inhibitory role during embryogenesis
DUSP1,-2,-6,-9	dual specificity phosphatase 1, -2,-6,-9	414	These phosphatases inactivate their target kinases by dephosphorylation. They negatively regulate members of the MAP-kinase superfamily (MAPK/ERK, SAPK/JNK, p38), which are associated with cellular proliferation and differentiation.
MAPK9,-10,-12,-14	mitogen-activated protein kinase 9,-10,-12,-14	414	MAP kinases act as an integration point for multiple biochemical signals, and are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation and development.
PTPRR	protein tyrosine phosphatase, receptor type, R	414	PTPs are signaling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation. Silencing of this gene has been associated with colorectal cancer. Sequesters mitogen-activated protein kinases (MAPKs) such as MAPK1, MAPK3 and MAPK14 in the cytoplasm in an inactive form.
FOSL1	FOS-like antigen, FBJ murine osteosarcoma viral oncogene B	414	Fos proteins interact with Jun proteins enhancing their DNA binding activity. FOS proteins have been implicated as regulators of cell proliferation, differentiation, and transformation.
RIPK2	receptor-interacting serine-threonine kinase 2	414	Serine/threonine/tyrosine kinase that plays an essential role in modulation of innate and adaptive immune responses. It is a potent activator of NF-kappaB and inducer of apoptosis in response to various stimuli.
SH3BP5	SH3-domain binding protein 5	414	Plays a negative regulatory role in BTK-related signaling in B-cells. May be involved in BCR-induced apoptotic cell death.
JUNB	jun B proto-oncogene	414	Transcription factor involved in regulating gene activity following the primary growth factor response.

4.3.4: Colorectal Cancer

CCA data included 21648 probes after filtering by p-values. The maximal score was reached at 2967 steps. The resulting community structure included 6879 singletons, 385 pairs, 149 triplets, and 253 modules. The maximum module size at this step was 160 (module size $(3 > \text{size} \leq 160)$. The top scoring modules are summarized in Table 3 and are presented in Appendix B and in high resolution as Supplementary Files. I reviewed functional annotation and visualized scoring modules 301, 144 and 762. There are three known oncogenes in module 301 (Figure 17): *SPI1*, *RUNX1* and *IRF4*. *CEBPB* and *CEBPE* interact with these oncogenes, affect cellular proliferation and alter tumor development and cancer risk[193,194]. Transcription factors *SPI1*and *RUNX1* participate in hematopoietic stem cell formation and can lead to the development of multiple cell lineages in cancer[195,196]. These genes show altered expression in the network, and specifically, the role of the highly differentially regulated transcription factor *SPIB* may play in colorectal cancer is an interesting area for further research. Module 144 (Figure 18) shows interaction between *CDK1*, a regulator of the cell cycle and proliferation, and genes associated with cancer: *PBK*, *HMGA2* and *FOXM1*. Putative candidates among neighboring genes include: *BRSK1*, *WEE1* and *CDC25C*, which are involved in cell-cycle checkpoints and are overexpressed in CCA. Specifically, *WEE1* and *CDC25C* are both significantly differentially regulated and are known to play a mutually antagonistic role in cell-cylce control. *BRSK1* is not well described, but exhibits key interactions with genes involved in cell-cycle control. Module 762 (Figure 19) consists of interactions among *SFRP1* and *SFRP2* genes and *FZD* genes in the *Wnt*-frizzled pathway. The *Wnt* pathway is involved in cell polarity and malignant cell transformation in colorectal cancer, and the *SFRP1* and *SFRP2*[197] genes are known to interfere with *Wnt* signaling. Given the topology of *SFRP1* and *SFRP2* as hubs in this module and their altered expression, these genes appear to play a central role in the *Wnt* pathway and CCA development.

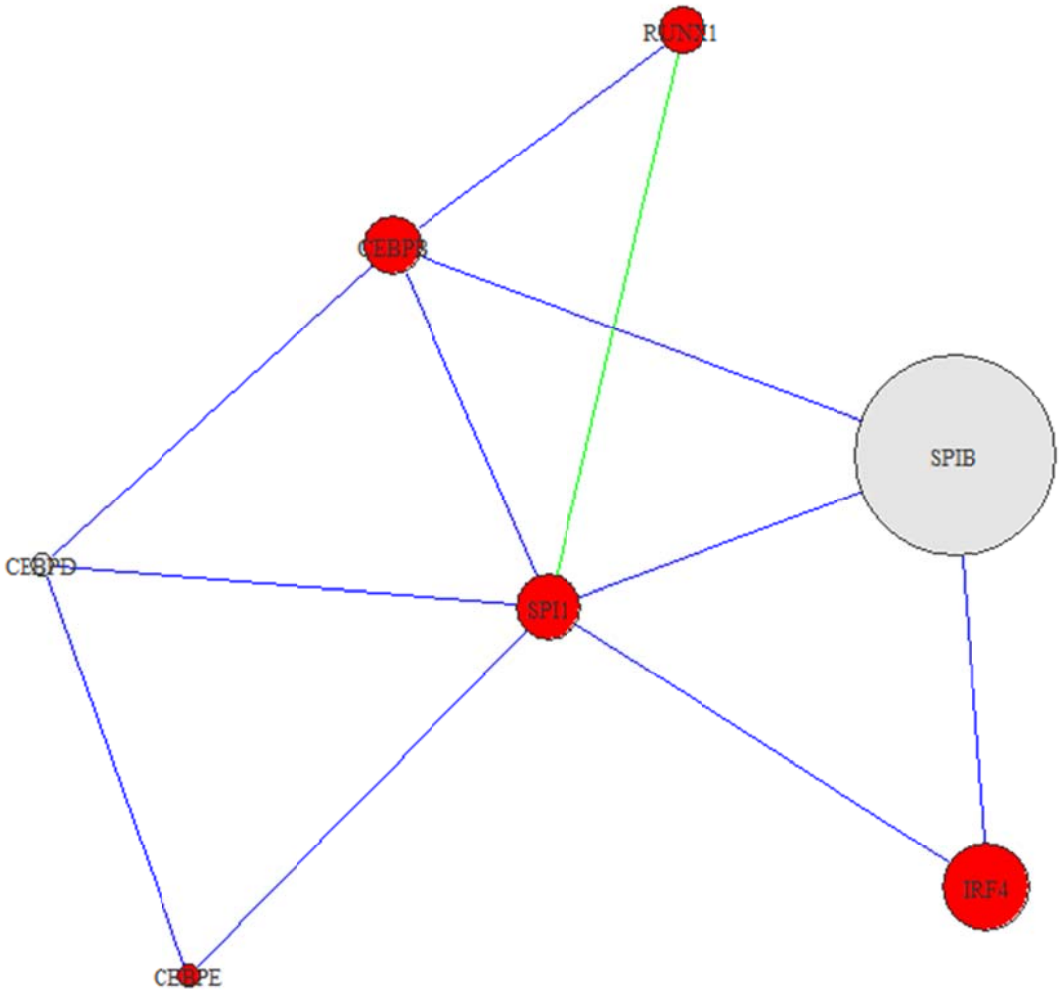


Figure 17: CCA Network Module 301. Module 301 shows interactions among cancer-related transcription factors. *SPIB* is of interest because this transcription factor is highly differentially regulated in this module and interacts closely with known cancer genes. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, and gree from KEGG.

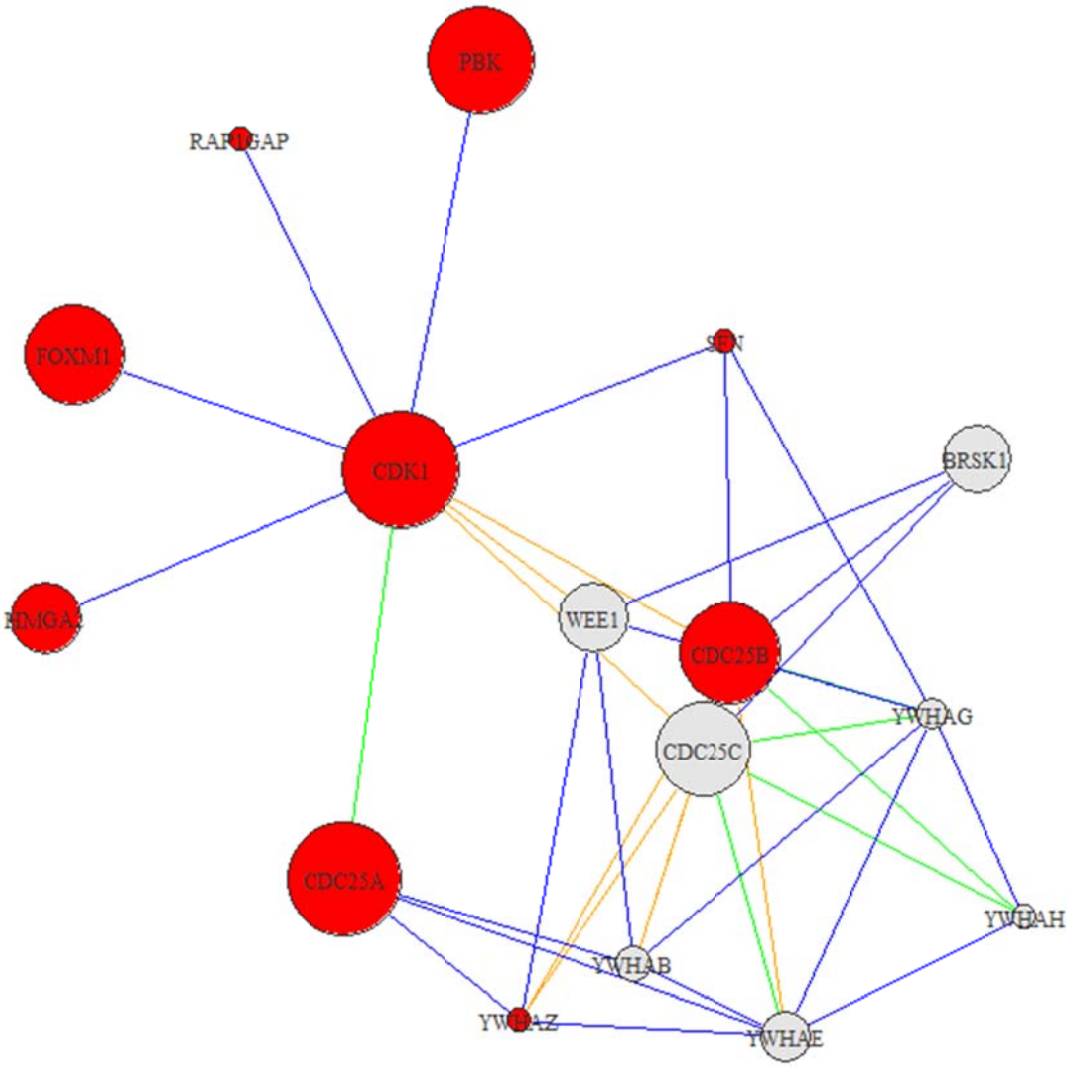


Figure 18: CCA Network Module 144. Module 144 shows interactions among cell cycle regulatory genes and *FOXM1* oncogene. *WEE1*, *CDC25C*, *YWHAE* and *BRSK1* are also involved in cell cycle control and interact closely with cancer-associated genes, but are not themselves well-described as cancer genes. Also of note, *WEE1* and *CDC25C* are known to play an antagonistic role in regulating the cell cycle. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, green from KEGG, and orange from both databases.

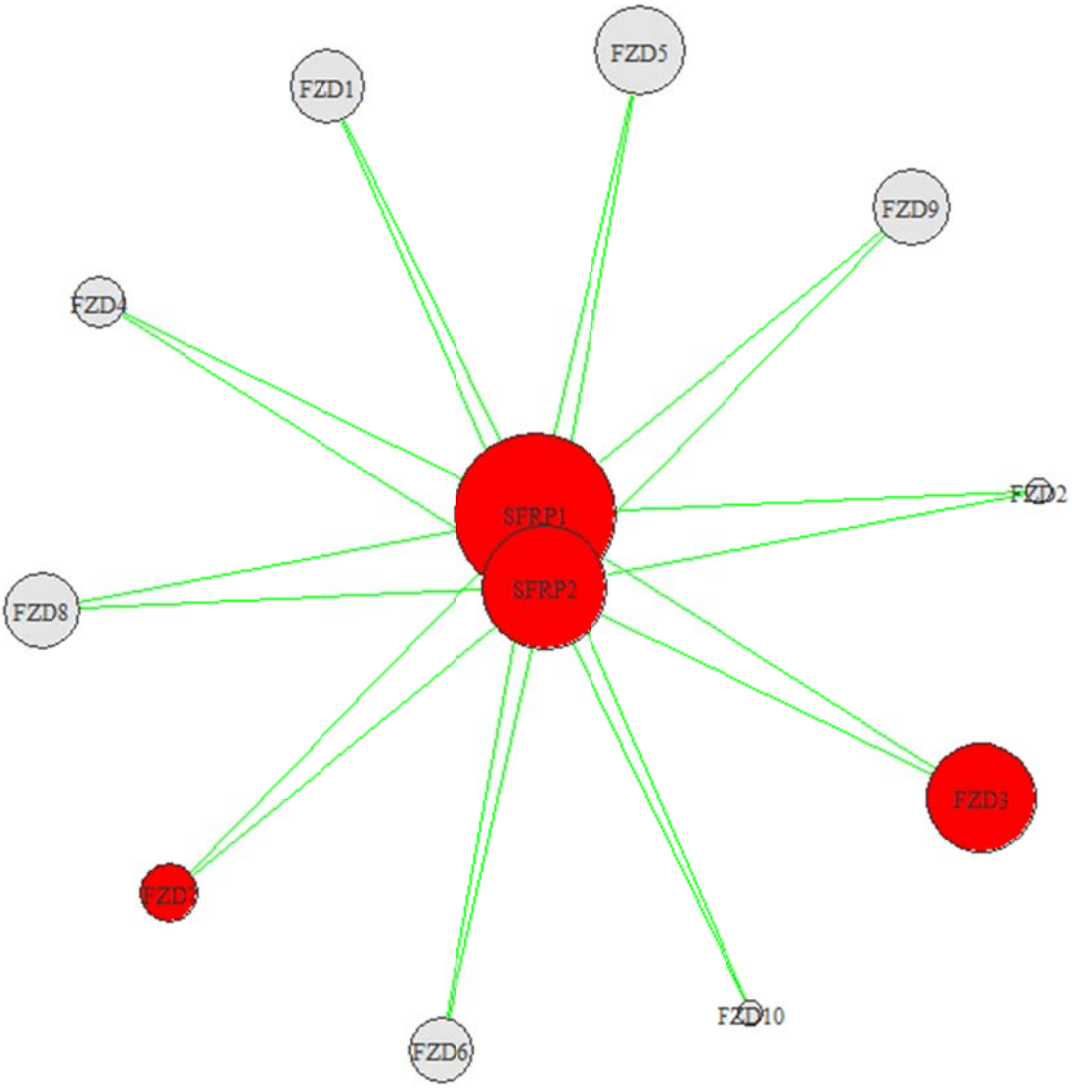


Figure 19: CCA Network Module 762. Module 762 shows interactions among *Wnt* pathway genes, including *SFRP1, SFRP2* and genes from the family of *Frizzled* genes. This module highlights specific *Wnt* interactions that show altered activity in colorectal cancer. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, green from KEGG, and orange from both databases.

Table 6: Key Genes described in CCA Modules

Gene	Gene Description	Module	Function
SPI1	spleen focus forming virus (SFFV) proviral integration oncogene	301	Binds to the PU-box, a purine-rich DNA sequence that can act as a lymphoid-specific enhancer. A transcriptional activator involved in the differentiation or activation of macrophages or B-cells. Binds RNA and modulates pre-mRNA splicing.
SPIB	Spi-B transcription factor	301	Transcriptional activator which binds to the PU-box, a purine-rich DNA sequence that can act as a lymphoid-specific enhancer. Required for B-cell receptor (BCR) signaling, necessary for normal B-cell development and antigenic stimulation
RUNX1	runt-related transcription factor 1	301	Core binding factor (CBF) is a transcription factor that binds to many enhancers and promoters and is involved in normal hematopoiesis development. Chromosomal translocations are well-documented and are associated with leukemia.
IRF4	interferon regulatory factor 4	301	A member of the IRF (interferon regulatory factor) family of transcription factors, important in the regulation of interferons in response to infection by virus, and in the regulation of interferon-inducible genes. IRF4 negatively regulates Toll-like-receptor (TLR) signaling. A translocation involving this gene and the IgH may be a cause of multiple myeloma.
CEBPB	CCAAT/enhancer binding protein (C/EBP), beta	301	Transcriptional activator in the regulation of genes involved in immune and inflammatory responses. Binds to an IL-1 response element in the IL-6 gene and plays a role in regulation of acute-phase reaction, inflammation and hemopoiesis.
CDK1	cyclin-dependent kinase 1	144	A member of the Ser/Thr protein kinase family that acts as a catalytic subunit of the protein kinase complex known as M-phase promoting factor (MPF), which is essential for G1/S and G2/M phase transitions of eukaryotic cell cycle.
PBK	PDZ binding kinase	144	Phosphorylates MAP kinase p38 and may be active only in mitosis. Can form a complex with TP53, leading to TP53 destabilization and attenuation of G2/M checkpoint in response to DNA damage.
HMGA2	high mobility group AT-hook 2	144	A transcriptional regulator that plays an key role in the meiotic G2/M transition and in cell cycle regulation via CCNA2.
FOXM1	forkhead box M1	144	Transcriptional factor regulating the expression of cell cycle genes essential for DNA replication and mitosis.
BRSK1	BR serine/threonine kinase 1	144	Serine/threonine-protein kinase that plays a key role in neuron polarization and centrosome duplication. Phosphorylates CDC25B, CDC25C, MAPT/TAU, RIMS1, TUBG1, TUBG2 and WEE1. Involved in the DNA damage checkpoint, probably by inhibiting CDK1 activity through phosphorylation and activation of WEE1, and inhibition of CDC25B and CDC25C.
WEE1	WEE1 homolog	144	A nuclear tyrosine kinase belonging to the Ser/Thr family of protein kinases. Catalyzes the inhibitory tyrosine phosphorylation of CDC2/cyclin B kinase, and appears to coordinate the transition between DNA replication and mitosis.
CDC25A	cell division cycle 25 homolog A	144	Tyrosine protein phosphatases and is required for progression from G1 to S phase of the cell cycle. It dephosphorylates CDK1 and CDK2 and it is involved in the DNA damage response. It has oncogenic properties that are not well-understood.
CDC25B	cell division cycle 25 homolog B	144	Tyrosine protein phosphatase required for G2/M phases of the cell cycle progression and abscission during cytokinesis. Dephosphorylates CDK1 and stimulates its kinase activity. CDC25B has oncogenic properties that are not well-understood.
CDC25C	cell division cycle 25 homolog C	144	Tyrosine protein phosphatase required for progression of the cell cycle by activating G2 cells into prophase. Directly dephosphorylates CDK1 and activates its kinase activity. It is also thought to suppress p53-induced growth arrest.
YWHAB, -E	tyrosine 3-monooxygenase /tryptophan 5-monooxygenase activation protein, beta, -epsilon	144	The 14-3-3 family of proteins interacts with CDC25 phosphatases, RAF1 and IRS1 proteins, suggesting a role in biochemical activities related to signal transduction, such as cell division, mitogenic signaling and regulation of insulin sensitivity. YWHAE has been implicated in the pathogenesis of small cell lung cancer.
RAP1GAP	RAP1 GTPase activating protein	144	T GTPase-activating-protein (GAP) that down-regulates activity of the ras-related RAP1 protein. RAP1 plays a role in diverse processes such as cell proliferation, adhesion, differentiation, and embryogenesis.
SFRP1,-2	secreted frizzled-related protein 1-2	762	Soluble frizzled-related proteins (sFRPS)are modulators of Wnts and Wnt signaling. They regulate differentiation and cell growth. Epigenetic silencing of SFRP genes leads to deregulation of the Wnt-pathway which is associated with cancer.
FZD2,-3,-5,-6, -8,-9	frizzled family receptor 2,-3,-5,-6, -8,-9	762	Most Frizzled receptors are coupled to the beta-catenin canonical signaling pathway, which leads to the activation of disheveled proteins, inhibition of GSK-3 kinase, nuclear accumulation of beta-catenin and activation of Wnt target genes.

4.3.5: *Evaluation: Overlap with GSEA*

To evaluate pathway annotation, I analyzed the three cancer datasets using GSEA against MSigDB canonical pathways. Overlap of top-scoring *Walktrap* results with GSEA results was evaluated by cross-validating the top GSEA pathways with pathways significantly over-represented in the data ($p \leq .01$). Notably in the BC data, module 224 exhibits significant over-representation in Cell Cycle, Pyrimidine Metabolism, and Apoptosis pathways which are among the top 10 enriched pathways in GSEA. Module 205 overlaps with the following highest-ranking GSEA pathways: Cell Cycle, Ubiquitin-mediated Proteolysis, DNA Replication and Apoptosis. HCC module 408 shows significant enrichment with the highest-ranking GSEA results, including: Tryptophan, Tyrosine, Phenylalanine, Beta-Alanine and Fatty Acid Metabolism, and Metabolism of Xenobiotics by Cytochrome P450 and Nuclear Receptors. Significant pathways over-represented in module 314 are Tryptophan, Tyrosine, Beta-Alanine, Lysine, Glycerophospholipid, Phenylalanine, Glycerolipid and Fatty Acid Metabolism. In CCA, module 144 overlaps with the top 10 ranked pathways in GSEA, including the ATM pathway, Cell Cycle, the P53 pathway, the ATR/BRCA pathway, and DNA replication, and module 487 shows overlap with Cell Cycle, P53, ATM and FAS pathways. Overall, the observed consistency with GSEA suggests that processes similar to those highlighted by GSEA are also found by searching for for enriched modules.

4.3.6: *Evaluation: Comparison with jActiveModules and Matisse*

The performance of *Walktrap* is compared with two highly cited platforms developed to find network modules using gene expression data in interaction networks, *jActiveModules* [149] and *Matisse* [151]. *jActiveModules* applies a simulated annealing algorithm to find modules across experimental conditions in gene expression data. An activity score is then calculated based on significance values associated with differential expression. *Matisse* applies a seed clustering algorithm that iteratively improves seed data, finds modules across expression data, and similarly determines a module score based on expression values. *Walktrap* modules do not include overlapping nodes and *jActiveModules* was configured to not allow overlap, while *Matisse* modules do include overlap. As a result, *Matisse* modules include more coverage of relevant interactions, but redundant sets of significant genes.

Parameters set to execute *jActiveModules* were *regional scoring*, *adjust score for size*, *overlap* = 0, and *number of modules* =1000. Parameters set for *Matisse* were *beta*=.95, *min seed size* =2, *min module size = 2*, *max module size* =200, search strategy all neighbors, and no regulation priors. The ability of these tools to identify cancer-related genes and interactions is evaluated using a list of derived from OMIN (Section 4.2.5). To assess the significance of each module, genes in the interaction network were randomly sampled to generate 5000 random distributions of cancer class labels for each module size. The performance of each platform is assessed by calculating a cancer-enrichment score for each module, summarized by a *z*-score assessing the number of known cancer genes in each module compared to a random distribution.

A comparison of of cancer-gene enrichment for the top twenty scoring modules generated by each platform is presented in Figure 20. *Walktrap* generally performed as well or better than *Matisse* or *jActiveModules* using the HCC and CCA data and performs consistently well overall. *Matisse* modules include overlap, so the corresponding set of top modules include redundancy and overlap between significant genes. By excluding overlap *Walktrap* does not find multiple modules including the same genes, but this design increases coverage of unique interactions across modules. I also consider module size; distribution of module sizes for each dataset and platform are shown in Figure 21. *jActiveModules* generated several large modules, including a module of 981 nodes and a module of 377 nodes. The majority of significant modules generated by *Matisse* were over 100 nodes. Generally, such large clusters demand further mining to discover the most relevant interactions and genes in each module. The smaller distribution of module sizes associated with *Walktrap* highlights a more specific and informative set of biological interactions that facilitates interpretation of modules; the functional annotation of large modules my be too general to be meaningful. Further, the performance of *Walktrap* computation was more efficient than the other tools, I was able to run all analyses on an 2.8 GHz, 64-bit machine using 8GB RAM, where other tools required additional computing resources. The efficiency of the algorithm is described by the original authors[7].

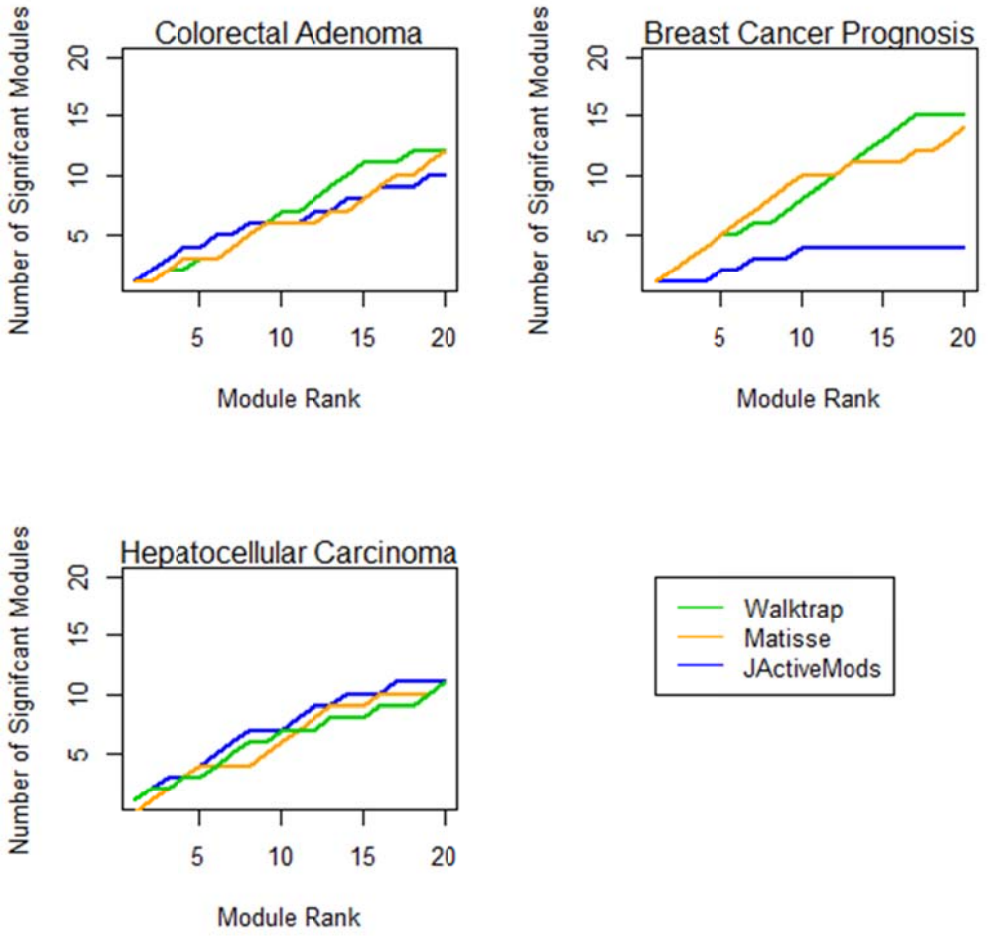


Figure 20: Comparison of Top Modules from *Walktrap*, *Matisse*, and *jActiveModules*. Yellow lines measure perfomance in colorectal cancer data (CCA), green for breast cancer (BC) and blue for hepatocellular carcinoma (HCC). Green lines show *Walktrap* performance, blue *jActiveModules,* and orange *Matisse*. *Walktrap* performs comparably to or better than the other approaches across datasets. In the BC data, resulted in one very large and significant module of 981 nodes, but few significant modules overall. *Matisse* includes overlapping significant genes within its modules where *Walktrap* does not and *jActiveModules* is configured not to inlcude overlap.

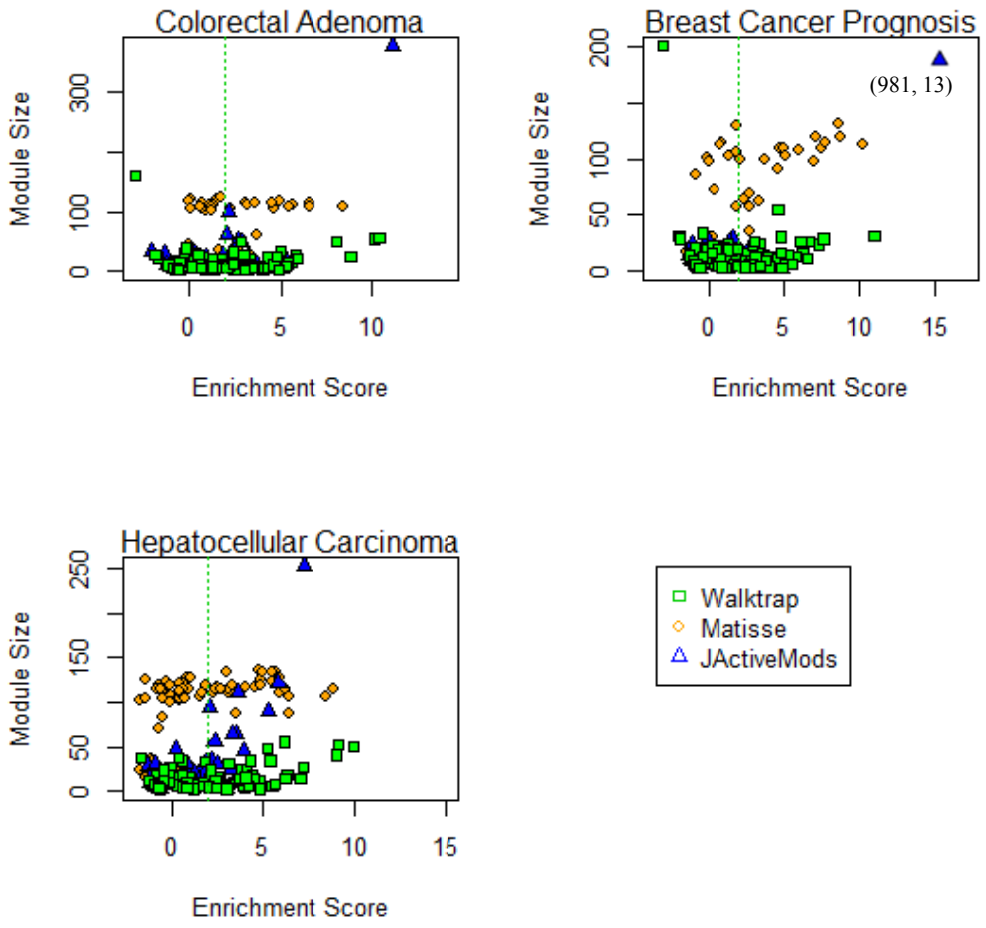


Figure 21: Distribution of Module Sizes and Scores. *Walktrap* markers are noted in green, *Matisse* in orange, and *jActiveModules* in blue. *Walktrap* with a size threshold of 200, identifies significant modules that are generally smaller. Smaller modules tend to have more specific and informative functional interpretation; the functional annotation of large modules may be too general to be meaningful.

4.4: Conclusion

Network analysis provides a framework to search for communities of genes associated with disease by modeling their coordinated behavior and biological knowledge of their interactions. I utilize *Walktrap*, a random walk algorithm optimized to search for communities in large networks, to mine for disease genes in a weighted interaction network.

This approach is used to discover cancer-associated modules in a network of biological interactions weighted by differential gene expression of breast cancer, hepatocellular carcinoma and colorectal cancer data.

This study identifies modules relevant to the etiology of multiple cancer outcomes, and suggests interactions among promising candidate genes for further study of molecular interaction that influence cancer or potential therapeutic interventions. Functional analysis of modules discovered in this analysis reveals strong enrichment of cancer related pathways and known cancer genes. Pathways enriched across the three data sets include those involved in cell cycle control, DNA replication, DNA damage and repair, amino acid metabolism, inflammation, and cell adhesion and migration. Specifically, several genes may represent targets for further research, including *CBLC* or *IRS2* which influence breast cancer survival; transcription factors *RPS6KA2* and *RPS6KA6* and the interaction among *MCM/CDC* and *ORC* cell cycle control genes in the onset of hepatocellular carcinoma; or cell-cycle genes *BRSK1*, *WEE1*, *CDC25C,* and the transcription factor *SPIB* in colorectal adenoma development. These genes and their interactions can serve as a strong basis for hypothesis generation regarding their functional roles and therapeutic value in cancer.

The *Walktrap* approach identifies biologically relevant modules associated with cancer and performs well compared to other module search platforms, *Matisse* and *jActiveModules*. Strong performance combined with smaller, more specific, and non-overlapping modules, facilitates the biological interpretation of these results. These modules reflect known pathways in cancer and present hypotheses for new studies. Future work may include an analysis across additional cancer and other complex disease data, or apply these methods to integrate additional classes of genomic data. In Chapter 5, I investigate expanding the network to include microRNA-mRNA regulatory information from cancer expression studies.

.

Chapter 5: Analysis of miRNA Data in Random Walk-Generated Expression Modules

5.1: Introduction

With technological advancement in measuring biological variability in genes, mutations and epigenetic interactions, there is a corresponding demand for analytical methods to integrate large and diverse data sets and investigate their association with complex phenotypes. These include, for example, mRNA expression, SNP, copy number, proteomic, genomic mappings and microRNA (miRNA) measurements. Modeling of these data requires sophisticated analytical approaches and high computational efficiency to merge and analyze the breadth and scope of these interactions. Biological networks are powerful frameworks to integrate disparate data types, and explore high dimensional data for interactions associated with disease outcomes. Network-based approaches have been applied to genomic studies to better understand the complexity of such data[20,107,112,124], and have been proven useful in analyses of regulatory interactions; for example, investigating transcription factors, methylation activity, and miRNA silencing of gene expression and downstream targets [120,133,134,198]. In Chapter 3, I examined the application of graph-based analyses to module discovery using evidence of biological interactions and cancer expression data. To better understand the post-translational behavior of cancer genomics; in this chapter, I integrate miRNA coexpression data into the molecular interaction network to find significant cancer-associated modules enriched with miRNA regulatory interactions.

While analysis of transcription factor networks is developing as a mature area of research, relatively new work employs a network approach to examine miRNA-mRNA interactions and their associations with disease outcomes. MiRNAs are short non-coding RNA molecules (between 19-22 nucleotides in length) that bind to mRNA post transcriptionally and interfere with mRNA translation[199] (Figure 22). Comparable to transcription factor regulation of gene expression, miRNAs typically bind to a "seed" region of their mRNA targets, usually nucleotides 2-7 in the 3′ UTR of the target mRNA. Multiple miRNAs can bind to a particular mRNA, and a miRNA may bind to as many as hundreds of distinct mRNAs in its mRNA "targetome". There are approximately 2,000 known mature human miRNAs (miRBase Release 19; August 2012)[200] that regulate greater than 60% of all

protein-encoding genes, where sequence similarity helps determine specificity. Based on their role in epigenetic control of gene expression, and their broad "targetome", these molecules are capable of regulating diverse cellular functions, including development, differentiation, proliferation, apoptosis and metabolism [198,201].

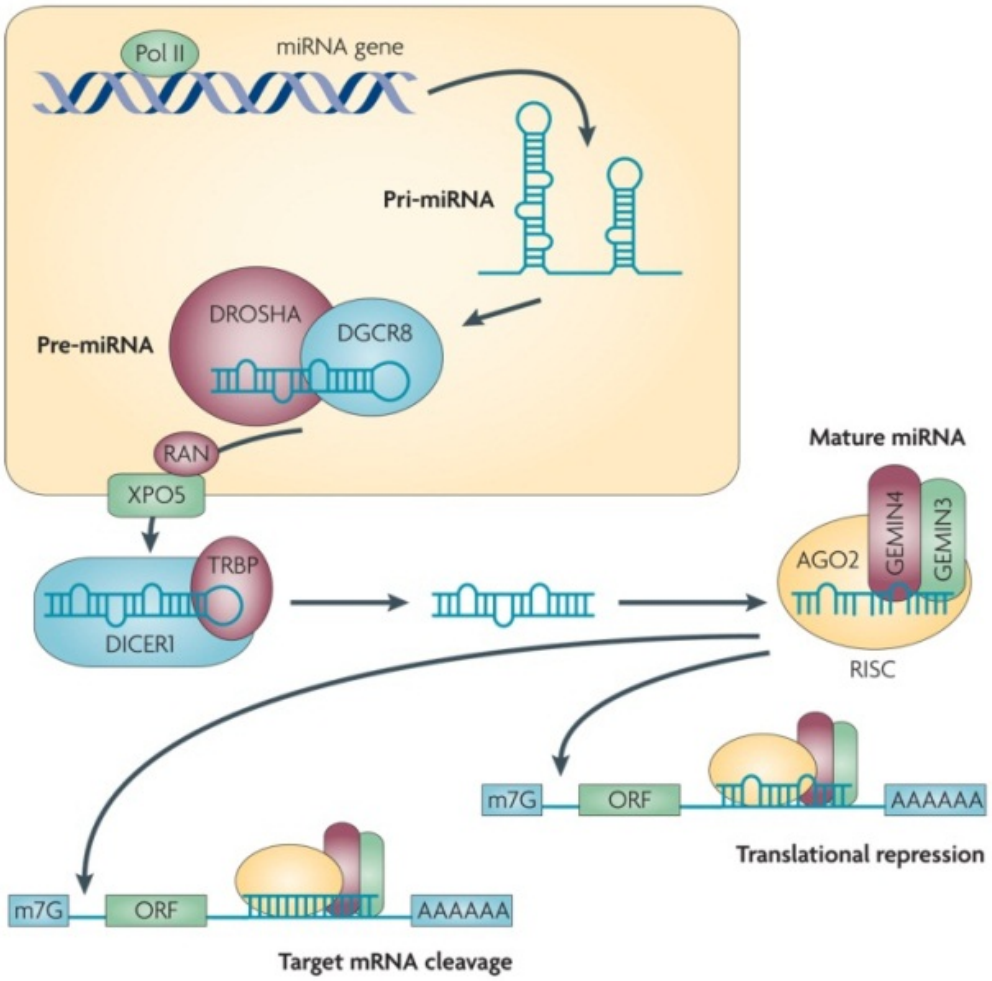


Figure 22: The miRNA Lifecycle. This figure from Ryan 2010[199] shows the miRNA lifecycle. RNA polymerase II (Pol II) produces a 500–3,000 nucleotide transcript, called the primary microRNA (miRNA), or pri-miRNA, that is then cropped to form a pre-miRNA hairpin by a multi-protein complex that includes DROSHA (~60–100 nucleotides) (a simplified view is shown here). This double-stranded hairpin structure is exported from the nucleus by RAN GTPase and exportin 5 (XPO5). Finally, the pre-miRNA is cleaved by DICER1 to produce two miRNA strands, a mature miRNA sequence, approximately 20 nucleotides in length, and a short-lived complementary sequence, which is denoted miR* and is sometimes referred to as the passenger strand or 3p strand. The single stranded miRNA is incorporated into RISC, which then binds to the 3′ untranslated region of the target mRNA sequence to regulate repression and cleavage.

MiRNAs play an important role in cancer where they play a role in regulating oncogenic and tumor-suppressor pathways[198,202]. Burchard et al. analyze a correlated set of miRNAs and mRNAs in hepatocellular carcinoma and find *miR-122* is under-expressed in tumor tissue[11]. They confirm that the putative targets of this miRNA, *SMARCD1*, *MAP3K3*, *CAT-1* are down-regulated *miR-122* with an increase in *miR-122* expression while secondary target *PPARGC1A* is up-regulated with a decrease in *miR-122*. These genes participate in mitochondrial biogenesis pathways, including fatty acid metabolism; and *miR-122* acts as a tumor suppressor where it can play a role to stabilize metabolic function in the liver and thus improve patient survival. In a breast cancer study including correlated miRNA-mRNA prognostic profiles, Buffa et al.[12] find *miR-210*, *miR-128a* and *miR-27b* to be differentially regulated and prognostic of breast cancer survival. Fu et al. study miRNA-mRNA pairs co-regulated in colorectal cancer and involved in the *Wnt* pathway and find *mir-21*, *mir-223*, *mir-224*, *mir-29a*, *mir-29b* to be upregulated and their predicted targets, *SFRP1*, *SFRP2*, *RNF138*, and *KLF4* to be downregulated. They experimentally confirm the relationship between *mir-29a* and *KLF4* at both the RNA and protein levels in colorectal cancer cells. Laios et al. investigate miRNAs and their association with pathways involved in ovarian cancer [203] and show *miR-214* induces cell survival and cisplatin resistance by targeting *PTEN* regulation of the *Akt* pathway, and *miR-15b* and *miR-16* were found to inhibit *BCL2*-mediated apoptosis. Findings by Gennarino et al.[204], show that *miR-519d*, *miR-190* inhibit and *miR-340* enhances *TGFβ* signaling, cell proliferation and ceullular migration in lung carcinoma. These studies demonstrate a diverse means of miRNA-based regulation of cancer-related pathways and suggest miRNA co-expression analysis as a general approach to identify miRNA targets in cancer.

Biological networks have been applied to model interactions between miRNAs and their targets and to identify miRNA subnetworks associated with cancer[198,202]. Satoh et al.[106] assemble a human miRNA targetome incorporating differentially expressed miRNAs and their predicted targets from thousands of human tissue samples. They use a neighboring network-search algorithm to find co-regulated miRNA-mRNA pairs in normal and cancer tissues and use expression data to validate miRNA-mRNA interactions. Dysregulated processes were found to be associated with differentially expressed miRNAs in invasive breast cancer cells, including key pathways regulated by *MYB* (*miR-15a*), *Rb/E2F* (*miR-106b*), *p53* (*let-7d*), *ZEB* and *EMT* (*miR-200b*). Overall, the most relevant pathological event in their human miRNA targetome was “cancer”, suggesting that miRNAs play a specialized

role in oncogenesis. Bandyopadhyay et al.[137] generate a bipartite network by mining experimentally verified cancer-miRNA relationships from the literature, miRNA-mRNA interactions based on predicted targets, experimentally-supported interactions, and co-expression models. They mine for cancer-miRNA modules and show that neighboring miRNAs are often similarly up- or down-regulated, suggesting coordinated activity of the miRNAs on target gene regulation in cancer tissues or cell lines. Specifically, they find downregulation of *miR-143* and *miR-145* in colon cancer, downregulation of *miR-127* in bladder carcinoma, and overexpression of *miR-99* is in pancreatic cancer. O'Day et al. [202] analyze a network of mRNAs and their predicted miRNAs in breast cancer and find a well-connected gene-interaction network including *MYC* as a hub interacting with critical cell-cycle genes and regulated by key miRNAs including: *miR-206*, *-34a, -200, -17-5p*, *-125a/b*, *-21*, *-155, -373/-520c*, *-31* and *let-7*. Nam et al.[205] extract network clusters from an integrated network to distinguish drug resistant states from drug sensitive states in breast cancer. They identify clusters that contribute to antiestrogen resistance which include miRNAS *miR-146a*, *-27a*, *-145*, *-21*, *-155*, *-15a*, *-125b*, and *let-7s*, and *miR-221/222*. Zhang et al.[206] perform a network cluster analysis to identify correlated miRNA-mRNA pairs to distinguish primary and metastatic prostate cancer tumor subtypes and *miR-106b*, *-191*, *-19b*, *-92a*, *-92b*, *-93*, and *-141* were found to be enriched in metastatic samples. Several studies also note that miRNA networks consist of well-connected miRNA hubs that are dysregulated in cancer[137,138,207], and that miRNAs tend to target hub-genes in human protein interaction (PPI) networks[106].

Modeling interactions among miRNAs and their correlated mRNA targets provides an additional layer of evidence to identify key gene interactions and increases confidence in the discovery of functional associations between genes and disease. In high-dimensional data, this additional knowledge source reduces the feature space to narrow the search for candidate genes. Combined evidence summarizing the coordinated activity of miRNAs and their predicted targets in cancer tissues, the significance of biological modules in networks weighted with cancer outcomes, and relevant functional annotation of those modules, increases the likelihood that they have true causal relationships with cancer. Thus, such network analysis using interaction, experimental and regulatory data can improve the search for miRNAs, miRNA-miRNA interactions, or target mRNAs associated with disease.

Earlier, in Chapter 4, I investigated the application of a weighted random walk and a modularity-driven clustering algorithm to search for modules of interacting genes significantly associated with cancer onset and progression. This framework merges gene expression data and protein interaction and metabolic interaction data in a weighted network,

a random walk algorithm with a module-searching component, and a bootstrap scoring metric to find significant modules. In this chapter, I further leverage the ability of this molecular interaction network to integrate interaction, experimental, and miRNA regulatory data to improve the search for modules associated with cancer. I assess the value of using miRNAs and their targets by comparing these results to previously published findings of miRNA-mRNA interactions in cancer and to the analysis of breast cancer and hepatocellular carcinoma outcomes using only mRNA data in Chapter 4. Finally, the biological relevance of these findings is evaluated by functional annotation and supported evidence in the literature.

5.2: Methods

5.2.1:Overview

A graph-based random walk algorithm is employed in an integrated interaction network to mine hepatocellular carcinoma (HCC) and breast cancer (BC) expression data, including correlated mRNA and miRNA expression, to search for modules of genes associated with cancer outcomes. First, metabolic, signaling and protein interactions from the Kyoto Encyclopedia of Genes and Genomes (KEGG)[35] and the Human Protein Reaction Database (HPRD)[31] are used to construct a network of biological interactions. I evaluate matching and integration methods to include miRNAs in the network analysis. Matching methods include selecting the best overall match for each miRNA, filtering the best three and five matches, or including all correlated matches. Methods to integrate miRNA/mRNA pairs to the interaction network include using miRNAs as an additional interaction type with edges weighted directly corresponding to fold change as in section 4.2.4, or using a linear transformation of the fold change values to create edge weights. The *Walktrap* random walk algorithm is applied to this wieghted network to discover cancer-associated module that are enriched with differential miRNA regulatory activity. To evaluate findings I review functional annotation of the results, and compare these results to published data and to the study of HCC and BC datasets not including miRNA data (as described in Chapter 4). These methods are summarized in Figure 23.

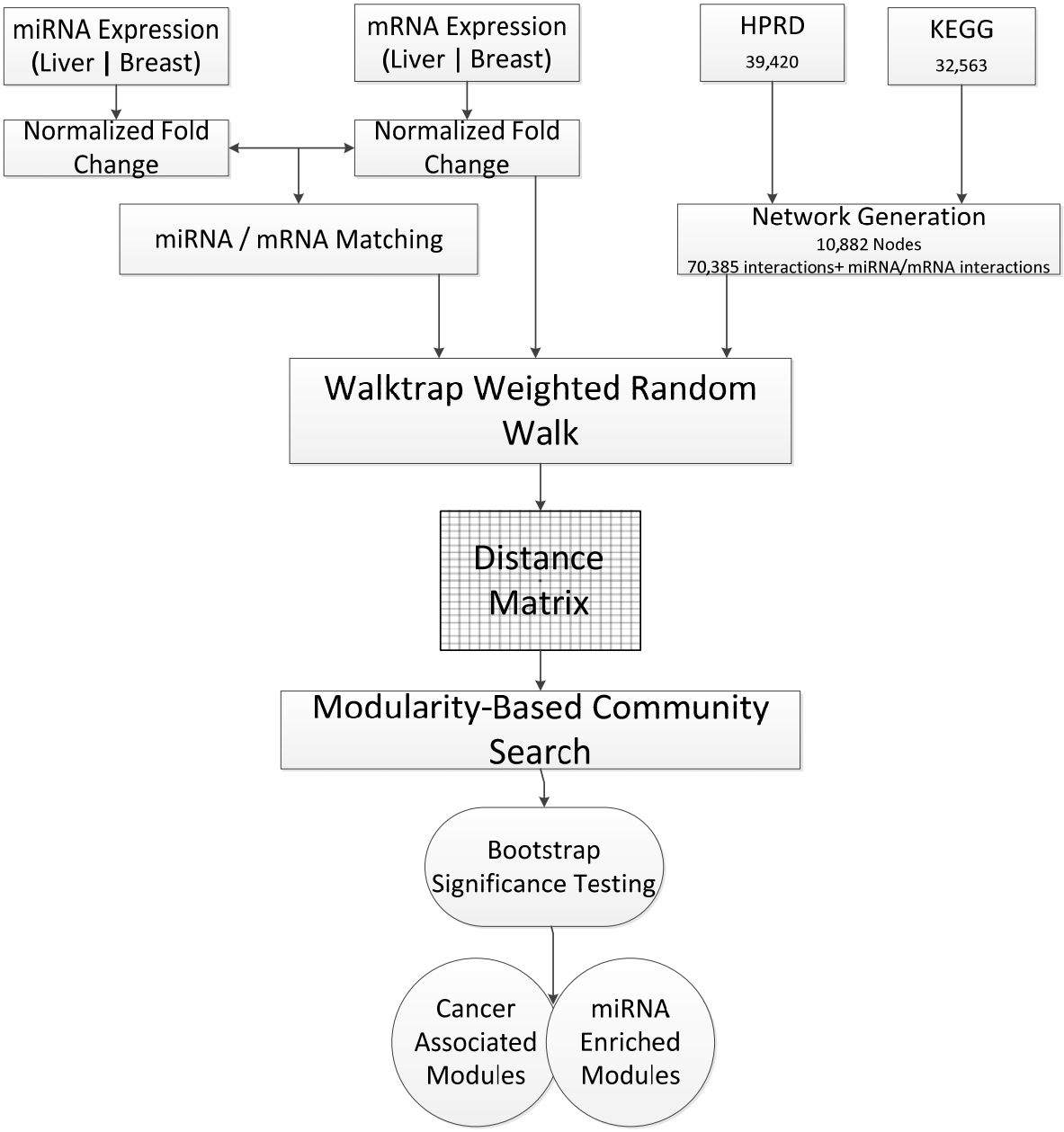


Figure 23: miRNA Network Analysis and Evaluation. This flowchart summarizes miRNA network analysis and evaluation. HCC and BC miRNA/mRNA expression data are downloaded from GEO. MiRNA-mRNA pairs are integrated into the interaction network and the *Walktrap* algorithm is applied to search for dense modules significantly associated with cancer outcomes and enriched with miRNA targets.

5.2.2: Gene Expression Data

Two cancer data sets including mRNA and miRNA expression data were downloaded from the Gene Expression Omnibus (GEO) [181]. GSE22058 includes genome-wide expression profiles of both miRNAs and mRNAs from a cohort of hepatocellular carcinoma patients (HCC) in Hong Kong, comparing expression levels of paired tumor tissue and normal adjacent tissue [11]. The platform used for measuring mRNA expression is the Rosetta/Merck Mouse 23.6K 3.0 A1 microarray, and the Rosetta human miRNA qPCR array is used for miRNA measurement (Rosetta Inpharmatics/ Merck Pharmaceuticals, Seattle, WA). GSE22220 is a study of early primary breast cancer (BC) including correlations between mRNA and miRNA expression in 210 tumor samples[12]. mRNA expression levels are assessed by the Illumina humanRef-8 v1.0 expression beadchip, and miRNAs expression levels using the Illumina Human v1 MiRNA expression beadchip (Illumina Inc, San Diego, CA). The study includes clinical features (ER status, adjuvant treatment, endocrine therapy or combination chemotherapy, and CMF) and follow-up at 10 years, to assess prognostic features. Prognosis indicators include relapse, which I extract to assess risk (no relapse= 131, relapse= 79). The data are summarized in Table 4. I calculate normalized, log-transformed fold change values and p-values for each data set. P-values were corrected for multiple testing using the Benjamini and Hochberg false discovery rate[182]. All analyses were performed in R using Bioconductor [53].

Table 7: Description of Cancer Expression Data

GEO Accession	Reference	Clinical Outcome	Cases	Controls
GSE22058	Burchard et al. 2010	HCC tumors (HCC)	192 hepatocellular tumors	192 paired adjacent non-tumor
GSE22220	Buffa et al. 2011	BC prognosis (BC)	210 mRNA and miRNA samples	prognosis scores for each sample

5.2.3: MiRNA-mRNA Matching

Correlation among samples is measured by calculating a Pearson Correlation coefficient comparing differential expression of miRNAs and mRNAs in the HCC and BC data sets. Significantly correlated pairs below a p-value of .05 were selected for further investigation and integration into the molecular network. Four different methods were examined to match miRNA to mRNAs. First, I took the best matches using an optimal matching algorithm[208]. This approach finds an exclusive best match based on correlation scores for each miRNA. Next, the best (up to three) matches and the best (up to five) matches for each miRNA were selected based on ranked correlation values. Both the best three and the best five matches are non-exclusive matches. Finally, in the last approach, all possible matches are included for each miRNA.

Further, I tested these approaches using a biological filter to select only those matches that had a seed match based on TargetScan Human Release 6.2 [41]. This tool finds predicted miRNA targets based on 7-mer and 8-mer seed matches. The TargetScan configuration for these queries incorporated conserved and non-conserved regions to include a broad base of potential mRNA-miRNA interactions. All predicted targets based on the TargetScan prediction algorithm using sequence alignment, conservation across species and flanking seqments were retained, no threshold for conservation score or context score were configured[41]. Altogether, across the four matching approaches, I consider filtered and non-filtered data, resulting in eight matching combinations. Code for matching and filtering miRNA-mRNA pairs is presented in Appendix B.

5.2.4:Network Construction

The interactome in this study was built by extracting human interactions from the Kyoto Encyclopedia of Genes and Genomes [35]and HPRD [31], and this network is used to assess the incorporation of miRNA-target interactions extracted from correlated expression samples. Details of the network construction are discussed in section 4.2.3.

5.2.5: Weighting Scheme

To calculate edge weights for mRNA-mRNA interactions in the network, I use an average of the absolute fold change values of the two adjacent nodes as discussed in section 4.2.4.

I include miRNA-mRNA interactions based on matching schemes discussed in section 5.2.3, using 1) optimal matching 2) the best three matches 3) the best five matches, or 4) all matches. Each network is assessed using matches filtered based on their seed match using TargetScan, or unfiltered matches. Two weighting schemes are used to add weights to edges representing interactions between miRNAs and their targets. In the first scheme, I add miRNA expression weights to the network applying the same weighing scheme as mRNA-mRNA matching, using a square of the mean fold-change values. The second scoring scheme applies a weight that is a linear transformation of the fold change based on the number of mRNA matches multiplied by the fold change of the miRNA (*n* x fold change). The matching and weighting schemes are summarized in Figure 24.

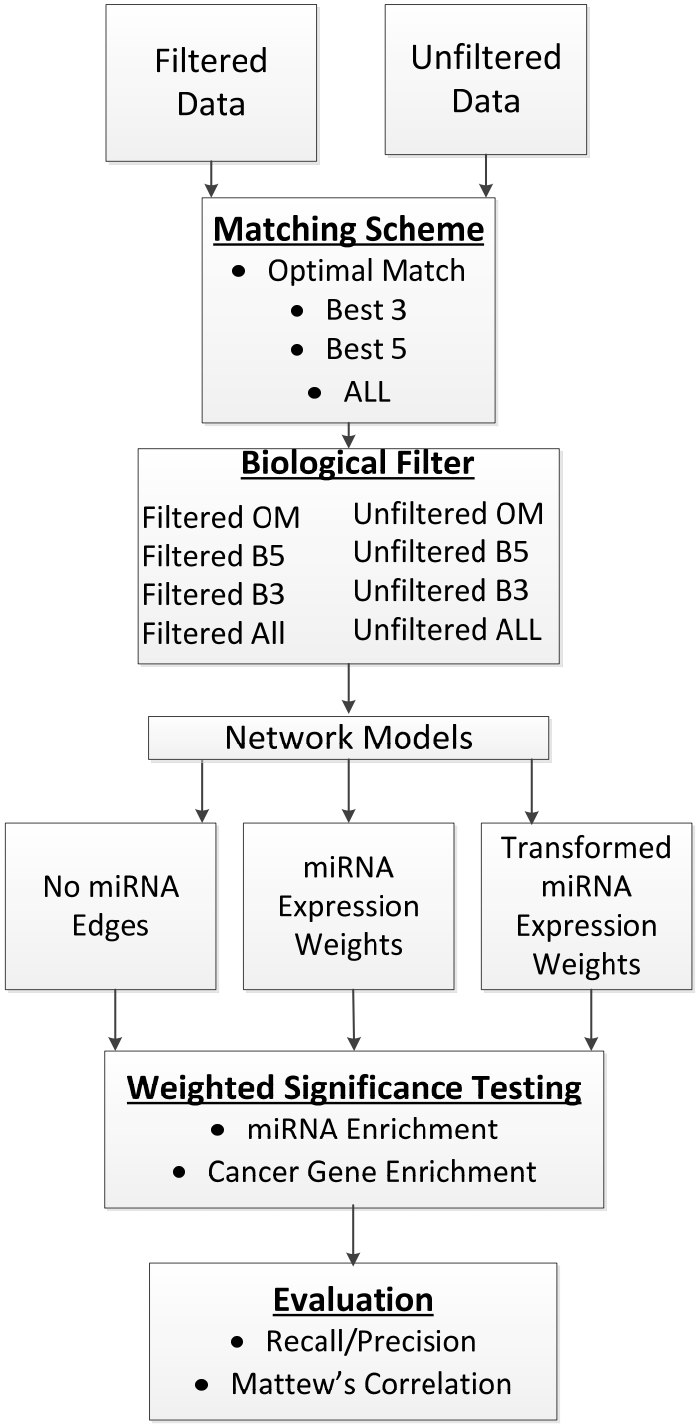


Figure 24: miRNA Match and Weight Scheme Evaluation: Flowchart showing the four matching schemes, and filtering used to select miRNA-mRNA pairs. These pairs are then added using a square of the mean fold change values, or after a linear transformation of the weights. After applying the *Walktrap* algorithm, the methods are evaluated by Precision/Recall and Matthews Correlation. The best 5 matches and using a square of the mean weights showed the best performance.

5.2.6: Community Analysis

Random walks have been shown to be valuable when applied to study genomic data in biological networks[112,167]. The random walk algorithm implemented here was chosen because it incorporates the topology of the network to calculate distance metrics, and optimizes the community search componenet by using the graph theoretic concept of modularity. Details of the random walk are described in section 2.5, modifications to adjust the merge stopping criteria based on module size, score and modularity are reviewed in section 4.2.5.

5.2.7: Module Scoring

The magnitude of the expression signal for each module was compared against a random distribution. Module weight was calculated by taking an average of the node weights; each node corresponds to a squared transformation of the maximum fold change for probes corresponding to each gene symbol. Higher-confidence modules greater than three nodes in size were tested for significance. A module score was then calculated by comparing the significance of the module weight to a distribution of 5000 random samples of expression values for each module size. Enrichment of miRNAs in each module is assessed by comparison with a random distribution of 5000 random samples of miRNA matches where 1 is a correlated/predicted target and 0 is not. Code for module scoring and significance testing is presented in Appendix A.

5.3: Results

5.3.1: Assessment of Weighting and Scoring Schemes

I evaluate several approaches to match miRNAs and weigh miRNA-enriched modules in the community network analysis. To assess the performance of weights and scoring schemes, I measure the precision and recall of these methods in the BC data to detect known cancer genes (the gold standard list was created by text mining and manual curation of OMIM, details discussed in sections 4.2.5 and 4.3.6). Evaluation data for these approaches

are described in Appendix B. The most sensitive approaches use all the network data but may have many false positives. More selective approaches using only the top filtered matches are more precise but generally are not sufficiently sensitive to identify many important modules. Results were filtered by top-ranked Precision and Recall values. I then calculated Matthew's correlation coefficient [209] to evaluate the best overall performance. The filtered data using non-transformed network weights and the best five matches performed best in the BC data. The BC data were used as the training set and the methods were then validated using the HCC data; the Precision, Recall and Matthews Correlation Coefficient figures performed best in the HCC data as well.

When incorporating the best five filtered miRNA-mRNA matches in the data, 95 edges are added to the BC data and 19 edges added to HCC data. I evaluate the results of these analyses compared with findings with the original studies from which the HCC and BC data were obtained, by Burchard[11]and Buffa[12], and compared with analyses of similar HCC and BC data sets using only mRNA data in Chapter 4. Finally, these results are evaluated by their functional annotation and biological relevance; I evaluate functional annotation of miRNAs and mRNA targets in the significant modules using ConsensusPathDB and evidence of the functional relevance of the interactions from previous literature.

5.3.2: Functional Annotation

Functional annotation of significant modules is assessed using ConsensusPathDB[55]. I queried genes in the top-scoring modules for over-representation analysis comparing against pathway gene sets (including: KEGG, WikiPathways [183], PID [34] and Reactome [30]), and a minimum overlap of two genes with the input gene list and the consensus pathway. Results were filtered using a default p-value of .01. Canonical cancer pathways and pathways associated with hallmarks of cancer are enriched in each cancer dataset: cell-cycle control (including *MAPK*, *JNK*, *TGFβ*, and *Wnt*), DNA replication/repair, cellular adhesion/migration, cell differentiation apoptosis, angiogenesis, evasion of the immune response and immortality. A summary of statistics and a sample of representative pathways for the top scoring modules are presented in Table 8.

BC modules are highly enriched with cell cycle control, transcriptional regulation, growth signaling, cytokine and chemokine signaling, T-cell and B-cell signaling, focal

adhesion and angiogenesis-related genes. A number of BC modules are also annotated with progesterone and estrogen hormone signaling, and levels of these hormones are known to correlate with BC risk. In HCC, detoxifying pathways, including cytochrome P450, nucleotide and fatty acid metabolism, cellular adhesion and interactions, DNA repair and cell-cycle signaling are among the most enriched pathways. Inflammation and deregulation of liver-related detoxifying pathways are frequent markers of carcinogenic toxicity, oxidative stress and tumorigenesis. Amino acid synthesis and metabolism pathways, related to the stability of DNA replication and repair are over-represented across all three cancer types, though most notably in HCC. These findings are consistent with overrepresented pathways in mRNA-only modules discussed in section 4.3.1.

Table 8: Functional Annotation for Significant Modules

Breast Cancer			
Module	Score	Size	Functional Annotation
22	56.72	21	SIGNALING MEDIATED BY HDAC CLASS I, II AND III, SIGNALING BY NOTCH, TRANSCRIPTIONAL MISREGULATION IN CANCER, REGULATION OF PGC-1A, CELL DIFFERENTIATION, CELL CYCLE, RANBP2 REGULATES TRANSCRIPTIONAL REPRESSION, VIRAL CARCINOGENESIS, TGF BETA SIGNALING, ADIPOGENESIS, C-MYB TRANSCRIPTION, BCR SIGNALING, P38 MAPK SIGNALING, TRANSCRIPTIONAL ACTIVITY OF SMAD2/SMAD3:SMAD4, HTERC TRANSCRIPTIONAL REGULATION, REGULATION OF TELOMERASE, FATTY ACID, TRIACYLGLYCEROL, AND KETONE BODY METABOLISM, GENE REGULATION BY PEROXISOME PROLIFERATORS VIA PPARA, ACUTE MYELOID LEUKEMIA, REGULATION OF RB, NUCLEAR ESTROGEN RECEPTOR ALPHA NETWORK, IL4-MEDIATED SIGNALING, HIF-1-ALPHA TRANSCRIPTION
63	63.79	143	ANDROGEN RECEPTOR SIGNALING, PROSTATE CANCER, INTEGRATED BC PATHWAY, TRANSCRIPTIONAL ACTIVITY OF SMAD2/SMAD3:SMAD4, TGF BETA SIGNALING, INTEGRATED CANCER PATHWAY, DNA DAMAGE RESPONSE, CELL CYCLE, P73 TRANSCRIPTION, INTEGRATED PANCREATIC CANCER PATHWAY, TRANSCRIPTIONAL ACTIVITY BY PML, MIRNA REGULATION OF DDR, RANBP2 REGULATES TRANSCRIPTIONAL REPRESSION, C-MYC PATHWAY, CHROMATIN REMODELING, GLUCOCORTICOID RECEPTOR NETWORK, BARD1 SIGNALING, VIRAL CARCINOGENESIS, RB TUMOR SUPPRESSOR/CHECKPOINT, TRANSCRIPTIONAL MISREGULATION IN CANCER, PATHWAYS IN CANCER, P53 PATHWAY, EFP CONTROLS CELL CYCLE AND BREAST TUMORS GROWTH, UBIQUITIN MEDIATED PROTEOLYSIS, NON-HOMOLOGOUS END-JOINING, ATM SIGNALING, REGULATION OF TELOMERASE, SIGNALING MEDIATED BY HDAC CLASS I/II, ARF INHIBITS RIBOSOMAL BIOGENESIS, LKB1 SIGNALING, WNT SIGNALING, AP-1 TRANSCRIPTION, AHR PATHWAY, NUCLEAR ESTROGEN RECEPTOR NETWORK, ADIPOGENESIS, REGULATION OF NUCLEAR BETA CATENIN, DEGRADATION OF CYCLIN D1, SEROTONIN RECEPTOR 4-6-7 AND NR3C SIGNALING, DNA REPAIR, BTG PROTEINS AND CELL CYCLE REGULATION, PTC1 REGULATES CELL CYCLE, FOXM1 TRANSCRIPTION, ADHERENS JUNCTION, E2F NETWORK, INTERFERON SIGNALING, NOTCH SIGNALING, AURORA A SIGNALING, NGF SIGNALING VIA TRKA, ID SIGNALING, AKAP95 IN MITOSIS/CHROMOSOME DYNAMICS, RNA POLYMERASE TRANSCRIPTION, SIGNALING BY EGFR IN CANCER, FAS SIGNALING (CD95), BRCA1 BRCA2 AND ATR IN CANCER, CDC25 AND CHK1 REGULATORY PATHWAY IN DDR, PI3K-AKT SIGNALING, SREBP SIGNALING, ERBB SIGNALING, CALCINEURIN-DEPENDENT NFAT SIGNALING, NF-KAPPA B SIGNALING, RETINOIC ACID RECEPTORS-MEDIATED SIGNALING, ALPHA-SYNUCLEIN SIGNALING, SCF-BETA-TRCP DEGRADATION OF EMI1, P38 MAPK SIGNALING, TSH SIGNALING, HIF-1-ALPHA TRANSCRIPTION, REGULATION OF APC/C ACTIVATORS
74	57.13	27	ENDOCYTOSIS, SIGNALING BY SCF-KIT, EGF-EGFR SIGNALING, SIGNALING BY EGFR IN CANCER, SIGNALING BY ERBB, IMMUNE SYSTEM, JAK-STAT SIGNALING, NEUROTROPHIN SIGNALING, SIGNALING BY NGF, TRANSCRIPTIONAL TARGETS OF DELTANP63, NOTCH SIGNALING, PTP1B SIGNALING, TRANSCRIPTIONAL TARGETS OF TAP63
212	61.48	30	CHEMOKINE SIGNALING, LIGAND-BINDING RECEPTORS, SIGNALING BY GPCR, CYTOKINE-CYTOKINE RECEPTOR INTERACTION, ENDOCYTOSIS, ADRENOCEPTORS, IL8- AND CXCR1-MEDIATED SIGNALING, ACTIVATION OF PKA, TOLL-LIKE RECEPTOR SIGNALING, CSK INHIBITS SIGNALING THROUGH THE T CELL RECEPTOR, EBV LMP1 SIGNALING, ARF6 SIGNALING
269	139.06	75	T CELL RECEPTOR SIGNALING, FC EPSILON RI SIGNALING, BCR SIGNALING, GPVI-MEDIATED ACTIVATION CASCADE, FC GAMMA R-MEDIATED PHAGOCYTOSIS, IMMUNE SYSTEM, SIGNALING BY CBL, NK CELL CYTOTOXICITY, DAP12 SIGNALING, INTERLEUKIN SIGNALING, SCF-KIT SIGNALING, LEUKOCYTE MIGRATION, GAB1 SIGNALOSOME, PLATELET ACTIVATION, CHEMOKINE SIGNALING, PI3K/AKT SIGNALING IN CANCER, PROLACTIN SIGNALING, GENERATION OF SECOND MESSENGER MOLECULES, T-CELL APOPTOSIS, INTERFERON TYPE I, KIT RECEPTOR SIGNALING, ACUTE MYELOID LEUKEMIA, SIGNALING BY EGFR IN CANCER, FOCAL ADHESION, FGFR SIGNALING, CHRONIC MYELOID LEUKEMIA, ENDOMETRIAL CANCER, VEGF SIGNALING, COLORECTAL CANCER, PHOSPHATIDYLINOSITOL SIGNALING, SIGNALING BY PDGF, APOPTOSIS, JAK-STAT SIGNALING, NEUROTROPHIN SIGNALING, NGF SIGNALING VIA TRKA, MTOR SIGNALING, CELL SURFACE INTERACTIONS, CHOLINERGIC SYNAPSE, TIE2 SIGNALING, ERBB SIGNALING, NF-KAPPA B SIGNALING, REGULATION OF ACTIN CYTOSKELETON, INSULIN SIGNALING, AMPK SIGNALING, TOLL-LIKE RECEPTOR SIGNALING, NEPHRIN INTERACTIONS, TRANSLOCATION OF ZAP-70 TO PECAM1 INTERACTIONS, G ALPHA SIGNALING, VIRAL CARCINOGENESIS, RANKL-RANK SIGNALING, PHOSPHOLIPID METABOLISM, CTLA4 SIGNALING, NEF SIGNAL TRANSDUCTION, PD-1 SIGNALING, INFLAMMATORY RESPONSE, G13 SIGNALING, CELL-CELL COMMUNICATION, IRS-MEDIATED SIGNALING, RAS SIGNALING, PATHWAYS IN CANCER, EPO RECEPTOR SIGNALING, RAC1 CELL MOTILITY, CELL ADHESION MOLECULES (CAMS), CXCR4 SIGNALING, TRKA RECEPTOR SIGNALING, METABOLISM OF LIPIDS/LIPOPROTEINS, NRAGE SIGNALS DEATH THROUGH JNK, NOTCH SIGNALING, LEPTIN SIGNALING, TGF BETA SIGNALING, IGF-1 SIGNALING, ANGIOGENESIS
292	51.78	20	CALCIUM SIGNALING, GASTRIN-CREB SIGNALING VIA PKC/MAPK, SIGNALING BY

			GPCR, REGULATION OF INSULIN, SECRETION BY ACETYLCHOLINE, PEPTIDE LIGAND-BINDING RECEPTORS, THROMBOXANE SIGNALING, REGULATION OF ACTIN CYTOSKELETON, ACTIVATION OF PKA, THROMBIN SIGNALING, CHREBP REGULATION BY CARBOHYDRATES AND CAMP, CSK INHIBITS SIGNALING THROUGH THE T CELL RECEPTOR, ACE INHIBITOR PATHWAY, ADP SIGNALING, COMPLEMENT AND COAGULATION CASCADES, SEROTONIN RECEPTOR 2 AND ELK-SRF-GATA4 SIGNALING, PROSTAGLANDIN SYNTHESIS AND REGULATION, ANGIOTENSIN II MEDIATED ACTIVATION OF JNK PATHWAY VIA PYK2 SIGNALING
327	30.42	4	CROSS-PRESENTATION OF PHAGOSOMES, LEUKOCYTE MIGRATION, CLASS I MHC ANTIGEN PROCESSING & PRESENTATION, IMMUNE SYSTEM
379	41.04	9	REGULATION OF IGF, SENESCENCE AND AUTOPHAGY, P53 SIGNALING, MYOMETRIAL RELAXATION AND CONTRACTION PATHWAYS, TRANSCRIPTIONAL MISREGULATION IN CANCER
516	30.42	4	NA
Hepatocellular carcinoma			
650	32.48	23	TIGHT JUNCTION INTERACTIONS, CELL-CELL JUNCTION ORGANIZATION, TRANSENDOTHELIAL MIGRATION, CELL ADHESION MOLECULES (CAMS), CELL-CELL COMMUNICATION, HEPATITIS C
647	21.37	6	CGMP EFFECTS, NITRIC OXIDE STIMULATES GUANYLATE CYCLASE, POTASSIUM CHANNELS
583	60.30	15	METABOLISM OF XENOBIOTICS BY CYTOCHROME P450, ARACHIDONIC ACID METABOLISM, RETINOL METABOLISM, CHEMICAL CARCINOGENESIS, LINOLEIC ACID METABOLISM, TAMOXIFEN METABOLISM, BENZO(A)PYRENE METABOLISM, SEROTONERGIC SYNAPSE, TRYPTOPHAN METABOLISM, STEROID HORMONE BIOSYNTHESIS, FATTY ACID METABOLISM, PPAR SIGNALING
582	30.01	42	PURINE METABOLISM, CGMP EFFECTS, MORPHINE ADDICTION, NITRIC OXIDE STIMULATES GUANYLATE CYCLASE, PLATELET HOMEOSTASIS, DARPP-32 EVENTS, SIGNALING BY GPCR, ALANINE, ASPARTATE AND GLUTAMATE METABOLISM, PYRIMIDINE METABOLISM, CALMODULIN INDUCED EVENTS, DAG AND IP3 SIGNALING, PLC-GAMMA1 SIGNALING
567	35.92	16	SIGNALING BY SCF-KIT, PI3K/AKT SIGNALING IN CANCER, GAB1 SIGNALOSOME, GLIOBLASTOMA, DAP12 SIGNALING, FGFR SIGNALING, BCR SIGNALING, SIGNALING BY EGFR IN CANCER, DAP12 INTERACTIONS, PDGF SIGNALING, SIGNALING BY NGF, PATHWAYS IN CANCER, REGULATION OF ACTIN CYTOSKELETON, PROSTATE CANCER, FOCAL ADHESION, CALCIUM SIGNALING, IMMUNE SYSTEM, SHC-MEDIATED CASCADE, IRS SIGNALING, CYTOKINE-CYTOKINE RECEPTOR INTERACTION, FRS2-MEDIATED CASCADE, ERBB SIGNALING, ADHERENS JUNCTION, ANGIOGENESIS, SEMAPHORIN INTERACTIONS, HELICOBACTER PYLORI INFECTION
398	27.97	15	TCR SIGNALING, BCR SIGNALING, IMMUNE SYSTEM, PRIMARY IMMUNODEFICIENCY, GENERATION OF SECOND MESSENGER MOLECULES, T-CELL APOPTOSIS, NK CELL CYTOTOXICITY, NF-KAPPA B SIGNALING
389	29.48	11	MITOTIC G1-G1/S, SCF(SKP2)-MEDIATED DEGRADATION OF P27/P21, CELL CYCLE, VIRAL CARCINOGENESIS, MIRNA REGULATION OF DDR, E2F1 DESTRUCTION, SMALL CELL LUNG CANCER, DNA REPLICATION, INTEGRATED CANCER PATHWAY, REGULATION OF APC/C ACTIVATORS, TSH SIGNALING, P53 SIGNALING, ID SIGNALING, PROSTATE CANCER, ATR RESPONSE TO REPLICATION STRESS, PI3K-AKT SIGNALING
348	22.69	7	NEUROTROPHIN SIGNALING, AXONAL GROWTH, CERAMIDE SIGNALING, SIGNALLING BY NGF, P75 NTR RECEPTOR SIGNALING, NF-KB SIGNALS SURVIVAL, NRIF SIGNALS CELL DEATH, ARMS-MEDIATED ACTIVATION, SIGNALLING TO ERKS, VASOPRESSIN-REGULATED WATER REABSORPTION, CELL DEATH SIGNALLING VIA NRAGE, NRIF AND NADE, RHO GTPASE CYCLE, MAPK SIGNALING
343	21.37	6	ARACHIDONIC ACID METABOLISM, EICOSANOID METABOLISM, PROSTAGLANDIN SYNTHESIS AND REGULATION, PROSTANOID METABOLISM
318	29.52	10	FRUCTOSE AND MANNOSE METABOLISM, PENTOSE PHOSPHATE PATHWAY, GLYCOLYSIS AND GLUCONEOGENESIS, GALACTOSE METABOLISM, INSULIN SIGNALING
309	22.69	7	ECM-RECEPTOR INTERACTION, FOCAL ADHESION, INFLAMMATORY RESPONSE, PI3K-AKT SIGNALING, COLLAGEN BIOSYNTHESIS, PLATELET ACTIVATION, NCAM1 INTERACTIONS, SIGNALING BY PDGF, SMALL CELL LUNG CANCER, INTEGRIN CELL SURFACE INTERACTIONS, PROTHROMBIN ACTIVATION, GPVI-MEDIATED CASCADE, PATHWAYS IN CANCER
232	19.51	5	SUMOYLATION AS A MECHANISM TO MODULATE CTBP-DEPENDENT GENE RESPONSES, TGF BETA SIGNALING, CHRONIC MYELOID LEUKEMIA, PATHWAYS IN CANCER
200	27.25	9	CALCIUM SIGNALING, GASTRIN-CREB SIGNALING VIA PKC/MAPK, GPCR SIGNALING, EICOSANOID METABOLISM
186	62.95	47	STEROID HORMONE BIOSYNTHESIS, RETINOL METABOLISM, METABOLISM OF XENOBIOTICS BY CYTOCHROME P450, CHEMICAL CARCINOGENESIS, GLUCURONIDATION, ASCORBATE AND ALDARATE METABOLISM, PORPHYRIN AND CHLOROPHYLL METABOLISM, STARCH AND SUCROSE METABOLISM, HEME DEGRADATION, TAMOXIFEN METABOLISM, THYROID HORMONE METABOLISM, OXIDATIVE STRESS INDUCED GENE EXPRESSION VIA NRF2, ESTROGEN METABOLISM, NICOTINE METABOLISM, CODEINE/MORPHINE METABOLISM, FATTY ACID OMEGA OXIDATION, AHR PATHWAY, ARYLAMINE METABOLISM, FLUOROPYRIMIDINE ACTIVITY, AFLATOXIN B1 METABOLISM, BENZO(A)PYRENE METABOLISM, TRYPTOPHAN METABOLISM, IL-10 SIGNALING
92	69.28	34	JAK-STAT SIGNALING, CYTOKINE SIGNALING, INTERLEUKIN SIGNALING, TSLP SIGNALING, IMMUNE SYSTEM, GHR SIGNALING, EPO RECEPTOR SIGNALING, PI3K-AKT

			SIGNALING, INFLAMMATORY RESPONSE, PROLACTIN SIGNALING, TPO SIGNALING, LEPTIN SIGNALING, FGFR SIGNALING, KIT RECEPTOR SIGNALING, INHIBITION OF CELLULAR PROLIFERATION BY GLEEVEC, AGE-RAGE PATHWAY, ERBB4 SIGNALING, PDGF SIGNALING, VIRAL CARCINOGENESIS, ACUTE MYELOID LEUKEMIA, HEPATITIS C, EGF-EGFR SIGNALING, CHEMOKINE SIGNALING, ERBB2 IN SIGNAL TRANSDUCTION AND ONCOLOGY
44	35.33	38	INTEGRATED BC PATHWAY, INTEGRATED PANCREATIC CANCER PATHWAY, COLORECTAL CANCER, PROSTATE CANCER, INTEGRATED CANCER PATHWAY, ANDROGEN RECEPTOR SIGNALING, MAPK SIGNALING, DNA DAMAGE RESPONSE, P53 SIGNALING, CELL CYCLE, CASPASE-MEDIATED CLEAVAGE OF CYTOSKELETAL PROTEINS, APOPTOSIS, TNF ALPHA SIGNALING, PATHWAYS IN CANCER, MIRNA REGULATION OF DDR, ATM SIGNALING, APOPTOSIS MODULATION BY HSP70, ERBB SIGNALING, ARF INHIBITS RIBOSOMAL BIOGENESIS, ADHERENS JUNCTION, SMAC-MEDIATED APOPTOTIS, WNT SIGNALING, P75 NTR MEDIATED SIGNALING, BTG FAMILY PROTEINS AND CELL CYCLE REGULATION, AUTODEGRADATION OF COP1, SENESCENCE AND AUTOPHAGY, TGF BETA SIGNALING, VIRAL CARCINOGENESIS, RB CELL SURVIVAL PATHWAY, PI3K-AKT SIGNALING, TELOMERASE CELLULAR AGING AND IMMORTALITY, NOTCH SIGNALING, CHROMATIN REMODELING, CELL DEATH SIGNALING VIA NRAGE, NRIF AND NADE, TNFR1 SIGNALING, INTERNAL RIBOSOME ENTRY PATHWAY, FAS SIGNALING (CD95), SIGNALING BY HIPPO, SIGNALING BY NGF, APOPTOSIS THROUGH DR3 AND DR4/5, CASPASE CASCADE IN APOPTOSIS, SIGNALING BY EGFR IN CANCER, TRANSCRIPTIONAL MISREGULATION IN CANCER, NF-KAPPA B SIGNALING, TRANSCRIPTIONAL ACTIVITY OF SMAD2/SMAD3:SMAD4, DNA REPAIR, SIGNAL TRANSDUCTION BY L1, TWEAK SIGNALING, MRNA PROCESSING
18	17.94	4	WNT SIGNALING

5.3.3: Breast Cancer

The optimal step size for clustering of BC data was reached with the maximal cluster size, at 4175 steps. BC clusters described in Table 8 include all modules significantly enriched with miRNA targets and these modules are presented in Appendix B and in high resolution as Supplementary Files. The network includes 5607 singletons, 457 pairs, 149 triplets and 326 communities with greater than three nodes. For the nine significant miRNA-enriched modules, I visualized and evaluated these clusters and present the most relevant, interpretable clusters. Module 379 (Figure 25) includes interactions between various growth factors, a number of which are associated with cancer, and the oncogene *NOV*. In this module, *IGFI* interacts with several binding proteins and the oncogene *NOV,* and is a target of the highly differentially regulated miRNA *miR-33b*. This growth factor is involved in growth and proliferation signaling and is suspected to alter cancer risk[190]. *IGFBP* are *IGF* binding proteins help increase the half-life of *IGF* and target it to specific tissues. *IGFBP7*, *IGFBP5*, *IGFBP3* appear to function as tumor suppressors [210-213] and are associated with growth pathways and apoptosis in breast carcinomas. Conversely, *IGFBP*'s can also increase risk due to proliferative activity.

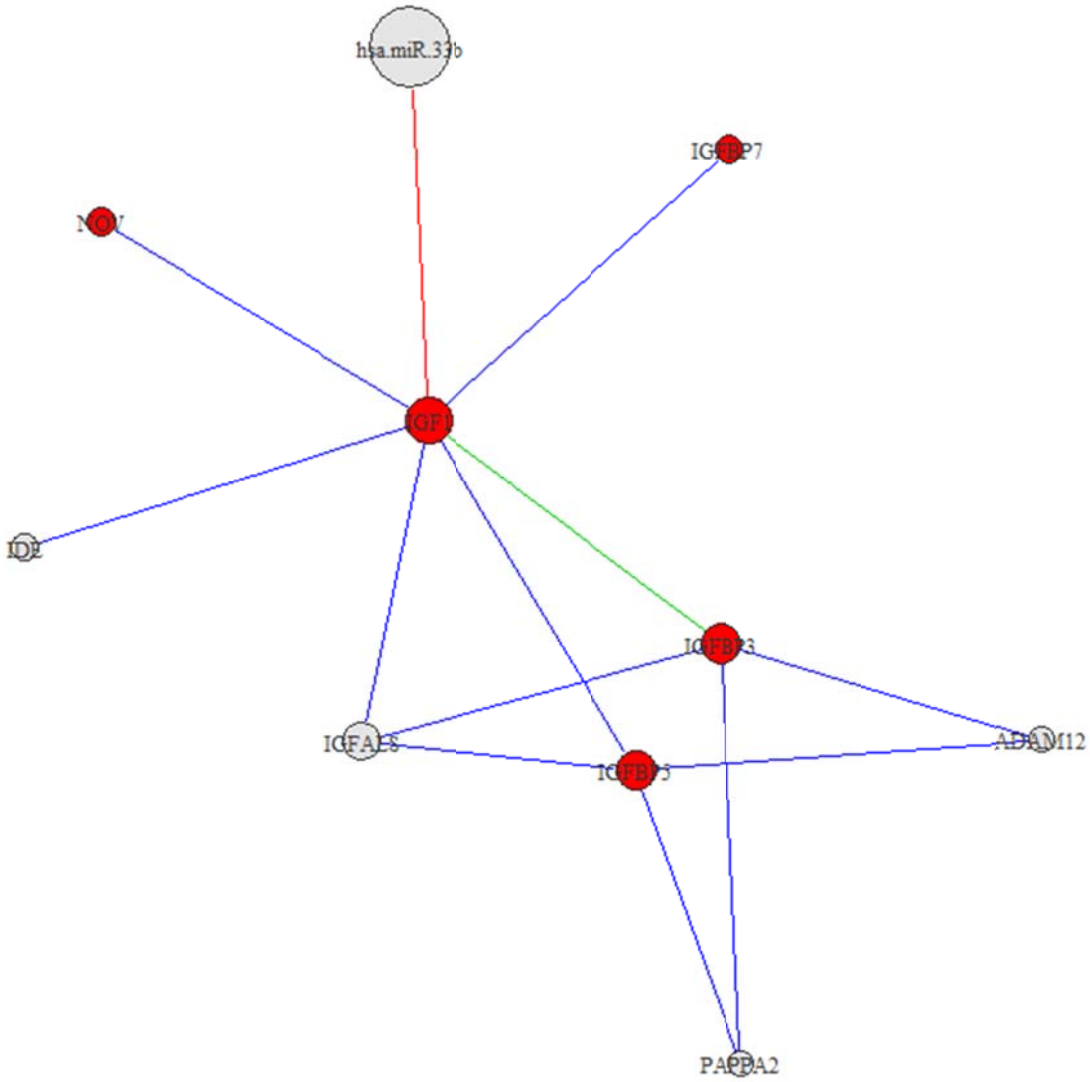


Figure 25: BC Network Module 379. Module 379 shows *IGF* receptor interaction with *IGFBP* genes and the *NOV* oncogene. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, green from KEGG, orange from both databases and red designate miRNA-mRNA interactions.

Module 74 (Figure 26) consists of interactions among several oncogenes and overlaps with growth signaling and transcription pathways. Multiple genes in this cluster have been identified as therapeutic targets for BC. *YES1*[214] (a target of *miR-770-5p*) *KIT*[215,216] and *CSF1R*[217] have been shown be potential therapeutic targets for BC survival. The associated *MATK* tyrosine kinase gene was also found to be expressed in BC cells, but not neighboring normal cells[218]. Other genes in this module are related to growth, proliferation and motility, including *PI3K/AKT*, *MAPK-JNK*, pathways involved in key cancer processes. *EPS15* is a

substrate for *EGFR* and can promote *PI3K*/*AKT* signaling[219]. These genes interact closely with *SRC* homology-3 intercellular proteins which mediate protein-protein interactions and signaling for diverse functions, such as growth signaling, motility, cell-polarization and transcription regulation. *ITCH,* ubiquitin ligase gene, has not been well-described as a cancer gene, but plays an important role in transcriptional activation, and is associated with the *MAPK*, *JNK*, and *JUN* signaling pathways [220].

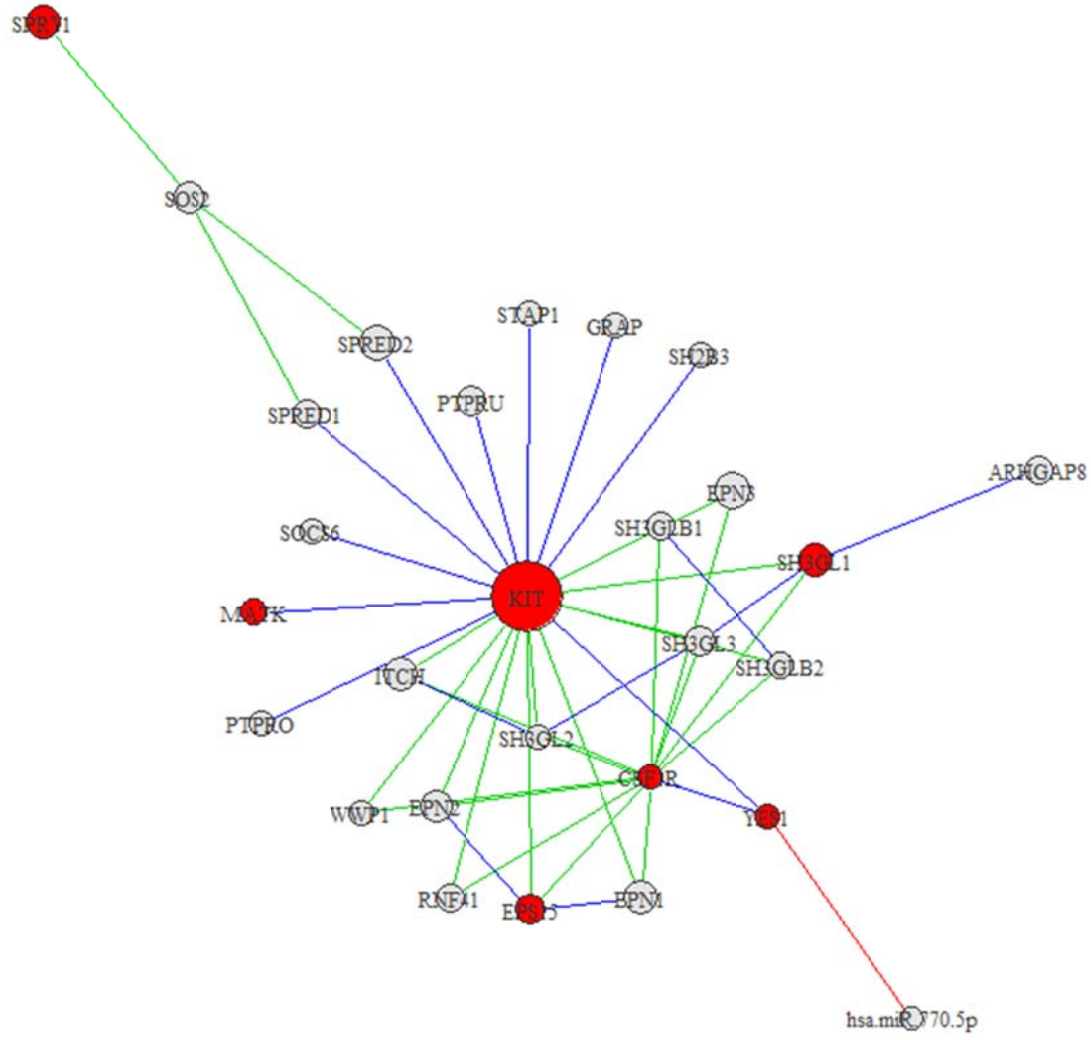


Figure 26: BC Network Module 74. Module 74 shows *KIT* oncogene and interactions with *ITCH*, *MATK* and *YES1*. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, green from KEGG, orange from both databases and red designate miRNA-mRNA interactions.

Module 22 (Figure 27) includes interactions between genes involved in gene fusions and those controlling transcriptional regulation and repression. *ETV6, ZBTB16, RUNX1T1*, *BCOR* and *BCL6* are involved in gene fusions in leukemia[221,222,223,224,225,226]. ETV6 acts as a transcriptional co-repressor with *SIN3A*, *NCOR1*, and *SMRT* [227]. Specifically, *ETV6* may act as a tumor suppressor[228] and is known to form a fusion gene with *RUNX1*[229]. Histone deacetylases, *HDAC*s interact with genes in this complex, as well as cancer associated genes *BCL6*, *NCOR1*, *RUNX1T1* and *SKI* to regulate transcription. *SIN3A* acts as a scaffold for *HDAC* complexes [230] and its activity has been shown to be dependent on association with *HDAC*s to repress *STAT3* transcription[227]. Another member of these complexes, the *SKI* oncogene, plays an important role in the TGFβ-signaling pathway and proliferation [231].

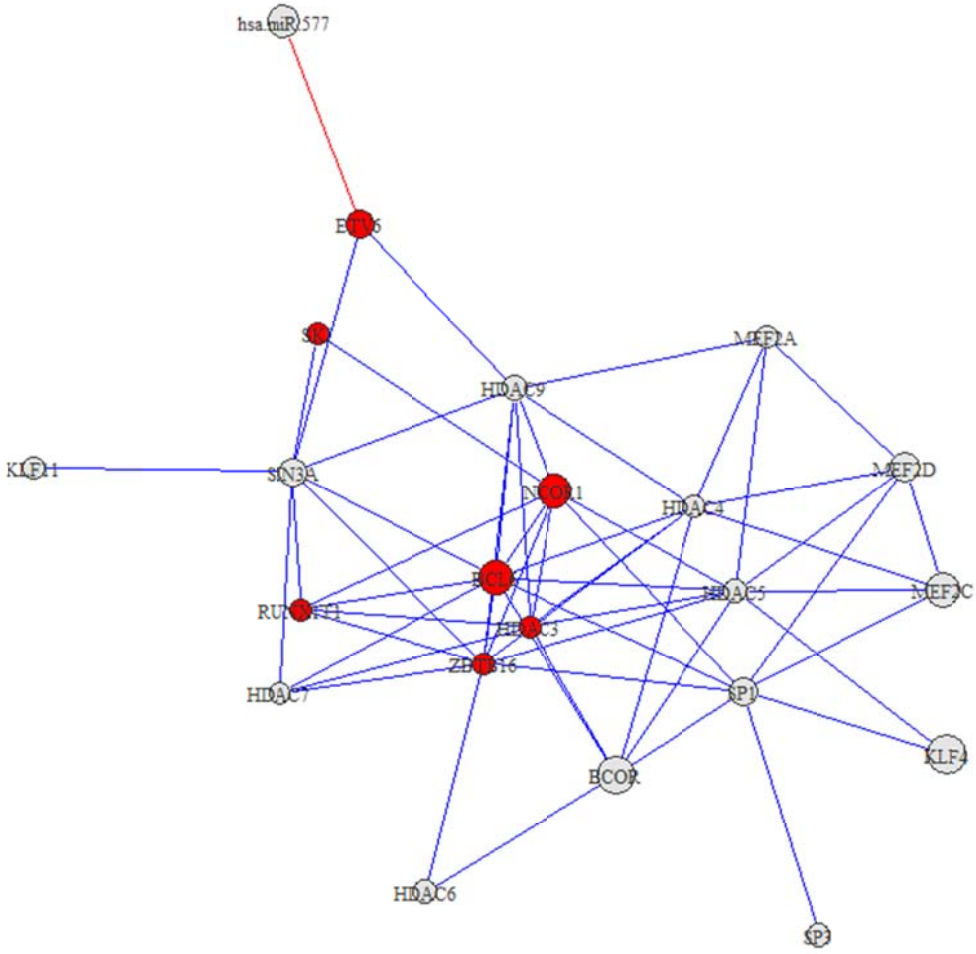


Figure 27: BC Network Module 22 Module 22 shows nteractions among *ETV6*, *BCL6*, *BCOR* and *HDAC* genes. Many of these genes are involved in gene fusions or form a complex with *HDAC* genes in cancer studies. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, green from KEGG, orange from both databases and red designate miRNA-mRNA interactions.

MiRNA *miR-223* intersects three modules, modules 292, 269 and 327 (Figure 28). Module 292 is associated with genes involved in T and B-cell signaling, cell transformation and invasion. *miR-223* is co-regulated with *PTAFR*. *PTFAR* is part of a family of G-protein coupled receptors, including *RGS4,* and is associated with inflammation and cell invasion[232]. *PTAFR* targets include *GNAQ*, *GNA11* and *GNA14*, part of a family of guanine-nucleotide binding cell-surface receptors, and intercellular signaling pathways. There is little literature describing *GNA14*, but other genes in this family, *GNAQ* and *GNA11* have known sites of oncogenic mutations[233]. *ADRAI1B* has the capacity to induce oncogenic transformation in cells and has been described as a protooncogene [234]. *F2R* is involved in the thrombotic response and has been shown to be necessary and sufficient to induce proliferation and invasion in a BC model in mice[235]. Further, *EDNRB* plays a role in allowing cancer cells to evade the T-cell immune response[236].

Module 269 is largely associated with proliferation, cell adhesion, cell-cell communication, and T-cell and B-cell signaling. The pi3-pi4 kinase family of proteins consists of phosphoinositide 3-kinases phosphorylate inositol lipids important in extracellular communication and cellular adhesion. *PIK3CG* is a target of *miR-223*, is involved in cytotoxicity of natural killer cells, and has been found to inhibit growth in tumor cells[237]. *PIK3CD* is a target of *miR-146a* and is involved in proliferation, adhesion and migration of mast cells [238]. In BC, *PIK3CA* affects cancer progression by interacting with *PTEN* and blocking cell-cycle arrest [239]. *FYN* is induced by *Ras-PIK3-AKT* signaling and has been found to be necessary for cancer progression, cell invasion and migration in several cancer types[240]. Various other genes in this module have implications in cancer. The *DOK1*, *DOK2* and *DOK3* genes are involved in transcriptional regulation and proliferation and are associated with tumorigenesis[241]. *VAV1* and *VAV2* are oncogenes involved in development, transcription, angiogenesis and cell signaling[242,243]. *BLNK* is a component of the B cell receptor pathway and acts as a tumor suppressor[244,245]. *MAP4K1* is an upstream activator of signaling pathways, including *MLK3*, *JNK*, *SERK1*, *SAPK*, *MEKK* pathways [246,247]. Interaction among *PTPRC*, *ZAP70*, *LCP2* and *SKAP1* is involved cell-cell communication and cell migration in immune cells and these interactions shown in more detail in BC module 398 (Section 4.3.3, Figure 32).

Genes in module 327 consists of *CYBB* and the *NCF* family of genes which are also involved in the T-cell response. *GNA14* is also a target of a miRNA gene in HCC module 200 (Figure 34).

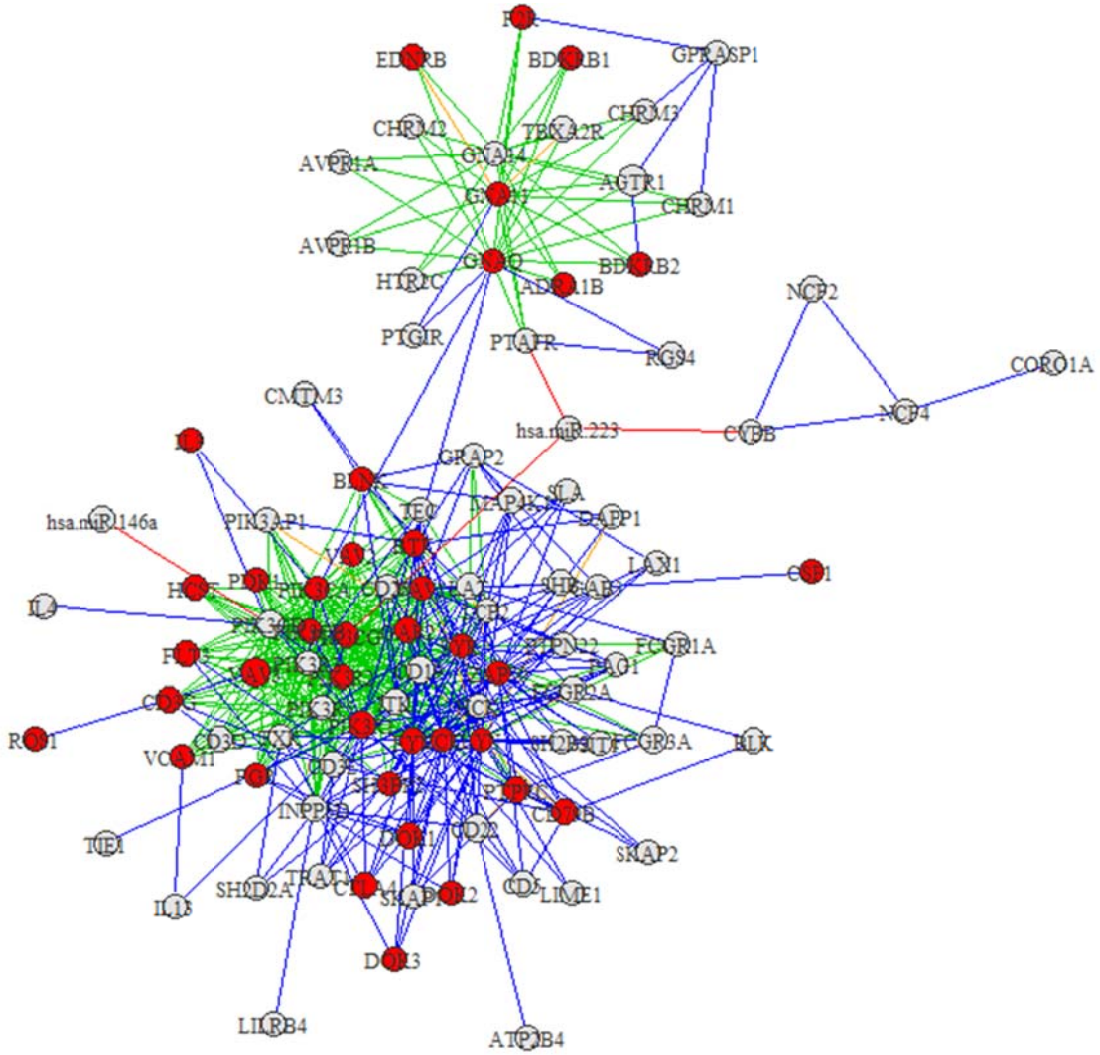


Figure 28: Intersection of BC Network Modules 292 and 269. Module 292 shows an intersection of G-coupling proteins, proliferation, T-cell response and thrombotic response genes. Module 269 consists of oncogenes and *PIK3** genes. Genes in module 327 consists of *CYBB* and the *NCF* family of genes which are also involved in the T-cell response. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, green from KEGG, orange from both databases and red designate miRNA-mRNA interactions.

Table 9: Key Genes described in BC miRNA Modules

BC			
NOV	nephroblastoma overexpressed	379	Protein likely to play a role in cell growth regulation.
IGF1	insulin-like growth factor 1	379	The insulin-like growth factor is structurally and functionally related to insulin but has higher growth-promoting activity.
IGFBP1,-3,-5,-7	insulin-like growth factor binding protein 1,-3,-5,-7	379	IGFBPs are members of the insulin-like growth factor (IGF)-binding protein (IGFBP) family. IGFBPs bind IGFs, regulate IGF availability in body fluids and tissue and modulate IGF binding to its receptors. IGFBPS can inhibit or stimulate the growth promoting effects of the IGFs and modulate cell adhesion and prostacyclin production.
IFGALS	insulin-like growth factor binding protein, acid labile subunit	379	Serum protein that binds insulin-like growth factors, increasing their half-life and their vascular localization. Its production is stimulated by growth hormone and it is involved in receptor-ligand binding and cell adhesion.
YES1	v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1	74	Non-receptor protein tyrosine kinase involved in the regulation of cell growth and survival, apoptosis, cell-cell adhesion, cytoskeleton remodeling, and differentiation. Stimulation by receptor tyrosine kinases (RTKs) including EGRF, PDGFR, CSF1R and FGFR recruits YES1 to the phosphorylated receptor. Regulates the G1 phase, G2/M progression and cytokinesis.
CSF1R	colony stimulating factor 1 receptor	74	Tyrosine-protein kinase that acts as cell-surface receptor for CSF1 and IL34 and plays an essential role in the regulation of survival, proliferation and differentiation of hematopoietic precursor cells. It promotes reorganization of the actin cytoskeleton, regulates cell adhesion and cell migration, and promotes cancer cell invasion. Phosphorylates PIK3R1, PLCG2, GRB2, SLA2 and CBL. Activated CSF1R also mediates activation of the AKT1 signaling pathway, MAP kinases MAPK1/ERK2 and/or MAPK3/ERK1, the SRC family kinases SRC, FYN and YES1 and STAT family members STAT3, STAT5A and/or STAT5B. .
KIT	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog	74	Tyrosine-protein kinase that acts as cell-surface receptor for the cytokine KITLG/SCF and plays an essential role in the regulation of cell survival and proliferation, hematopoiesis, stem cell maintenance, gametogenesis, mast cell development, migration and function, and melanogenesis. Phosphorylates PIK3R1, PLCG1, SH2B2/APS and CBL. Activates the AKT1 signaling pathway by phosphorylation of PIK3R1, the regulatory subunit of phosphatidylinositol 3-kinase. Activated KIT also transmits signals via GRB2 and activation of RAS, RAF1 and the MAP kinases MAPK1/ERK2 and/or MAPK3/ERK1. Promotes activation of STAT family members STAT1, STAT3, STAT5A and STAT5B. Mutations in this gene are associated with gastrointestinal stromal tumors, mast cell disease, acute myelogenous lukemia, and piebaldism.
MATK	megakaryocyte-associated tyrosine kinase	74	Plays an important role in signal transduction of hematopoietic cells, regulates tyrosine kinase activity of SRC-family members and plays an inhibitory role in the control of T-cell proliferation. May be involved in some cases of breast cancer.
EPS15	epidermal growth factor receptor pathway substrate 15	74	Protein is present at clatherin-coated pits and is involved in receptor-mediated endocytosis of EGF. This gene is rearranged with the HRX/ALL/MLL gene in acutemyelogeneous leukemias. Involved in cell growth regulation, regulation of mitogenic signals and control of cell proliferation.
ITCH	itchy E3 ubiquitin protein ligase	74	A member of the Nedd4 family of HECT domain E3 ubiquitin ligases that plays a role in erythroid and lymphoid cell differentiation and the regulation of immune responses. Mediates antiapoptotic activity of EGFR through ubiquitination and proteasomal degradation of p15 BID.
SH3GL1-3	SH3-domain GRB2-like 1, -2, -3	74	Implicated in endocytosis. May recruit other proteins to membranes with high curvature.
ETV6	ets variant 6	22	Transcriptional repressor important in hematopoiesis and maintenance of the developing vascular network. Involved in a large number of chromosomal rearrangements associated with leukemia and congenital fibrosarcoma.
BCL6	B-cell CLL/lymphoma 6	22	A zinc finger transcription factor and contains an N-terminal POZ domain. This protein acts as a sequence-specific repressor of transcription, and has been shown to modulate the transcription of START-dependent IL-4 responses of B cells.

RUNX1T1	runt-related transcription factor 1; translocated to, 1	22	A member of the myeloid translocation gene family which binds to histone deacetylases interacts with DNA-bound transcription factors to facilitate transcriptional repression. The t(8;21)(q22;q22) translocation is one of the most frequent karyotypic abnormalities in acute myeloid leukemia. Can repress transactivation mediated by TCF12.
HDAC3, -4,-5,-9	histone deacetylase 3, -4, -5,-9	22	Histone Deacetylases (HDACs) are a group of enzymes that catalyze the removal of acetyl groups from lysine residues in histones and non-histone proteins, which alters chromosome structure and affects transcription factor access to DNA. HDAC3 can also down-regulate p53 function and thus modulate cell growth and apoptosis and it is regarded as a potential tumor suppressor gene. HDACs play a critical role in transcriptional regulation, cell cycle progression, cell growth arrest, cell differentiation and death and this has led to substantial interest in HDAC inhibitors as possible antineoplastic agents.
SIN3A	SIN3 transcription regulator homolog A	22	A transcriptional repressor that interacts with MXI1 to repress MYC responsive genes and antagonize MYC oncogenic activities. Can repress transcription by tethering SIN3A to DNA, and in parallel with histone deacetylation.
NCOR1	nuclear receptor corepressor 1	22	Mediates transcriptional repression by acting as part of a complex which promotes histone deacetylation and the formation of repressive chromatin structures which may impede the access of basal transcription factors
ZBTB16	zinc finger and BTB domain containing 16	22	A member of the Krueppel C2H2-type zinc-finger protein family and encodes a zinc finger transcription factor that is involved in cell cycle progression, and interacts with a histone deacetylase. Instances of gene rearrangement at this locus have been associated with acute promyelocytic leukemia (APL)
MEF1D	myocyte enhancer factor 2D	22	Transcriptional activator which binds specifically to MEF2. Plays diverse roles in the control of cell growth, survival and apoptosis via p38 MAPK signaling in muscle-specific and/or growth factor-related transcription
KLF4	Kruppel-like factor 4	22	Transcription factor that plays an important role in maintaining embryonic stem cells. Involved in cellular differentiation of epithelial contributes to the down-regulation of p53/TP53 transcription
MEF1C	myocyte enhancer factor 2C	22	Transcription activator which binds specifically to MEF2 element in the regulatory regions of many muscle-specific genes. Controls cardiac morphogenesis and myogenesis, and is involved in vascular development.
PTAFR	platelet-activating factor receptor	292	A G-protein-coupled receptor for platelet-activating factor (PAF). PAF is a phospholipid that plays a significant role in oncogenic transformation, tumor growth, angiogenesis, metastasis, and pro-inflammatory processes. Binding of PAF to the PAF-receptor (PAFR) stimulates numerous signal transduction pathways including phospholipase C, D, A2, mitogen-activated protein kinases (MAPKs), and the phosphatidylinositol-calcium second messenger system.
RGS4	regulator of G-protein signaling 4	292	Inhibits signal transduction by increasing the GTPase activity of G protein alpha subunits thereby driving them into their inactive GDP-bound form.
GNA11,-14, -Q	guanine nucleotide binding protein (G protein), alpha -11 -14, -Q	292	Guanine nucleotide-binding proteins (G proteins) are involved as modulators or transducers in various transmembrane signaling systems
ADRAI1B	adrenoceptor alpha 1B	292	Alpha-1-adrenergic receptors (alpha-1-ARs) are members of the G protein-coupled receptor superfamily. They activate mitogenic responses and regulate growth and proliferation of many cells. There are 3 alpha-1-AR subtypes: alpha-1A, -1B and -1D, all of which signal through the Gq/11 family of G-proteins and different subtypes show different patterns of activation. This gene encodes alpha-1B-adrenergic receptor, which induces neoplastic transformation when transfected into NIH 3T3 fibroblasts and other cell lines. Thus, this normal cellular gene is identified as a protooncogene.
PTGIR	prostaglandin I2 (prostacyclin) receptor	292	The protein encoded by this gene is a member of the G-protein coupled receptor family 1 and has been shown to be a receptor for prostacyclin. Prostacyclin, the major product of cyclooxygenase in macrovascular endothelium, elicits a potent vasodilation and inhibition of platelet aggregation through binding to this receptor.
TBXA2R	thromboxane A2 receptor	292	This gene encodes a member of the G protein-coupled receptor family. The protein interacts with thromboxane A2 to induce platelet aggregation and regulate hemostasis.

AGTR1	angiotensin II receptor, type 1	292	Angiotensin II is a potent vasopressor hormone and a primary regulator of aldosterone secretion. It is an important effector controlling blood pressure and volume in the cardiovascular system.
EDNRB	endothelin receptor type B	292	The endothelinB receptor (ETB receptor) is a member of the endothelin receptor group of G-protein-coupled receptors located primarily in vascular endothelial cells where they play a role in vasoconstriction, vasodilation, bronchoconstriction and cell proliferation.
CYBB	cytochrome b-245, beta polypeptide	327	Critical component of the membrane-bound oxidase of phagocytes that generates superoxide. Also functions as a voltage-gated proton channel that mediates the H(+) currents of resting phagocytes. Participates in the regulation of cellular pH.
PIK3CA,-D,-G	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit gamma	269	PI 3-Kinases (phosphoinositide 3-kinases, PI3Ks) are family of lipid kinases capable of phosphorylating the 3'OH of the inositol ring of phosphoinositides. They are responsible for coordinating a diverse range of cell functions including proliferation, cell survival, degranulation, vesicular trafficking and cell migration.
FCGR2A, -3A	Fc fragment of IgG, low affinity IIIa, receptors	269	Receptor for the Fc region of IgG that mediates antibody-dependent cellular cytotoxicity (ADCC) and other antibody-dependent responses, such as phagocytosis
PTEN	phosphatase and tensin homolog	269	Tumor suppressor that acts as a lipid phosphatase and as a dual-specificity protein phosphatase, dephosphorylating tyrosine-, serine- and threonine-phosphorylated proteins. Antagonizes the PI3K-AKT/PKB signaling thereby modulating cell cycle progression and cell survival. Dephosphorylates tyrosine-phosphorylated focal adhesion kinase and inhibits cell migration and integrin-mediated cell spreading.
FYN	FYN oncogene related to SRC, FGR, YES	269	A member of the protein-tyrosine kinase oncogene family implicated in the control of cell growth. Involved in the regulation of cell adhesion and motility through phosphorylation of CTNNB1 (beta-catenin) and CTNND1 (delta-catenin).
DOK1, -2, -3	docking protein 2, 56kDa	269	DOK proteins are enzymatically inert adaptor or scaffolding proteins that provide a docking platform for the assembly of multimolecular signaling complexes. DOK1 is a negative regulator of the insulin signaling and integrin activation DOK2 may modulate the cellular proliferation induced by IL-4, as well as IL-2 and IL-3, modulating Bcr-Abl signaling and EGF-stimulated MAP kinase activation. DOK3 is a negative regulator of JNK signaling and may modulate ABL1.
VAV3	vav 2 guanine nucleotide exchange factor	269	Guanine nucleotide exchange factor for the Rho family of Ras-related GTPases. Its recruitment by EPHA2 is critical for EFNA1-induced RAC1 GTPase activation and vascular endothelial cell migration and assembly. Important in angiogenesis.
PTPRC	protein tyrosine phosphatase, receptor type, C	269,	Protein tyrosine phosphatase that regulates a variety of cellular processes including cell growth, differentiation, mitosis, and oncogenic transformation. PTPRC has been shown to be key regulator of T- and B-cell antigen receptor signaling. Upon T-cell activation, recruits and dephosphorylates SKAP1 and FYN. Dephosphorylates and modulates LYN activity.
ZAP70	zeta-chain (TCR) associated protein kinase 70kDa	269,	Protein tyrosine kinase that plays a role in T-cell development and lymphocyte activation. This enzyme functions in the initial step of TCR-mediated signal transduction in combination with the Src family kinases, Lck and Fyn.
SKAP2	src kinase associated phosphoprotein 2	269,	A src family kinase which is an adaptor protein that is thought to play an essential role in the src signaling pathway
BLNK	B-cell linker	269,	Cytoplasmic linker or adaptor protein that plays a critical role in B cell function and development and bridges SYK kinase to a multitude of signaling pathways. Deficiency in this protein has also been shown in pre-B acute lymphoblastic leukemia. Plays a role in the activation of ERK/EPHB2, MAP kinase p38 and JNK, AP1, NF-kappa-B and NFAT, PLCG1, PLCG2 and Ca(2+).
MAP4K1	mitogen-activated protein kinase kinase kinase kinase 1	269	May play a role in the response to environmental stress. Appears to act upstream of the JUN N-terminal pathway. May play a role in hematopoietic lineage decisions and growth regulation
BLK	B lymphoid tyrosine kinase	269	Nonreceptor tyrosine-kinase of the src family of proto-oncogenes that are typically involved in cell proliferation and differentiation. Plays a role in B-cell receptor signaling and B-cell development. It also stimulates insulin synthesis and secretion in response to glucose and enhances the expression of several pancreatic beta-cell transcription factors.

5.3.4: Hepatocellular Carcinoma

The optimal step size for the clustering of breast cancer data was reached with the maximal size, at 2776 steps. HCC clusters in Table 8 include all modules significantly enriched with miRNA targets and these modules are presented in Appendix B and in high resolution as Supplementary Files. The network includes 7228 singletons, 394 pairs, 136 triplets and 223 communities wuth greater than three nodes. For the seventeen significant miRNA-enriched modules, I visualized and evaluated these clusters and present a number of the most relevant, interpretable clusters. Module 309 (Figure 29) describes a community of regulatory genes that are associated with extracellular matrix organization and focal adhesion. *THBS1*, *FN1* and *SPARC* are differentially expressed and interact with collagen genes to maintain focal adhesion and cellular organization. *THBS1* also is a potent anti-angiogenic regulator[248] and exhibits coordinated activity with *miR-346*. *THBS1* [248,249], *FN1* [250] *SPARC* [251], *COL1A1* [252] which are known to be involved in cancer [248,249,253].

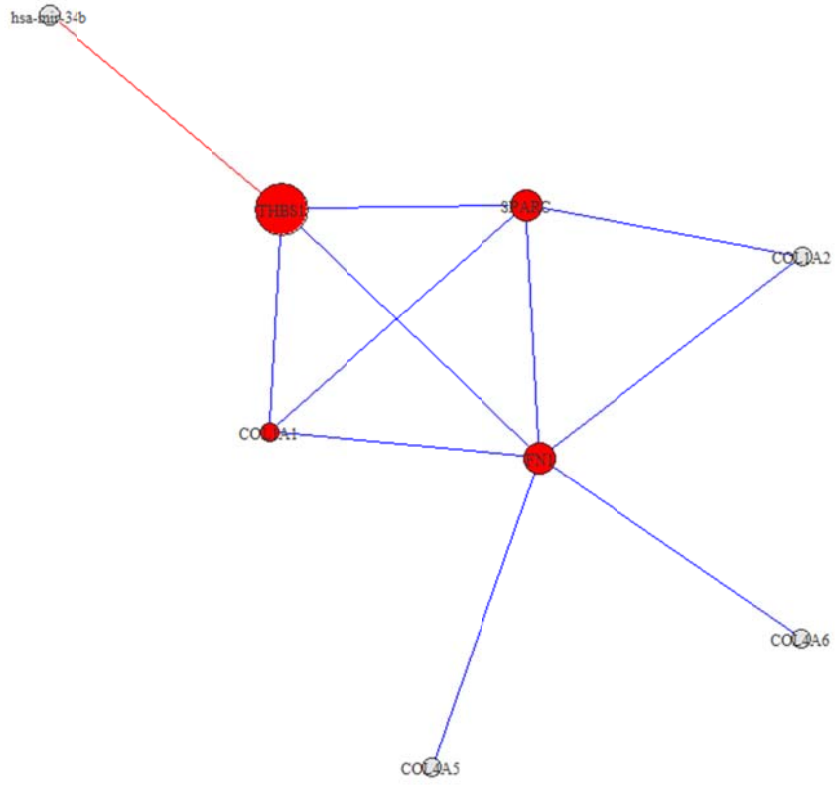


Figure 29: HCC Network Module 309. Module 309 shows *THBS1* interactions with *SPARC*, *FN1* and collagen genes, primarily involved in cell-cell interactions and cell adhesion. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, green from KEGG, orange from both databases and red designate miRNA-mRNA interactions.

Module 567 (Figure 30) consists of interactions among genes involved in proliferation and differentiation. The module consists of cancer-related genes including *PDGFRA*, *MET*, *ErbB2*, *ErbB3*, *ErbB4*, *FGFR2/3*, *RET* and *KDR* which are key genes in growth and proliferation signaling. *PDGFRA,* which is a target of *miR-184,* is highly differentially regulated and has been shown to cause tumorigenesis in gliomas[254]. *MET* is implicated in HCC by hindering apoptosis in liver cells [255] and dysregulated *SH3GL2* has been associated with HCC via downregulation *miR-330*[256]. These genes are further linked to tumor progression and angiogenesis via the *KDR* gene (*VEGF* receptor). *SH3GL2* gene acts as a hub in this module, but the role this gene plays in cancer is not well-understood, making the gene a good candidate for further study.

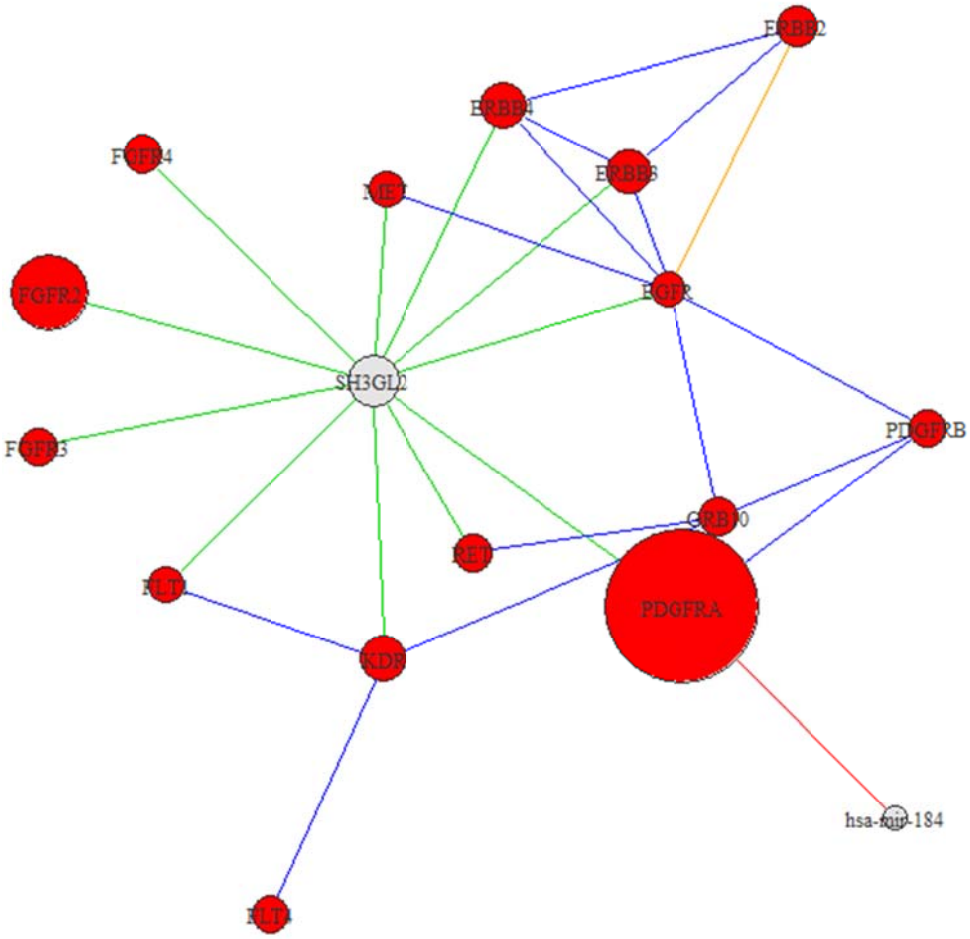


Figure 30: HCC Netwok Module 567. Module 567 shows *PDGFRA* interactions with *MET* and *ErbB** proliferation genes, *RET*, *KDR* and *FGF** genes. *SH3GL2* is a hub in this module of cancer genes, but its role in cancer is not well-studies. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, green from KEGG, orange from both databases and red designate miRNA-mRNA interactions.

Module 232 (Figure 31) consists of interactions between *ZEB2* and *ZEB1*, *CTBP1* and *TGFβ*-signaling gene *TGIF1*. *ZEB1/2* mediate *TGFβ*-signaling via interactions with *SMAD* genes and have been shown to control expression of the cell-adhesion regulatory gene *e-cadherin*. The *ZEB1/ZEB2* genes are associated with *miR-200* family members[106], increasing *miR200* levels induced mesenchymal-to-epithelial transition (*MET*) in human cancer cell lines by targeting *ZEB1* and *ZEB2*; conversely, reducing *miR-200* levels compromised epithelial-to-mesenchymal transition (EMT), and enhanced tumor progression[257]. Further, *CTBP1,* a co-repressor with *ZEB1*/*2*, binds to the C-terminal region of the *EA1* protein, which is an important site mediating the oncogenic activity of adenoviruses. *miR-137* has been found to interact with *CTBP1* in melanoma cells to increase expression of *CTBP1* target genes, *e-cadherin* and *BAX*, which play a role in cell migration and DNA repair [258].

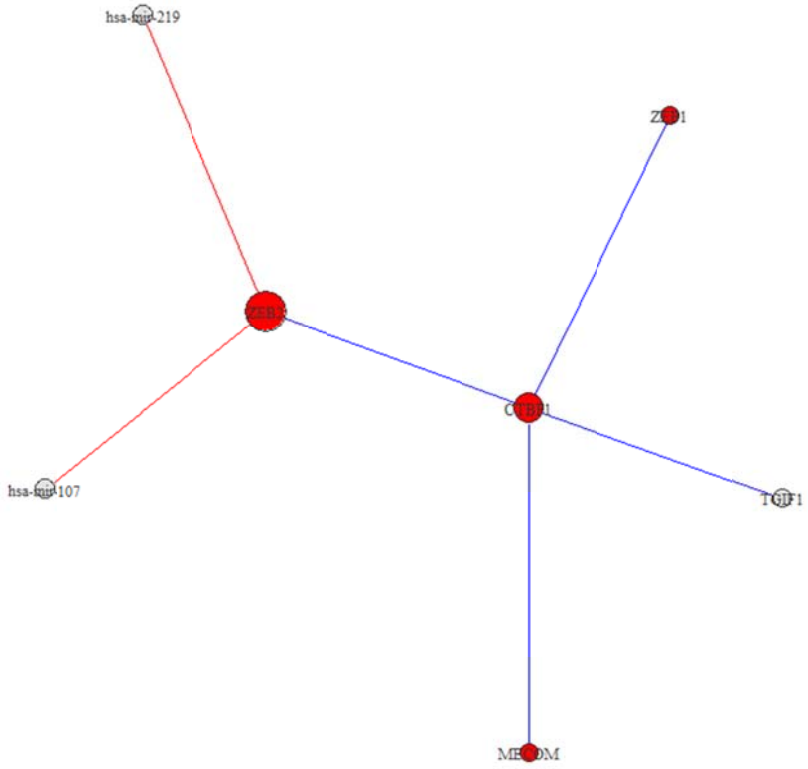


Figure 31: HCC Network Module 232. Module 232 shows nteractions between *ZEB* genes and co-repressor *CTBP1*. *ZEB* genes are known to be associated with several miRNAs in cancer, they are implicated in dysregulated cell adhesion, and here they interact with other known cancer genes, *CTBP1* and *MECOM*. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, green from KEGG, orange from both databases and red designate miRNA-mRNA interactions.

Module 44 interacts with module 398 via *miR-218* (Figure 32). Module 44 shows interaction among genes involved in T-cell signaling, histone and chromatin modification, and tumor suppressor genes including *TP53*, *BRCA1* and *FHL*. *FHL* is targeted by *miR-218*, associates with the *BRCA1* tumor suppressor[259], and has been shown to be associated with anti-proliferation and anti-apoptotic effects in liver cancer[260]. *PBK* is commonly upregulated in BC and thought to influence tumor progression via histone modification[261].

In module 398, *miR-218* interacts with protein tyrosine kinase *PTPRC* (*CD45*), which is known to mediate *JAK/STAT* signalling and is a genetic risk factor for hepatitis C infections[262,263], which lead to higher incidence of HCC[264]. Together, *PTPRC, ZAP70, LCP2* and *SKAP1* form a cluster of key genes in T-cell signaling, cell-cell communication and cellular adhesion[265],[266,267]. The role of *SKAP1* in HCC is not well documented and this gene is an interesting candidate for further study. *SMARCD1*, which is involved in chromatin remodeling, is also shown to be associated with HCC in the Burchard study[11]. Module 398 overlaps with *PTPRC*, *ZAP70* and *LCP* interactions in BC module 269 (Figure 28).

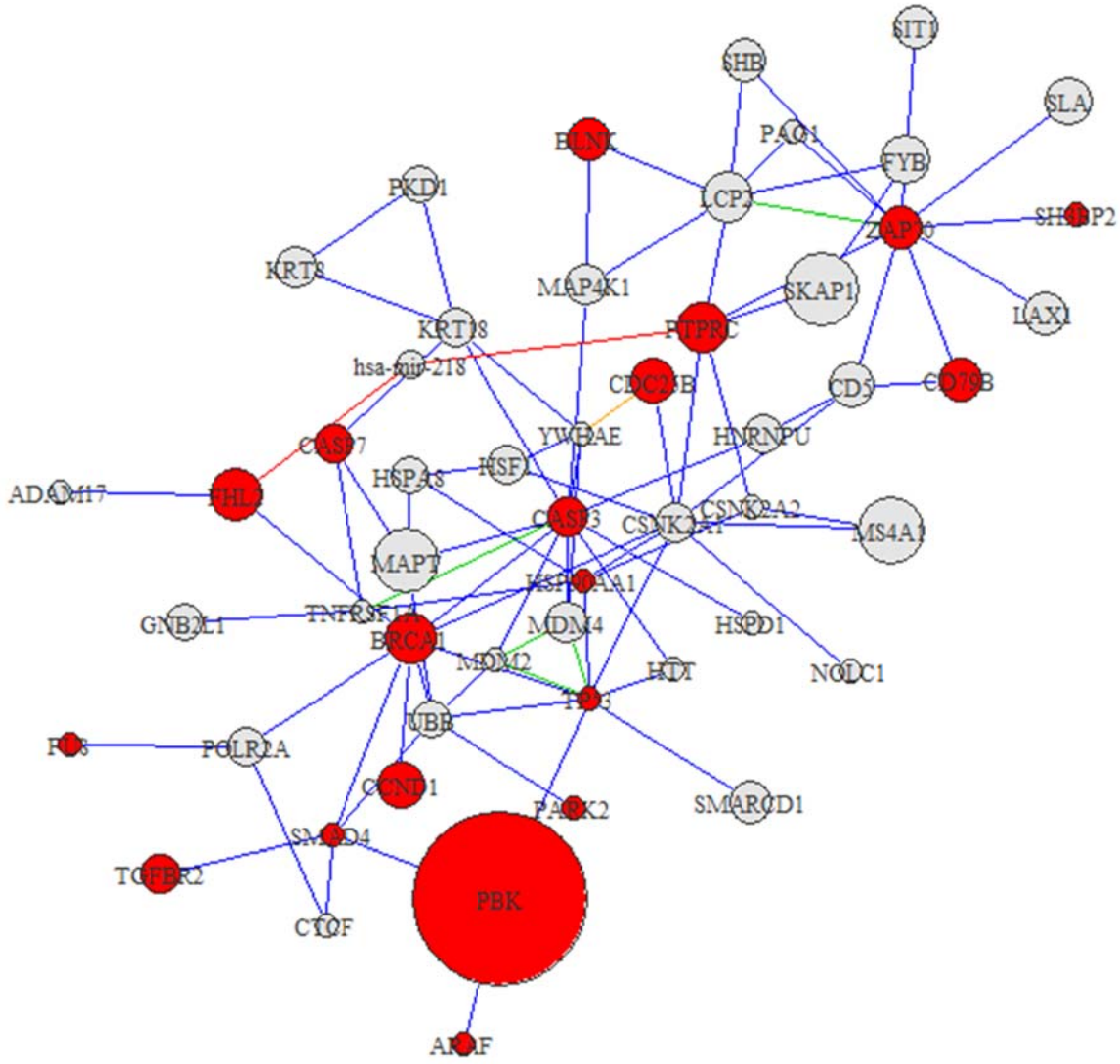


Figure 32: Intersection of HCC Network Modules 44 and 398. Module 44 shows *PBK* and interactions with tumor repressors *TP53*, *BRCA1*, and interactions with, cell-cycle control genes and *SMARCD1*, also highlighted in the Burchard study. Module 398-*ZAP70* and interactions with genes involved in B-cell and T-cell signaling. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, green from KEGG, orange from both databases and red designate miRNA-mRNA interactions.

Module 389 (Figure 33) presents interactions among key cyclins associated with cancer, including *CDK2*, *CCNE1* and *CCNA2*, which are critical regulators of cell cycle control. Altered expression of *CDC25B* [268], *CDC25A*[269], and *SKP2*[270,271] play an important role in cell cycle progression and oncogenesis. *miR-22* is corregulated with *miR-22* and siRNA silencing of *MYBL2* increases expression of *CDK2* and c-*MYC*. *MYBL2* is an mediates interactions between *miR-22* and cell-cycle control genes associated with HCC

susceptibility [272,273], and an attractive target for HCC therapy[274]. Further *FOXM1* is thought to regulate a transcriptional cluster that determines progression into G2 mitosis in hepatocytes[275], and deletion of *FOXM1* prevents development of HCC, and this activity is associated with a decreased level of *CDK1B* [276] expression.

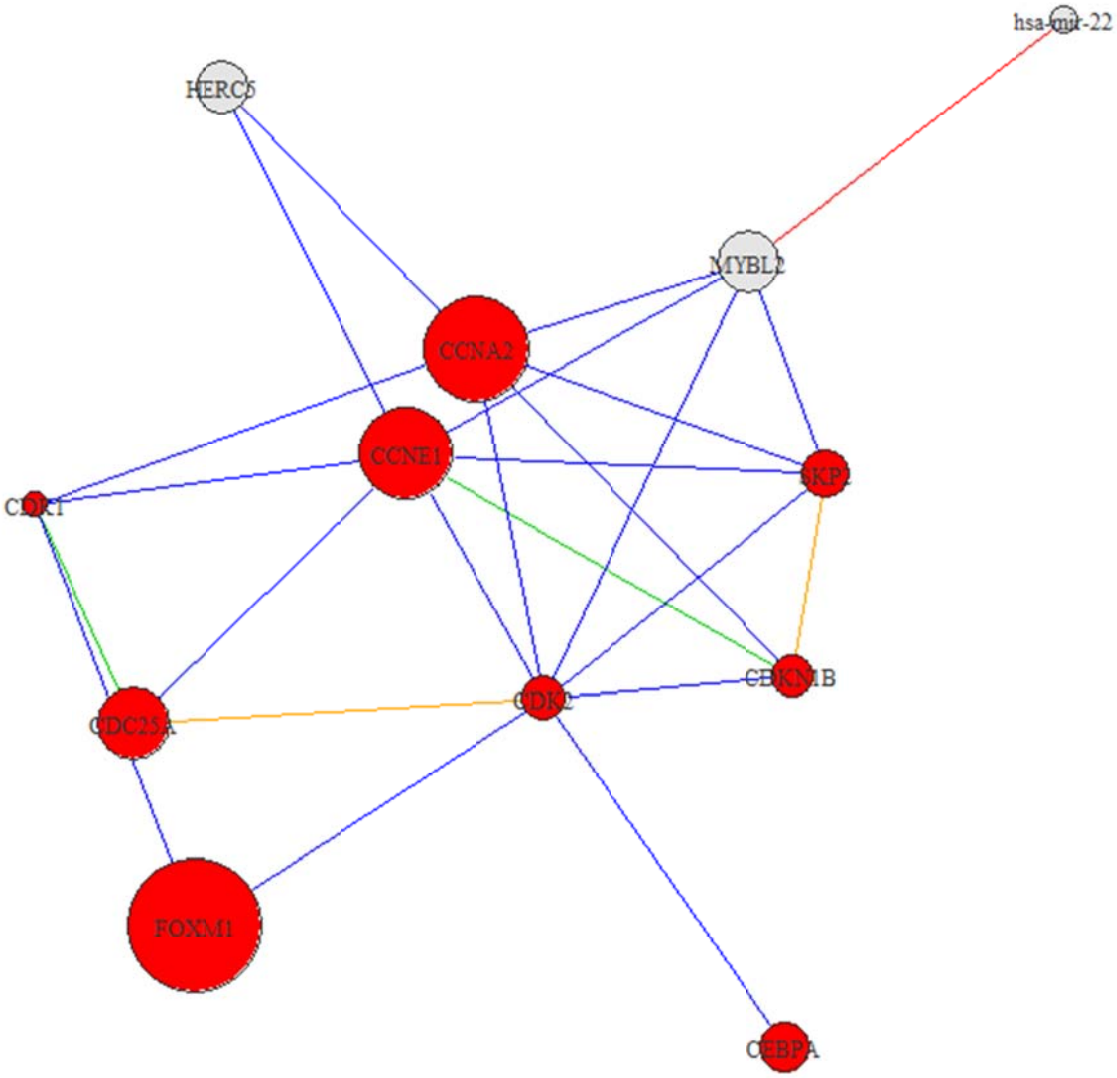


Figure 33: HCC Network Module 389. This module shows *FOXM1* and interactions with cell-cycle control genes. *MYBL2* is an interesting candidate for further research based on its connectivity to several cancer geens and as a target of *miR-22*. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, green from KEGG, orange from both databases and red designate miRNA-mRNA interactions.

MiRNA-183 co-regulates genes in modules 583, 186 and 200 (Figure 33). Module 583 presents interactions among the *CYP* family of detoxification and carcinogen-metabolizing genes and shows correlations between *miR-183* and *CYP1A2*, and *miR-196b* and *CYP4A11*. *CYP4A11* is a major fatty-acid omega- hydroxylase active in the liver [277,278] and *CYP4A22* polymorphisms have been found to be associated with HCC [279]. Module 186 similarly shows interaction among several miRNAs and detoxifying and carcinogen-metabolizing genes. *CYP3A4* is a target of miRNAs *miR-183* and *miR-96*, which have been found to co-regulate a cluster in prostate cancer[280]; and variants of *CYP3A4* have been associated with tumor aggressiveness[281]. In module 200, the target of *miR-183* and *miR-96*, and *miR-7* is *GNA14*[280]. GNA14 also plays a central role in BC module 292, although this gene is not well described in the literature. However, upregulated *GNA14* affects the activity of *PTGER3*[282] and *HTR2B*[283], which have proliferative effects in cells and are involved in liver regeneration[283]; and *CYSLTR1* which is involved in fatty acid metabolism and inflammation[284]. Such interaction among clusters reveals the importance of the *CYP* detoxification genes, particularly *CYP1A2*, which is highly differentially regulated and known to be associated with cancer, and candidate cancer genes *CYP4A22*, *CYP4A11 and GNA14*. The cluster also highlights *mir-183*, *mir-96*, and *miR-196b* as potential therapeutic targets in HCC.

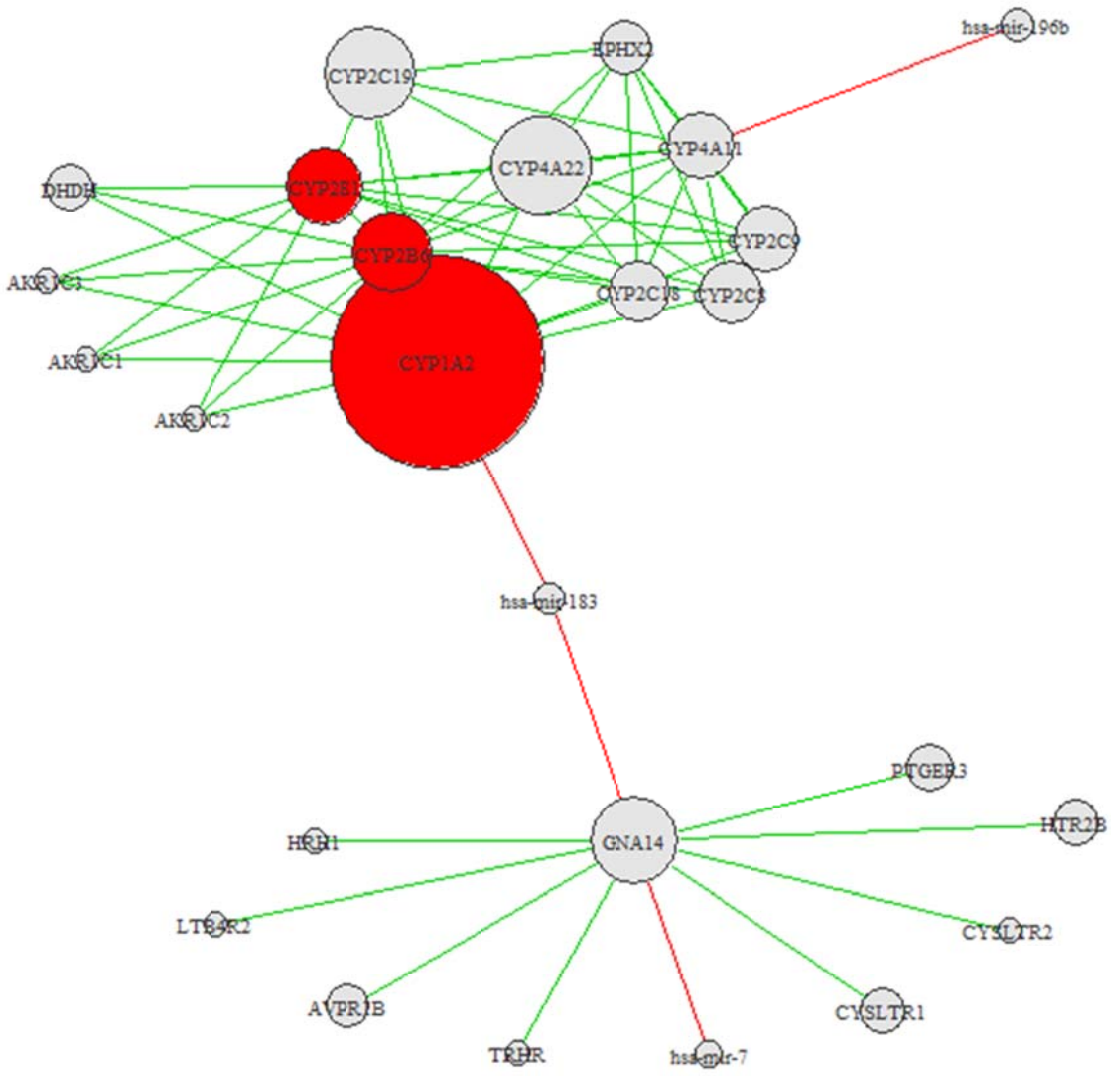


Figure 34: Intersection of HCC Network Modules 583, 200 and 186. Module 583 shows interaction among *CYP** metabolism and detoxification genes. Module 200 describes interaction between *GNA14*, which is not well described in the literature and fatty-acid metabolism and proliferation genes. Module 186 (figure in Appendix) describes interactions among *CYP**, *FMO**, *HSD** and *UGT** metabolism and detoxification genes. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, green from KEGG, orange from both databases and red designate miRNA-mRNA interactions.

Modules 318 and 92 are connected by association with *miR-151* and *miR-222* (Figure 35). In module 92, *LIFR* is highly differentially regulated and a target of *miR-128b*, *miR-151*, *miR-93* and *miR-339*. *LIFR* and *OSMR* are receptors for the *LIF* and *OSM* genes which function in cell differentiation, proliferation, survival and the inflammatory response[285,286] and it has been shown that *OSM*, when associated with *LIFR*, can induce a proliferative

response[287]. This module also includes *GHR*, a target of *miR-222*, which is involved in development and growth and interacts with interleukin signaling and *STAT* signaling genes. Together, the genes in these interacting modules are related to the immune response, transcription and growth signaling, and cell cycle control. Among these interactions, *LIFR* and *OSMR* are not well researched for their role in cancer, but these genes and their interaction are promising candiates for further cancer-based research.

Module 318 shows the interaction of *FBP1* with several *PFKFB* proteins, which play an important role in glycolysis and gluconeogenesis[288,289]. Interactions with this enzyme help to maintain a steady metabolic state, and to maintain an anaerobic source of metabolism for tumor cells. Depletion of *PFKFB3* has been shown to decrease tumor size, and it is believed to be responsible for maintaining glycolytic activity for cancer cells, reduced cancer cell proliferation[290], and the gene is a possible target for therapeutic intervention [289].

Module 348 (Appendix B) shows *NGFR* as a hub protein interacting with *FSCN1* and *ARHGD1B*, both involved in regulation of the actin cytoskeleton, influencing capacity of cells for metastasis. *NGFR* is a regulator of apoptosis and has been shown to play a key role in the differentiation and proliferation of hepatocytes in the diseased liver [291]. In this module, it is associated with three miRNAs, *miR-185*, *miR-186* and *miR-191*.

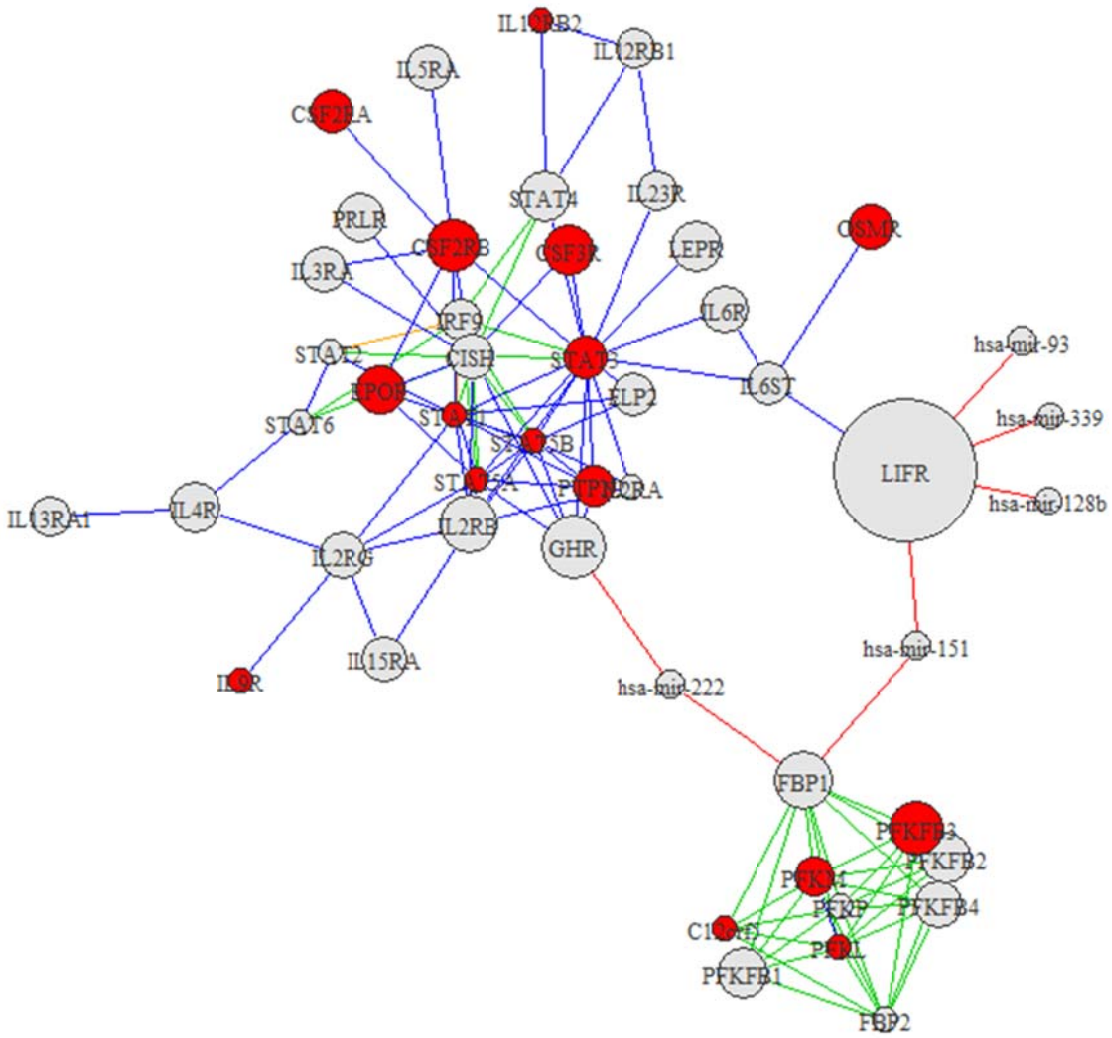


Figure 35: Intersection of HCC Network Modules 318 and 92. Module 318 shows interaction of *PFK** genes and *FBP1*, genes involved in glycolysis and energy metabolism, and FBP1 has been shown to afford a metabolic advantage to tumor cells. Module 92-*LIFR* and interactions with *OSMR*, *STAT* and interleukin signaling genes, involved in growth, the cell cycle and response to viral infections. *LIFR* is a receptor for *LIF* which is important in cell development and proliferation, and here this receptor is highly upregulated and the target of several miRNAs. Red nodes designate cancer-associated genes based on descriptions in OMIM. Node sizes correspond to the absolute values of the fold change of differentially regulated genes (up- or down-regulated). Blue edges are dervied from HPRD, green from KEGG, orange from both databases and red designate miRNA-mRNA interactions.

Table 10: Key Genes described in HCC miRNA Modules

HCC			
THBS1	thrombospondin 1	309	An adhesive glycoprotein that mediates cell-to-cell and cell-to-matrix interactions. Binds to fibrinogen, fibronectin, laminin, type V collagen and integrins alpha-V/beta-1 and plays roles in platelet aggregation, angiogenesis, and tumorigenesis.
FN1	fibronectin 1	309	Fibronectin is a glycoprotein involved in cell adhesion and migration processes including embryogenesis, wound healing, blood coagulation, host defense, and metastasis.
SPARC	secreted protein, acidic, cysteine-rich (osteonectin)	309	This protein regulates cell growth through interactions with the extracellular matrix and cytokines. Binds calcium and copper, several types of collagen, albumin, thrombospondin, PDGF and cell membranes. It is associated with tumor suppression and metastasis based its effects on cell shape which can promote tumor cell invasion.
PDGFRA	platelet-derived growth factor receptor, alpha polypeptide	567	Tyrosine-protein kinase that acts as a cell-surface receptor for PDGFA, PDGFB and PDGFC and plays an essential role in the regulation of embryonic development, cell proliferation, survival and chemotaxis.
MET	met proto-oncogene (hepatocyte growth factor receptor)	567	Receptor tyrosine kinase that transduces signals from the extracellular matrix into the cytoplasm by binding to hepatocyte growth factor/HGF ligand. Regulates processes including proliferation, morphogenesis and survival.
EGFR/ ErbB1-4	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, -3, -4,	567	The ErbB family includes: EGFR (ErbB1, HER1), ErbB2 (HER2), ErbB3 (HER3) and ErbB4 (HER4). Involved in a signaling cascade that drives many cellular responses, including changes in gene expression, cytoskeletal rearrangement, anti-apoptosis and increased cell proliferation. Amplification of these genes have been reported in numerous cancers.
FGFR2	fibroblast growth factor receptor 2	567	Tyrosine-protein kinase that acts as cell-surface receptor for fibroblast growth factors and plays an essential role in cell proliferation, differentiation, migration and apoptosis, and in the regulation of embryonic development. Mutations in FGFR genes may cause of several developmental disorders, and upregulation of FGFR may lead to cell transformation and cancer.
KDP	WNK lysine deficient protein kinase 1	567	Serine/threonine kinase that plays a key role in electrolyte homeostasis, cell signaling, survival, and proliferation
SH3GL2	SH3-domain GRB2-like 2	567	Implicated in synaptic vesicle endocytosis. May recruit other proteins to membranes with high curvature.
ZEB2	zinc finger E-box binding homeobox 2	232	A member of the Zfh1 family of 2-handed zinc finger/homeodomain proteins that functions as a DNA-binding transcriptional repressor and interacts with activated SMADs. Represses transcription of E-cadherin.
ZEB1	zinc finger E-box binding homeobox 1	232	Zinc finger transcription factor that inhibits interleukin-2 (IL-2) gene expression and regulates activity of ATP1A1. Represses E-cadherin and induces an epithelial-mesenchymal transition (EMT) by recruiting SMARCA4/BRG1. Represses BCL6 transcription with corepressor CTBP1. Promotes tumorigenicity by repressing stemness-inhibiting miRNAs
CTBP1	C-terminal binding protein 1	232	A protein that binds to the C-terminus of adenovirus E1A proteins. This phosphoprotein is a transcriptional repressor and is involved in cellular proliferation. It can form a complex including CTBP2 that regulates gene expression during development.
MECOM	MDS1 and EVI1 complex locus	232	A transcriptional regulator and oncoprotein that may be involved in hematopoiesis, apoptosis, development, differentiation and proliferation. Interacts with CTBP1, SMAD3, CREBBP, KAT2B, MAPK8, and MAPK9. May undergo translocation with the AML1 gene, resulting in onset of leukemia. May play a role in apoptosis through regulation of JNK and TGF-beta signaling.
SMAD4	SMAD family member 4	44	A member of the Smad family of signal transduction proteins which are phosphorylated and activated by transmembrane serine-threonine receptor kinases in response to TGF-beta signaling. These genes forms complexes with other activated Smad proteins, which then regulate the transcription of target genes. Mutations or deletions in this gene are associated with pancreatic cancer, juvenile polyposis syndrome, and hereditary hemorrhagic telangiectasia syndrome.
TGFBR2	transforming growth factor, beta receptor II	44	A member of the Ser/Thr protein kinase family and the TGFB receptor subfamily that acts to phosphorylate proteins, which regulate the transcription genes related to cell proliferation. Mutations in this gene have been associated with Marfan Syndrome, Loeys-Deitz Aortic Aneurysm Syndrome, and the development of various types of tumors.

TP53	tumor protein p53	44	The tumor protein p53 responds to diverse cellular stresses to regulate target genes that induce cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. p53 mutants that frequently occur in human cancers fail to bind the consensus DNA binding site, and hence cause the loss of tumor suppressor activity. Whilst the activation of p53 often leads to apoptosis, p53 inactivation facilitates tumor progression; inactivating p53 mutations occur in over 50% of cancers.
BRCA1	breast cancer 1, early onset	44	A nuclear phosphoprotein that plays a role in maintaining genomic stability, and it also acts as a tumor suppressor. The encoded protein combines with other tumor suppressors, DNA damage sensors, and signal transducers to form a large multi-subunit protein complex known as the BRCA1-associated genome surveillance complex (BASC). BRCA1 mutations are responsible for approximately 40% of inherited breast cancers and more than 80% of inherited breast and ovarian cancers.
FHL1	four and a half LIM domains 1	44	A member of the four-and-a-half-LIM-only protein family. Expression of these family members occurs in a cell- and tissue-specific mode and these proteins are involved in many cellular processes
CASP3, -7	caspase 7, apoptosis-related cysteine peptidase	44	A member of the cysteine-aspartic acid protease (caspase) family. Sequential activation of caspases plays a central role in the execution-phase of cell apoptosis.
CDC25B	cell division cycle 25 homolog B	44	Tyrosine protein phosphatase which functions as an inducer of mitotic progression. Required for G2/M phases of the cell cycle progression and abscission during cytokinesis in an ECT2-dependent manner. Directly dephosphorylates CDK1 and stimulates its kinase activity. CDC25B has oncogenic properties, although its oncogenic role is not well-understood.
CCND1	cyclin D1	44	This cyclin forms a complex with and functions as a regulatory subunit of CDK4 or CDK6, whose activity is required for cell cycle G1/S transition. This protein has been shown to interact with tumor suppressor protein Rb and its expression is regulated positively by Rb. Mutations, amplification and overexpression of this gene are observed frequently in tumors.
CSNK2A1	casein kinase 2, alpha 1 polypeptide	44	Subunit of a serine/threonine-protein kinase complex that regulates numerous cellular processes, such as cell cycle progression, apoptosis and transcription, and viral infection. CSNK2A1 is required for p53/TP53-mediated apoptosis. Phosphorylates CASP9 and CASP2, NOL3; RNA polymerases; and numerous transcription factors including NF-kappa-B, STAT1, CREB1, IRF1/2, ATF1, SRF, MAX, JUN, FOS, MYC and MYB. Phosphorylates Hsp90 and its co-chaperones FKBP4 and CDC37. Regulates Wnt signaling by phosphorylating CTNNB1 and LEF1. Phosphorylates proteins involved in viral life cycles..
UBB	ubiquitin B	44	Ubiquitin is a highly conserved protein required for intracellular protein degradation of proteins. Ubiquitin also binds to histone H2A but does not cause histone H2A degradation, suggesting involvement in regulation of gene expression.
MDM4	Mdm4 p53 binding protein homolog	44	Inhibits p53/TP53- and TP73/p73-mediated cell cycle arrest and apoptosis.
SMARCD1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 1	44	A member of the SWI/SNF family of proteins, whose members display helicase and ATPase activities and which are thought to regulate transcription of certain genes by altering their chromatin structure.
MS4A1	membrane-spanning 4-domains, subfamily A, member 1	44	This protein may be involved in the regulation of B-cell activation and proliferation
SLA	Src-like-adaptor	398	Adapter protein which negatively regulates positive selection and mitosis of T-cells. May link signaling proteins such as ZAP70 with CBL, leading to a CBL dependent degradation of signaling proteins
PTPRC	protein tyrosine phosphatase, receptor type, C	398	Protein tyrosine phosphatase that regulates a variety of cellular processes including cell growth, differentiation, mitosis, and oncogenic transformation. PTPRC has been shown to be key regulator of T- and B-cell antigen receptor signaling. Upon T-cell activation, recruits and dephosphorylates SKAP1 and FYN. Dephosphorylates and modulates LYN activity.
ZAP70	zeta-chain (TCR) associated protein kinase 70kDa	398	Protein tyrosine kinase that plays a role in T-cell development and lymphocyte activation. This enzyme functions in the initial step of TCR-mediated signal transduction in combination with the Src family kinases, Lck and Fyn.
LCP2	lymphocyte cytosolic protein 2	398	Involved in T-cell antigen receptor mediated signaling
SKAP2	src kinase associated phosphoprotein 2	398	Src family kinases that acts as an adaptor protein and is thought to play a key role in the src signaling pathway.

BLNK	B-cell linker	398	Cytoplasmic linker or adaptor protein that plays a critical role in B cell function and development and bridges SYK kinase to a multitude of signaling pathways. Deficiency in this protein has also been shown in pre-B acute lymphoblastic leukemia. Plays a role in the activation of ERK/EPHB2, MAP kinase p38 and JNK, AP1, NF-kappa-B and NFAT, PLCG1, PLCG2 and Ca(2+).
CCNE1	cyclin E1	389	This cyclin forms a complex with and functions as a regulatory subunit of CDK2, whose activity is required for cell cycle G1/S transition. Overexpression of this gene, which results in chromosome instability has been observed in many tumors.
CDKN1B	cyclin-dependent kinase inhibitor 1B (p27, Kip1)	389	Important regulator of cell cycle progression and G1 arrest. Potent inhibitor of cyclin E- and cyclin A-CDK2 complexes and is involved in the assembly, stability, and modulation of CCND1-CDK4 complex activation. Degradation of this protein, triggered by CDK-dependent phosphorylation and subsequent ubiquitination by SCF complexes, is required for the cellular transition from quiescence to the proliferative state.
FOXM1	forkhead box M1	389	Transcriptional factor regulating the expression of cell cycle genes essential for DNA replication and mitosis. Plays a role in the control of cell proliferation. Plays also a role in DNA breaks repair participating in the DNA damage checkpoint response
CYP4A11, -22	cytochrome P450, family 4, subfamily A, polypeptide 111	583	A member of the cytochrome P450 superfamily of enzymes which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids and hydroxylation of fatty acids. CYP4A11 oxidizes arachidonic acid to 20-HETE, while CYP4A22 shows no activity towards arachidonic acid and prostaglandin A1.
CYP2E1	cytochrome P450, family 2, subfamily E, polypeptide 1	583	A member of the cytochrome P450 superfamily of enzymes which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. Expression of this protein is induced by ethanol, the diabetic state, and starvation. The enzyme metabolizes both endogenous substrates, such as ethanol, acetone, and acetal, as well as exogenous substrates including benzene, carbon tetrachloride, ethylene glycol, and nitrosamines which are premutagens found in cigarette smoke. This enzyme is involved in processes as gluconeogenesis, hepatic cirrhosis, diabetes, and cancer. Bioactivates many xenobiotic substrates to their hepatotoxic or carcinogenic forms.
CYP1A2	cytochrome P450, family 1, subfamily A, polypeptide 2	583	A member of the cytochrome P450 superfamily of enzymes which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. Expression of this protein is induced by some polycyclic aromatic hydrocarbons (PAHs), some of which are found in cigarette smoke. The enzyme is able to metabolize some PAHs to carcinogenic intermediates. Other xenobiotic substrates for this enzyme include caffeine, aflatoxin B1, and acetaminophen
CYP3A4	cytochrome P450, family 3, subfamily A, polypeptide 4	183	A member of the cytochrome P450 superfamily of enzymes that catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. Expression of this protein is induced by glucocorticoids and some pharmacological agents. This enzyme is involved in the metabolism of approximately half the drugs in use today, including acetaminophen, codeine, cyclosporin A, diazepam and erythromycin. The enzyme also metabolizes carcinogens.
CYP2A6	cytochrome P450, family 2, subfamily A, polypeptide 6	186	A member of the cytochrome P450 superfamily of enzymes that catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. This protein localizes to the endoplasmic reticulum and its expression is induced by phenobarbital. The enzyme is known to hydroxylate coumarin, and also metabolizes nicotine, aflatoxin B1, nitrosamines, and some pharmaceuticals inlcuding the anti-cancer drugs cyclophosphamide and ifosphamide.
CYP2A7	cytochrome P450, family 2, subfamily A, polypeptide 7	186	A member of the cytochrome P450 superfamily of enzymes that catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. It oxidizes compounds including steroids, fatty acids, and xenobiotics.
CYP1A1	cytochrome P450, family 1, subfamily A, polypeptide 1	186	A member of the cytochrome P450 superfamily of enzymes that catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. Expression of this protein is induced by some polycyclic aromatic hydrocarbons (PAHs), some of which are found in cigarette smoke and it is able to metabolize some PAHs to carcinogenic intermediates. It oxidizes a variety of structurally unrelated compounds, including steroids, fatty acids, and xenobiotics.
GNA14	guanine nucleotide binding protein (G protein), alpha 14	186	Guanine nucleotide-binding proteins (G proteins) are modulators or transducers in various transmembrane signaling systems

SULT1E1	sulfotransferase family 1E, estrogen-preferring, member 1	186	Sulfotransferase enzymes catalyze the sulfate conjugation of many hormones, neurotransmitters, drugs, and xenobiotic compounds. This protein transfers a sulfo moiety to and from estrone, which may control levels of estrogen receptors.
UGT2B7, -10, -11	UDP glucuronosyltransferase 2 family, polypeptide B10, -11	186	UDPGT is of major importance in the conjugation and subsequent elimination of potentially toxic xenobiotics and endogenous compounds.
FBP1	fructose-1,6-bisphosphatase 1	318	Fructose-1,6-bisphosphatase 1, a gluconeogenesis regulatory enzyme, catalyzes the hydrolysis of fructose 1,6-bisphosphate to fructose 6-phosphate and inorganic phosphate
PFKFB1,-2,-3,-4	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 1,-2,-3,-4	318	Family of bifunctional 6-phosphofructo-2-kinase:fructose-2,6-biphosphatase enzymes involved in the synthesis and degradation of fructose 2,6-bisphosphate
PFKM, -L	phosphofructokinase, liver, muscle	318	Phosphofructokinase isozymes catalyzes the phosphorylation of D-fructose-6-phosphate to fructose-1,6-bisphosphate.
LIFR	leukemia inhibitory factor receptor alpha	92	Cytokine receptor that interacts with gp130 to form a complex that mediates the action of the leukemia inhibitory factor, a cytokine involved in cellular differentiation, proliferation and survival. May have a common pathway with IL6ST.
OSM	oncostatin M	92	A member of a cytokine family that includes leukemia-inhibitory factor, granulocyte colony-stimulating factor, and interleukin 6. This gene encodes a growth regulator which inhibits the proliferation of a number of tumor cell lines.
GHR	growth hormone receptor	92	A member of the type I cytokine receptor family, which is a transmembrane receptor for growth hormone.
STAT3, -4	signal transducer and activator of transcription 3, -4	92	Members of the STAT family of transcription factors. In response to cytokines and growth factors, STAT family members are phosphorylated by receptor associated kinases, and act as transcription activators that mediate the expression of a various genes in response to cell stimuli, and play a key role in cellular processes such as cell growth, differentiation and apoptosis.
PTPN2	protein tyrosine phosphatase, non-receptor type 2	92	A protein tyrosine phosphatase that regulates a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation. Substrates include EGFR and Shc, suggesting a role in growth factor cell signaling.
CSF3R,-2RA, 2RB	colony stimulating factor 3, -2A, -2B receptor (granulocyte)	92	Receptors for granulocyte colony-stimulating factor (CSF3), essential for granulocytic maturation. They play a key role in the proliferation, differientation and survival of neutrophilic cells. They may also function in cell adhesion or recognition events.
EPOR	erythropoietin receptor	92	Erythropoietin cytokine receptor. Mediates erythropoietin-induced erythroblast proliferation and differentiation and erythroid cell survival. Upon binding, this receptor activates Jak2 which activates pathways including: Ras/MAP kinase, phosphatidylinositol 3-kinase and STAT transcriptions. Dysregulation of this gene may affect the growth of certain tumors.
IL2RB	interleukin 2 receptor, beta	92	Receptor for interleukin-2 involved in receptor mediated endocytosis and transduction of mitogenic signals.
IRF9	interferon regulatory factor 9	92	Transcription regulatory factor that mediates signaling by type I IFNs (IFN-alpha and IFN-beta). Following type I IFN binding to cell surface receptors, Jak kinases (TYK2 and JAK1) are activated, leading to tyrosine phosphorylation of STAT1 and STAT2.
IL15RA	interleukin 15 receptor, alpha	92	High-affinity receptor for interleukin-15. Signal transduction involves STAT3, STAT5, STAT6, JAK2 (By similarity) and SYK. This receptor is reported to enhance cell proliferation and expression of apoptosis inhibitor BCL2L1/BCL2-XL and BCL2.
IL6ST	interleukin 6 signal transducer	92	A signal transducer shared by many cytokines, including interleukin 6 (IL6), ciliary neurotrophic factor (CNTF), leukemia inhibitory factor (LIF), and oncostatin M (OSM). Knockout studies suggest it plays a key role in regulating myocyte apoptosis.
FSCN1	fascin homolog 1, actin-bundling protein	348	A member of the fascin family of actin-binding proteins that plays a critical role in cell migration, motility, adhesion and cellular interactions. Expression of this gene is known to be regulated by several miRNAs, and overexpression of this gene may play a role in the metastasis of multiple types of cancer by increasing cell motility.
ARHGDIB	Rho GDP dissociation inhibitor (GDI) beta	348	Members of the Rho (or ARH) protein family and other Ras-related small GTP-binding proteins are involved in diverse cellular events, including cell signaling, proliferation, cytoskeletal organization, and secretion
NGFR	nerve growth factor receptor	348	This receptor can bind to NGF, BDNF, NT-3, and NT-4 and can mediate cell survival and cell death of neural cells. It is an oncogene commonly mutated in cancers.

Table 11. Significant MicroRNAs and their Targets

mircoRNA	target	module id
Breast Cancer		
miR-33b	IGF1	379
miR-770	YES1	74
miR-577	ETV6	22
miR-934	TP53BP2	63
miR-223	PTAFR, CYBB, PIK3CG	292, 327, 269
miR-146a	PIK3CD	269
Hepatocellular Carcinoma		
miR-34b	THBS1	309
miR-184	PDGFRA	567
miR-219, miR-107	ZEB2	232
miR-382	PTGIS	343
miR-218	PTPRC, FHL2	398, 218
miR-22	MYBL2	389
miR-96	CYP2D6	186
miR-196b	CYP4A11	583
miR-204	CLDN2	650
miR-301	UGT2B7	186
miR-222	FBP1	318
miR-151	FBP1	343
miR-128, miR-339, miR-93, miR-152	LIFR	92
miR-222	GHR	92
miR-194, miR-448, miR-142	PDE1A	582
miR-186, miR-185, miR-191	NGFR	348
miR-183	CYP1A2, GNA14, CYP3A4	583, 200, 186
miR-7	GNA14	200

5.3.5: Overlap with other Studies

To consider the extent to which my approach independently reproduced previous findings, I search for overlap of my and those of previous studies. I compare these results to the original studies that produced the HCC and BC coexpression data used in this analysis. Burchard et al.[11] study a training set of 96 matched samples and test sets of 180 and 40 samples. Other miRNAs that are highly differentially regulated include *miR-139, miR-99a, miR-10a, miR-199a/miR-199a*, miR-450, miR-378, miR-125b, miR-214, miR-422b, miR-424, miR-451,* and *miR-101*. They find *miR-122* expression positively correlated with mitochondrially localized proteins and metabolic functions including fatty acid metabolism,

and valine, leucine and isoleucine degradation. *miR-122* has been found to be under-expressed in HCC and the miRNA is associated with metabolic function in tumors and HCC metastasis [292-294]. They validate putative direct targets of *miRNA-122: SMARCD1*, *MAP3K3*, and *CAT-1,* which were reduced with increased expression of *miRNA-122;* while putative secondary target genes, *PPARGC1A,* and *SDH* subunits A and B, were increased with decreased expression of *miRNA-122*. The most connected secondary target was *PPARGC1A,* with 27 functional similarities, including *MED1*, *SMARCD1, LCMT1*, *PPP1CC*, *ATF4*, *MAP3K3*, and *MAPKAP2*. Dysregulation of normal mitochondrial functions may contribute to cancer metabolism and hepatocarcinogenesis, as the relationship between mitochondrial dysfunction and cancer is well documented[295]. *SMARCD1* stimiulates fatty-acid oxidation with *PPARGC1A*, also proposed to be a primary target of *miR-122*. Other proposed links with *miR-122* include *CAT-1,* Cyclin G1 and *N-myc* which were not reproducible, and *BCl-w* which was found in the BC modules. Only the target *SMARCD1* was among the enriched clusters in HCC and was reproduced using the *Walktrap* approach. However, related metabolic functions isolated by Burchard were highly enriched in HCC modules 583 and 343.

Buffa et al.[12] analyze coordinated expression of prognostic miRNAs and predicted target genes in 207 early-invasive cancers. They integrate mRNA and miRNA data to elucidate miRNA function in vivo and to identify interactions between miRNAs and targeted mRNAs for enhanced marker and therapeutic discovery. They evaluated predicted targets and their statistical significance to identify the following prognostic miRNAs: *miR-767-3p*, *miR-128a*, *miR-769-3*, and *miR-135a* in ER+ samples; *miR-27b*, *miR-144*, *miR-210*, *miR-342*, *miR-150* and *miR-30c* in ER- samples; and *miR-29c*, *miR-642* and *miR-548d* in all samples. The targets are implicated in pathways that play important roles in tumor growth and metastasis; altered pathways represent activity in apoptosis, *FGF* receptor signaling, *PTEN* and *FOX01,* tumor repressors, glutamate receptors, the *Wnt* pathway, immunity, proliferation, glycolysis, DNA repair, mitochondrial respiration, notch signaling, map kinase and *JNK* signaling. Most of these pathways were also enriched in the *Walktrap* modules. *miR-210* targets the cited study include *ISCU*, *CBX7* and *IGF1R*. *IGF1* was a hub and a direct target of *miR-33b* in the *Walktrap* modules. I found *miR-128* and *miR-7* to be enriched in the HCC data, but not in the BC data; and *miR-150* is among the miRNAs enriched in the BC clusters and is associated with immune functions.

Liu et al. [296] investigate miRNAs as alternate biomarkers to detect early-stage HCC. They find *miR-15b* and *miR-130b* to perform very well as predictive serum biomarkers, better than the state of the art method using AFP as a biomarker. MiRNAs can serve as valuable

biomarkers, as different cancer types have distinct miRNA expression profiles, and miRNA expression levels may be reliably detected in plasma or serum with high stability. However, *miR-15b* and *miR-130* did not appear in the *Walktrap* list of miRNAs associated with HCC, *miR-15b*, *-21*, *-130b*, *-183*, *-224* and *-301* were found to be consistently highly expressed in all HCC samples. *miR-301* and *miR-183* were also reproduced in the *Walktrap* analysis. Among genes found to be regulated by these miRNAs are *E2F*, *RUNX3* and *Bim*. *RUNX1* (HCC module 301) and *RUNX1T1* (HCC module 22), are members of significant *Walktrap* modules, and these transcription factors bind to the same type of Runt DNA-binding domain as *RUNX3*.

To summarize, I identified candidate genes including miRNAs, their targets and related genes that overlap with genes in the cancer studies that produced the data. The reproducible findings help validate these results. However, the overlap was not extensive, and many genes highlighted in *Walktrap* modules were not discussed in the Burchard and Buffa investigations. This is, in part, expected, and in some instances, may be explained by the fact that I applied matching criteria based on biological filters and significance values that filtered these matches. Further, cancers studies, and in particular, miRNA studies, yield varied independent findings as revealed by review of the literature. This may be a function of the large number of combinations and permutations of miRNA-mRNA pairings and the complexity of such gene regulation.

5.3.6: Overlap with mRNA-only Analysis

Signficant miRNA-cancer modules were also compared with the modules identified in Chapter 4 using independent BC and HCC datasets (datasets BC: GSE7390, HCC: GSE14520, CC: GSE8671), to validate the miRNA findings against experiments using only miRNA data (HCC: GSE22850 and BC: GSE22058). Considerable overlap was evident among miRNAs that were prominently enriched in modules from the BC and HCC datasets and significant clusters of Chapter 4 that did not include miRNA data. The reproducibility of these results in independent data sets adds further validity to these findings. BC data in module 379 overlaps with GSE7390.429 genes: *IGF1, NOV*, *IGFALS*, *IGFBP*'s and *IDE* interactions. The *LIFR* and *OSMR* (GSE7390.89, GSE8671.410) genes and interleukin interactions overlap among GSE14520 and miRNA-associated HCC clusters. GSE22058.567 overlaps with GSE7390.82, including *PDGFRA*, *ErbB* signaling genes, *KDR*, *GRB10*,

FLT1/4 and *RET* genes. GSE7390.143 and GSE22058.389 overlap among *CCNA2*, *CDK2*, *CDKN1A* and *SKP2* genes, which are among the most differentially regulated genes in both clusters. *FN1* and collagen interactions and their interactions overlap in GSE22058.309 and GSE14520.111. *SPARC1*, *THBS1* and collagen genes also overlap with GSE14520.328. *CYP**, *FMO** and *UGT** genes in modules 583 and 186 in GSE22058 also overlap with modules GSE14520.10 and GSE14520.408. GSE22220.379 overlaps with genes *IGF1*, *NOV*, *IGFBP7* and *IDE* in GSE14520.429.

Considering this inclusive sample of significant miRNA-coregulated clusters, many significant modules found in this chapter overlap with those of independent datasets with similar outcomes in Chapter 4. Such overlap shows that there is a reproducible signal in these significant clusters, which is validated by the coordinated differential activity of miRNA and mRNA expression data.

5.3.7: Sensitivity Analysis

To examine sensitivity of the community search methods to changes in node weights, the top 5% of significant miRNAs and their mRNA targets were retained using their original fold change weight values. All other nodes in the network were given a nominal non-significant fold change value of .01. The random walk community search was applied to the modified network to determine the extent to which 1) significant miRNAs and their targets were identified in the altered network, and 2) the module composition of significant miRNA matches changed.

Significant miRNAs were still detected in significant modules. However, module composition did change, given that the edge weights were modified. New modules consist of a different composition of neighboring interactions within the global network. If genes **A**, B, C, D, E and miRNA **X** were in a module previously (where **A** and **X** were significant genes of interest), the new module may consist of interaction among **A**, B, F, E, G and miRNA **X**. Therefore, as expected, community structures appear to be sensitive to weights in the global network. This behavior may be explained by the tendency of communities to be determined by average fold change values with adjacent nodes, and the influence of both weights and the degree connectivity on the random walk process.

5.4: Conclusion

Given their prognostic importance and functional relevance, miRNA analysis is proving to be a key component in the development of future cancer diagnostics and therapies. Developing efficient and effective frameworks to model miRNAs and their correlated and predicted targets in the context biological processes is an important area of current research. Network-based approaches provide a framework for the integration and analysis of diverse genomic data, including mRNA and miRNA regulatory data with high-throughput interaction information and cancer outcomes.

This study shows graph based models and the graph-based analytic methods are useful tools to integrate and model diverse types of genomic information. I implement a random walk algorithm in a weighted network including mRNA, PPI, metabolic, signaling and miRNA interactions. Several matching schemes are evaluated to integrate miRNAs in the network and multiple weighting schemes are investigated to score miRNA-enriched modules. The optimal network integration approach incorporates the top five matching miRNAs per gene target and the best scoring strategy uses fold change based edge weights. Results demonstrate that modules associated with cancer and enriched with miRNA targets can identify important candidate genes and therapeutic targets. Significant modules highlight differentially regulated genes of interest in based on their potential prognostic and therapeutic value in cancer, such as *miR-22 miR-196b miR-151 miR-93,GNA14, CYP4A11, SKAP1, SH3GL2, MYBL2* and *LIFR* in hepatocellular carcinoma and *miR-33b*, *mir-223, mir-770 YES1, ETV6, PTAFR*, and *CYPB* in breast cancer

Network analysis integrating miRNA data enhances the search for relevant disease modules by adding an additional layer of evidence to correlated and differentially expressed targets associated with cancer outcomes. The search for the association of miRNA/mRNA co-regulation and disease outcomes can be improved with better and more certain data. Further, more meta-analysis studies would be useful to summarize the effects of miRNAs on specific cancer types and stages, as well as to help track the reproducibility of such findings. Such studies can provide more resolution for the widespread variability across miRNA studies an their potential in personalized medicine, an important contribution to research where miRNAs are evaluated as therapeutic and prognostic targets.

Chapter 6: Conclusion

Outstanding issues in the analysis of high-dimensional data in genomics include lack of sufficient statistical power, multiple testing, and overfitting data. Developing better platforms to model interaction data and integrate prior evidence to address these issues is an important area of current research. In this dissertation, I have implemented and evaluated a graph-based approach to analyze genomic data in an effort to improve upon current methods. Graph-based approaches are a powerful framework for genomic studies because they are tailored to model complex relationships, and support qualitative and quantitative analysis of interaction data. I focus upon network centrality concepts and modularity in the analysis of cancer expression data to prioritize important interactions and functional groups of genes associated with cancer onset and progression.

The principal hypothesis of this work is that the use of a graph-based approach to study large-scale genomic data, focusing on network characteristics and module generation in biological networks, provides a powerful framework for data integration and improves performance and interpretation in the analysis of the coordinated behavior of genes in complex disease. I address this research question first by investigating the value of network features in a biological network, namely centrality, cohesion and modularity, to predict cancer genes. I then test the graph-based platform applying a random walk-algorithm to an interaction network weighted by expression data to search for genes associated with cancer outcomes. Finally, I investigate integrating multi-scale data and modeling regulatory relationships by including microRNA (miRNA) data in the interactome to study the influence of miRNA-mRNA regulatory activity on cancer onset and progression. These methods are evaluated by measuring their ability to extract known cancer genes, and examining functional annotation and the literature to determine the relevance of significant interactions.

Chapter 3 presents an analysis of network characteristics of cancer genes. A custom parser is developed to extract metabolic and signaling interactions from the KEGG pathway database. To provide a gold standard for cancer gene status, the OMIM database is mined for evidence of cancer-association of all genes in the metabolic and signaling networks. Centrality features and clustering coefficients are calculated for nodes in the network and a linear classifier is used to determine if these features are predictive of cancer gene status. Logistic regression estimates quantify the predictive ability of centrality and clustering coefficient and show more predictive power in signaling networks compared to metabolic

networks. In the assessment of the value of centrality features to predict cancer genes, centrality characteristics, in particular degree and closeness centrality, proved predictive of the status of cancer genes. Metabolic and signaling networks exhibit significant topological differences in terms of degree, clustering coefficients and community cohesiveness of cancer genes; and centrality features demonstrate greater predictive value in signaling networks. Further, cancer genes were found to be more cohesive than non-cancer genes, and significantly clustered in modules. Cancer genes in signaling communities tend to be more cohesive than those in metabolic communities and represent cell cycle, adhesion *Wnt*-signaling and *TGFβ* signaling pathways among other cancer-related processes. When investigating the metabolic network, communities of cancer genes frequently show methylation activity, amino acid synthesis and metabolism, and interact with signaling pathways. Network relationships can provide predictive value in identifying novel cancer genes, and definition of communities of cancer genes can help elucidate complex interactions influencing the onset and progression of cancer. These results provide an empirical basis for the application of algorithms using similar network-based measures to prioritize disease genes or predict disease states.

In Chapter 4, pathway interactions, protein interactions and expression data are merged in a biological network to search for cancer-associatied modules. The interactome is constructed from KEGG and HPRD data and the network is augmented with weights from three cancer expression studies. I implement *Walktrap*, a random walk-based community detection algorithm to identify modules predisposing to disease onset in hepatocellular carcinoma (HCC), adenoma development in colorectal cancer (CCA), and prognosis in breast cancer (BC). For each data set, the best scoring partitions under a maximum cluster size (max=200) were selected. Significant modules are rich in functional annotation associated with known cancer processes. These modules include interactions among transcription factors (*SPIB*, *RPS6KA2* and *RPS6KA6*) and cell-cycle regulatory genes (*BRSK1*, *WEE1* and *CDC25C*) that interact closely with other known cancer genes, are functionally related to cancer, and show promise as therapeutic targets. This approach is evaluated by comparing the cancer gene enrichment of modules discovered by *Walktrap* compared to those results from two other highly cited module-finding platforms, *Matisse* and *jActiveModules*. Overall, *Walktrap* performs as well or better than these tools across all datasets. Further, a size restriction is imposed in the module-finding algorithm, and the resulting modules are generally smaller and more interpretable compared to *Matisse* and *jActiveModules*. These

results demonstrate that the *Walktrap* algorithm performs well against related tools and can identify modules significantly enriched with cancer genes, their joint effects and promising candidate genes. Findings from this work can be used to develop hypothesis for further cancer-based studies.

In Chapter 5, miRNA data is mereged into the interactome built in Chapter 4 to investigate the ability of graph-based methods to integrate diverse data types and to use this information to search for high-confidence candidate genes. Two cancer data studies, one breast cancer survival one hepatocellular carcinoma study, which include miRNA and mRNA coexpression data, are used to integrate regulatory information in the network. Variations of matching methods were evaluated including using: the single optimal miRNA-mRNA match, retaining the best five or three pairs, or all miRNA-mRNA pairs. Multiple methods to integrate miRNA-mRNA matchings as edges in the network were also examined, including using fold change, a transformation of fold change to boost the importance of miRNA matches, and excluding miRNAs in the network but including their targets in enrichment analysis. Using Precision, Recall and Matthew's Correlation to measure performance, the best five filtered matches produced the best matching strategy, and using fold change without transformation produced the best network integration strategy. The resulting modules include differentially regulated candidate genes based on their potential prognostic and therapeutic value in cancer, such as *miR-22 miR-196b miR-151 miR-93,GNA14, CYP4A11, SKAP1, SH3GL2, MYBL2* and *LIFR* in hepatocellular carcinoma and *miR-33b*, *mir-223, mir-770 YES1, ETV6, PTAFR*, and *CYPB* in breast cancer. Further, overlap was evident in the functional annotation and specific gene groupings when comparing miRNAs and targets found by the *Walktrap* method with those of the original studies. These findings overlapped in part with earlier cancer-based miRNA studies; however, *Walktrap* identified primarily novel interactions not supported by previous work. These results demonstrate that modules associated with cancer and enriched with miRNA targets can identify important genes involved cancer pathways, and novel miRNAs associated with cancer.

6.1: Limitations

Several limitations were encountered in the course of this research. First, a there is a limitation considering the generation of the null hypothesis for significance testing of modules. The methods presented in this dissertation did not consider network topology when determining module significance. Specifically, the connectivity of genes within a module were not considered when evaluating the significance of a module; only the cumulative weights of the nodes were used as a basis for module activity scores, and the random distributions, independent of network structure. Determining the direction of the bias in this approach, or if inclusion of network topology would add additional bias would require further systematic research. I did not find a computationally efficient and scalable method to generate a random distribution for a null-hypothesis including network structure for each test case. To address this limitation, modules were ranked by significance and incorporated up to the top-ranked 25 modules. This limitation does not bias comparisons with other tools, as the same scoring metric was used for *Walktrap* modules and those discovered by *Matisse* and *jActiveModules*. It should also be noted that there may be a study bias in the search for relevant modules in that well-studied genes may be appear more significant in the network due to the fact that their their relationships with other genes are better studied.

A significant limitation of this work is lack of wet laboratory support, as the hypotheses generated by the module searches in Chapters 4 and 5 could not be verified experimentally. Such support would have enabled experiments to validate the biological relevance of novel genes and interactions highlighted in signficant modules. Without such support, I rely on evidence from the literature and previous work to validate my results. Hence, these methods generate hypotheses for several good candidate genes and interactions but further experiments to examine their effects and interactions in normal and cancer tissues are needed. Establishing experimental evidence to support the significance for these findings is an important area for future research.

The search for the association of protein-protein interactions, signaling and metabolic interactions, and miRNA/mRNA co-regulation with disease outcomes can be improved with better and more certain data. The interaction networks in my experiments were assembled using KEGG and HPRD data, but I also found that Reactome and INTACT would be suitable to generate interactomes for this study. There are many differences between these databases, Reactome, for one, consists of a smaller set of direct PPI data than HPRD but includes an extended network of indirect interactions that may be of interest. Appending additional PPI

data could be and area for future work. Further, the miRNA-mRNA filter applied in target scan is neither certain nor complete. The filter improves the search for likely regulatory pairs, but the pairs may be false positives, or may not interact in the given data, and important pairs may be omitted. Continued investigation and validation of miRNA-mRNA pairs and improved tools for functional prediction would improve network models of these relationships.

Findings in Chapter 5 did not show extensive overlap with previous findings, including the published findings from which the data were obtained. This is partially explained due by the fact that I applied stringent biological filters which preliminarily eliminated some possible matches that were found in the original analyses. I took this approach to focus my search on the most relevant pairings in the network, but in doing so, some information may have been lost. Further, based on my review of the literature, many findings across miRNA studies are not reproducible, raising the question of whether miRNA analyses yield results that are more or less consistent than expression or SNP studies. Considering that each miRNA may have hundreds of possible targets, the contextual variability of miRNA expression and function seems to be substantial.

The generalizability of extending the networks to more diverse data types, beyond miRNA data, has not been tested. I have not applied the random walk algorithm to data types apart from expression data; for example, transcription factor, DNA methylation, mutation, and copy number variation. Further, when integrating several expression data sets, as when merging mRNA and miRNA expression, the use of different normalization strategies affects the scoring of such modules and is a critical factor in the experimental design. A systematic study of normalization strategies for merging of multiple expression data sets would improve this work.

6.2: Contributions

Contributions of this study to the fields of biomedical informatics, genomics and cancer biology include: the implementation and evaluation of methods for high-dimensional data analysis; applications of network algorithms in biology, and; approaches to data integration in biology. The graph-based random walk is used to integrate prior biological evidence, including biological interactions, experimental and miRNA-mRNA regulatory information; and to detect significant network modules in large biological datasets. Such an

approach narrows the feature space and enables the search for the combined effect of genes associated with cancer in data where the effects of these genes might have independently been considered non-significant. Leveraging existing knowledge of the relationship between genes and their biological pathways facilitates interpretation of significant findings and the generation of high-quality hypotheses for further study.

Building on previous network-based approaches used to discover modules in integrated interaction networks, this study focuses on the importance of biologically interpretable and focused modules. Network analysis strategies vary based on their weight, scoring, and module-finding approaches. Where many previous studies have used correlation values, this study focuses on fold change values to focus the analysis on outcomes rather than the strength of correlation between genes. Prior work using seed-based algorithms and node weights may be sensitive to low values of single nodes, even if the adjacent nodes have high values. Applying edge weights by using the average of adjacent nodes allows for breaks of links in the network, and prevents loss of data due to low-significance intermediate genes. Thus, where an intermediate gene may not have a high differential fold change, if neighboring genes show high-signficance, the continuity of the chain of genes, including the intermediate gene, is preserved. Module activity is based on the cumulative weights of the fold-chance values in the module. I develop scoring metrics using a bootstrapping method to determine module significance and evaluate *Walktrap* modules.

Random walk algorithms vary based on their distance metrics, their heuristics, and optimization strategies. Where most previous work using random walks has focused on gene prioritization, this research contributes to the scope of work covering module-finding with an optimized algorithm for discovering communities in large networks. Several stopping thresholds were implemented to optimizing community-finding, including size, score and maximizing modularity. I apply a workflow considering size, score and optimal modularity to determine a stopping point for the merge process to improve the search for significant and interpretable modules. Evaluation results show that this approach shows strong performance when compared with similar tools and yields smaller, more interpretable modules.

To estimate the enrichment of cancer genes in modules, I created a gold standard list for the annotation of cancer genes. I considered several other lists, including the Sanger Cancer Gene Census[297], the Cancer Gene Atlas[298] and the Waldman Gene List by Locus [299], and a compilation of cancer gene lists compiled by Higgins et al.[300]. However, these lists were either too restrictive: for example the Cancer Gene Census and Waldman Gene List include 487 and 510 genes, respectively; or too inclusive in their coverage of cancer-

associated genes, where the Cancer Gene Atlas list of "possible" cancer genes consists of 8395 genes. The Cancer Gene Atlas list of 1174 genes had the best coverage; however there were cancer-associated genes in my network not in this list, and my best approximation to a gold standard to cover all genes in this study was achieved via text-mining of OMIM for evidence of cancer association. To assemble this list, I queried each gene in the network for cancer-associated terms and manually verified each match. This approach improves upon previous approaches to summarize cancer gene data, based on the specificity and coverage of queries and manual verification. I reviewd 6639 genes and assembled a gold standard reference composed of 1239 cancer genes. A recent study has gathered current data to construct a comprehensive consensus-driven list of annotated cancer genes[301], but this reference was not available during the course of my evaluation.

This research demonstrates the ability of the network to integrate pathway, protein-protein interactions, expression measurements and miRNA-mRNA regulatory data. Several strategies were considered to integrate miRNA data and the optimal strategy uses a subset of the best five pairings and searching for enrichment of miRNA. Inclusion of miRNA evidence increases confidence in candidate genes and their interactions based on the mutual importance of the miRNA and its target in cancer data.

Findings from this study identify candidate genes that are implicated in breast cancer, hepatocellular carcinoma and colorectal cancer. In Chapter 4, several genes were identified as targets for further research, including *CBLC* and *IRS2* which are associated with breast cancer survival; transcription factors *RPS6KA2*, *RPS6KA6* and the interaction among *MCM/CDC* and *ORC* cell cycle genes, associated with the onset of hepatocellular carcinoma; and cell-cycle genes *BRSK1*, *WEE1*, *CDC25C* and the transcription factor *SPIB*, associated with colorectal adenoma development. In Chapter 5, *GNA14*, *CYP4A11*, *SKAP1*, *SH3GL2*, *MYBL2* and *LIFR* were identified as candidate genes related to miRNA hepatocellular carcinoma and *IFGAL*, *SIN3A*, *PTAFR* and *CYPB*, and *HDAC* genes were associated with breast cancer. Candidate miRNAs include *miR-33b* and *miR-223* in breast cancer and *miR-184, miR-93, and miR-183* in hepatocellular carcinoma.

Finally, the random walk approach also provides a generalizable example to integrate diverse information and find communities of closely related entities to guide other applications of graph-based research.

6.3: Future Directions

Future work may include studies focused on evaluation and interpretation of modules discovered in biological networks. This dissertation makes a primary contribution to this area by focusing on the size of modules and emphasizing the interpretability of modules; however, further studies on the systematic interpretation of modules beyond functional annotation would be very useful to track the progress and success of methods in this area of research. Specifically, measuring the accuracy of such modules based on a suitable and common gold standard and further discussion of what constitutes the composition and size of an "interpretable" module.

More research is needed to define the null hypothesis for network analysis. A limitation in this study is that I did not consider network topology when assessing module significance. Future research can investigate accounting for the connectivity of the genes within a module when evaluating the significance of a module. The effects of this omission have not been studied in detail. More systematic studies are needed to suggest the best and most efficient solutions to this problem and to review possible biases in various approaches to account for network structure. A detailed review of the implications of significance testing in network modularity analysis, and how to standardize such testing, would be helpful to guide future work.

A general analysis of the reproducibility of miRNAs data across studies, and the current generalizability of miRNA work, would be useful. A possible approach would be to build classifiers with published miRNA signatures and apply these across published findings to judge the predictive value of these genes. Meta-analysis studies would be useful to summarize the scope of associations between miRNAs and specific cancer types and stages, as well as to help track the reproducibility of such findings. Such reviews could also provide a context for the widespread variability in past miRNA work and would help in the interpretation of these and studies.

Further laboratory validation is needed to examine the therapeutic potential of genes and interactions highlighted in this study. The connectivity and novelty of these genes in densely weighted networks suggest that they are good candidate genes for further study and provide a strong basis for experimental follow-up. These findings are presented in the context of ther neighboring interactions and functional annotation which facilitates the design of further experiments to test the relevance of their interactions and their influence in cancer

pathways. Constructing and testing hypothesis based on these genes and their interacting partners is a promising area for future studies.

Network analysis provides a powerful framework to include and analyze multi-scale genomic data. Further studies can explore the generalizability of these methods across a larger number of cancer data sets and investigate integration of additional data types; for example, mutation, SNP, methylation or transcription factor information. Further work is needed to examine network-weighting and normalization strategies to enable the integration of diverse interaction data and associated annotation information reflecting significance values, experimental measurements, and binary relations. In analyzing data across studies, methods to compare and quantify the similarity or differences between modules would also be helpful to evaluate reproducibility and consistency of module membership.

6.4: Summary

Graph and network based analyses present a unique capacity to represent and study relationships between entities in the network. These approaches are more tailored than traditional statistical and analytical methods to represent and analyze genomic complex information. Further, graph-based frameworks can incorporate attributes, including annotation, scores, and experimental measurements to nodes and edges, that are not easily included using other methods.

In this dissertation, I implement a graph-theoretical approach to model genomic data and identify network modules associated with cancer outcomes. I use a network approach to study integration of prior evidence and biological interactions in the context of cancer genomics. I examine centrality and modularity features in the network and to establish a basis for empirical work using these network properties to define cancer genes. I analyze cancer expression data using integrated interaction and miRNA regulatory data, and show that the *Walktrap* algorithm performs well against similar module discovery tools and discovers significant cancer-associated modules that highlight candidate genes for further study. These modules present potential cancer genes in the context of their biological interactions and functional annotation to better understand their relevance, build hypotheses, and design laboratory experiments. In comparison to single gene and pathway analysis, a modular approach also allows for the discovery of new genes of interest based on their relationships with more prominent cancer genes, and identification of complex genetic

interactions across pathway definitions. The results of this investigation show that using graph-based methods provides a powerful suite of tools to integrate prior evidence and study the coordinated behavior of genetic risk factors in the analyses of complex disease.

Investigation of the null hypothesis for network models, integration of more types of multi-scale data, and a systematic meta-analysis of miRNA studies would be valuable future contributions to the field of network analysis and graph-based studies. A closer examination in the laboratory of novel genes and interactions prioritized in this study is also an important next step to investigate the functional and therapeutic role of these genes of interest in breast cancer, hepatocellular carcinoma and colorectal cancer.

This work contributes to the fields of biomedical and health informatics, genomics and cancer biology an implementation and evaluation of a graph-based approach to model prior complex genomic data and to identify important genes of interest and their interactions in large-scale cancer data, where previous methods based on single candidate genes and *a priori* defined pathways have had limited success. While these methods underscore the promise of network-based research, the field is still nascent and more promising research is anticipated to apply these powerful tools to better isolate and understand the intricate genomic interactions and biological processes that underlie complex disease.

Bibliography

1. Hastie T, Tibshirani R, Friedman JH. *The elements of statistical learning : Data mining, inference, and prediction : With 200 full-color illustrations.* New York: Springer; 2001.

2. Ideker T, Sharan R. Protein networks in disease. *Genome Res*. 2008;18(4):644-652. doi: 10.1101/gr.071852.107; 10.1101/gr.071852.107.

3. Vidal M, Cusick ME, Barabasi AL. Interactome networks and human disease. *Cell*. 2011;144(6):986-998. doi: 10.1016/j.cell.2011.02.016; 10.1016/j.cell.2011.02.016.

4. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: A network-based approach to human disease. *Nat Rev Genet*. 2011;12(1):56-68. doi: 10.1038/nrg2918; 10.1038/nrg2918.

5. Petrochilos D, Abernethy N. Assessing network characteristics of cancer associated genes in metabolic and signaling networks. *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on*. 2012:290-297. doi: 10.1109/CIBCB.2012.6217243.

6. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 1999;27(1):29-34.

7. Pons P, Latapy M. Computing communities in large networks using random walks. *JGAA*. 2006;10(2):191-218.

8. Roessler S, Jia HL, Budhu A, et al. A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. *Cancer Res*. 2010;70(24):10202-10212. doi: 10.1158/0008-5472.CAN-10-2607.

9. Desmedt C, Piette F, Loi S, et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin Cancer Res*. 2007;13(11):3207-3214. doi: 10.1158/1078-0432.CCR-06-2765.

10. Sabates-Bellver J, Van der Flier LG, de Palo M, et al. Transcriptome profile of human colorectal adenomas. *Mol Cancer Res*. 2007;5(12):1263-1275. doi: 10.1158/1541-7786.MCR-07-0267.

11. Burchard J, Zhang C, Liu AM, et al. microRNA-122 as a regulator of mitochondrial metabolic gene network in hepatocellular carcinoma. *Mol Syst Biol*. 2010;6:402. doi: 10.1038/msb.2010.58; 10.1038/msb.2010.58.

12. Buffa FM, Camps C, Winchester L, et al. microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. *Cancer Res*. 2011;71(17):5635-5645. doi: 10.1158/0008-5472.CAN-11-0489; 10.1158/0008-5472.CAN-11-0489.

13. Christley RM, Pinchbeck GL, Bowers RG, et al. Infection in social networks: Using network analysis to identify high-risk individuals. *Am J Epidemiol*. 2005;162(10):1024-31.

14. Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the internet topology. *COMPUTER COMMUNICATION REVIEW*. 1999;29(4):251-262.

15. Ripeanu M, Iamnitchi A, Foster I. Mapping the gnutella network. *IEEE Internet Comput*. 2002;6(1):50-57.

16. Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. *Computer Networks & ISDN Systems*. 1998;30(1-7):1-7.

17. Nikolsky Y, Bryant J, eds. *Protein networks and pathway analysis.* Dordrecht; New York: Humana Press; 2009.

18. Chang AN. Prioritizing genes for pathway impact using network analysis. *Methods Mol Biol*. 2009;563:141-56.

19. Junker BH, Schreiber F, eds. *Analysis of biological networks.* Hoboken, N.J.: Wiley-Interscience; 2008.

20. Pavlopoulos GA, Secrier M, Moschopoulos CN, et al. Using graph theory to analyze biological networks. *BioData Min*. 2011;4:10-0381-4-10. doi: 10.1186/1756-0381-4-10; 10.1186/1756-0381-4-10.

21. Alberts B. *Molecular biology of the cell.* New York: Garland Science; 2008.

22. Weinberg RA. *The biology of cancer.* New York: Garland Science; 2007.

23. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. *Nature*. 2000;407(6804):651-4.

24. Barabási A, Albert R. Emergence of scaling in random networks. *Science*. 1999;286(5439):509-512.

25. Albert R, Hawoong Jeong, Barabasi A. Error and attack tolerance of complex networks. *Nature*. 2000;406(6794).

26. Kitano H, Omholt SW. SCIENCE'S COMPASS - BOOKS ET AL. - CELL BIOLOGY: Foundations of systems biology. *Science*. 2002;295(5563):2220.

27. Ciobanu G, Rozenberg G, eds. *Modelling in molecular biology.* Berlin; New York: Springer; 2004; No. Natural Computing Series.

28. Harris MA, Clark J, Ireland A, et al. The gene ontology (GO) database and informatics resource. *Nucleic Acids Res*. 2004;32(Database issue):D258-61. doi: 10.1093/nar/gkh036.

29. Bader GD, Betel D, Hogue CW. BIND: The biomolecular interaction network database. *Nucleic Acids Res*. 2003;31(1):248-250.

30. Robertson M. Reactome: Clear view of a starry sky. *Drug Discov Today*. 2004;9(16):684-685. doi: 10.1016/S1359-6446(04)03217-9.

31. Peri S, Navarro JD, Amanchy R, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*. 2003;13(10):2363-2371. doi: 10.1101/gr.1680803.

32. Franceschini A, Szklarczyk D, Frankild S, et al. STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*. 2013;41(Database issue):D808-15. doi: 10.1093/nar/gks1094; 10.1093/nar/gks1094.

33. Kerrien S, Aranda B, Breuza L, et al. The IntAct molecular interaction database in 2012. *Nucleic Acids Res*. 2012;40(Database issue):D841-6. doi: 10.1093/nar/gkr1088; 10.1093/nar/gkr1088.

34. Schaefer CF, Anthony K, Krupa S, et al. PID: The pathway interaction database. *Nucleic Acids Res*. 2009;37(Database issue):D674-9. doi: 10.1093/nar/gkn653.

35. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 1999;27(1):29-34.

36. http://biocarta.com/genes. Biocarta. http://biocarta.com/genes.

37. Caspi R, Altman T, Dreher K, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*. 2012;40(Database issue):D742-53. doi: 10.1093/nar/gkr1014; 10.1093/nar/gkr1014.

38. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*. 2005;33(Database issue):D514-7. doi: 10.1093/nar/gki033.

39. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA. A text-mining analysis of the human phenome. *Eur J Hum Genet*. 2006;14(5):535-542. doi: 10.1038/sj.ejhg.5201585.

40. Choi C, Krull M, Kel A, et al. TRANSPATH--a high quality database focused on signal transduction. *Comp Funct Genomics*. 2004;5(2):163-168. doi: 10.1002/cfg.386; 10.1002/cfg.386.

41. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*. 2005;120(1):15-20. doi: 10.1016/j.cell.2004.12.035.

42. Griffiths-Jones S. miRBase: The microRNA sequence database. *Methods Mol Biol*. 2006;342:129-138. doi: 10.1385/1-59745-123-1:129.

43. Grunau C, Renault E, Roizes G. DNA methylation database "MethDB": A user guide. *J Nutr*. 2002;132(8 Suppl):2435S-2439S.

44. Xin Y, Chanrion B, O'Donnell AH, et al. MethylomeDB: A database of DNA methylation profiles of the brain. *Nucleic Acids Res*. 2012;40(Database issue):D1245-9. doi: 10.1093/nar/gkr1193; 10.1093/nar/gkr1193.

45. Gnad F, Gunawardena J, Mann M. PHOSIDA 2011: The posttranslational modification database. *Nucleic Acids Res*. 2011;39(Database issue):D253-60. doi: 10.1093/nar/gkq1159; 10.1093/nar/gkq1159.

46. Miller ML, Jensen LJ, Diella F, et al. Linear motif atlas for phosphorylation-dependent signaling. *Sci Signal*. 2008;1(35):ra2. doi: 10.1126/scisignal.1159433; 10.1126/scisignal.1159433.

47. Newburger DE, Bulyk ML. UniPROBE: An online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res*. 2009;37(Database issue):D77-82. doi: 10.1093/nar/gkn660; 10.1093/nar/gkn660.

48. Portales-Casamar E, Thongjuea S, Kwon AT, et al. JASPAR 2010: The greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2010;38(Database issue):D105-10. doi: 10.1093/nar/gkp950; 10.1093/nar/gkp950.

49. Lefebvre C, Rajbhandari P, Alvarez MJ, et al. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol Syst Biol*. 2010;6:377. doi: 10.1038/msb.2010.31; 10.1038/msb.2010.31.

50. Hucka M, Finney A, Sauro HM, et al. The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*. 2003;19(4):524-531.

51. Luciano JS. PAX of mind for pathway researchers. *Drug Discov Today*. 2005;10(13):937-942. doi: 10.1016/S1359-6446(05)03501-4.

52. Shannon P, Markiel A, Ozier O, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-2504. doi: 10.1101/gr.1239303.

53. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10):R80. doi: 10.1186/gb-2004-5-10-r80.

54. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57. doi: 10.1038/nprot.2008.211.

55. Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herwig R. ConsensusPathDB: Toward a more complete picture of cell biology. *Nucleic Acids Res*. 2011;39(Database issue):D712-7. doi: 10.1093/nar/gkq1156.

56. Cowley MJ, Pinese M, Kassahn KS, et al. PINA v2.0: Mining interactome modules. *Nucleic Acids Res*. 2012;40(Database issue):D862-5. doi: 10.1093/nar/gkr967; 10.1093/nar/gkr967.

57. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res*. 2009;37(Web Server issue):W305-11. doi: 10.1093/nar/gkp427; 10.1093/nar/gkp427.

58. Newman MEJ. *Networks : An introduction.* Oxford; New York: Oxford University Press; 2010.

59. Freeman LC. Centrality in social networks conceptual clarification. *Social Networks Social Networks*. 1978;1(3):215-239.

60. Brandes U. A FASTER ALGORITHM FOR BETWEENNESS CENTRALITY. *Journal of Mathematical Sociology*. 2001;25(2).

61. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. *Nature*. 1998;393(6684):440-2.

62. Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A. The architecture of complex weighted networks. *Proc Natl Acad Sci U S A*. 2004;101(11):3747-52.

63. Fisher J, Henzinger TA. Executable cell biology. *Nat Biotechnol*. 2007;25(11):1239-1249. doi: 10.1038/nbt1356.

64. Sun N, Zhao H. Genomic approaches in dissecting complex biological pathways. *Pharmacogenomics*. 2004;5(2):163-179. doi: 10.1517/phgs.5.2.163.27488.

65. Keseler IM, Mackie A, Peralta-Gil M, et al. EcoCyc: Fusing model organism databases with systems biology. *Nucleic Acids Res*. 2013;41(Database issue):D605-12. doi: 10.1093/nar/gks1027; 10.1093/nar/gks1027.

66. Efroni S, Schaefer CF, Buetow KH. Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS ONE*. 2007;2(5):e425. doi: 10.1371/journal.pone.0000425.

67. Djebbari A, Quackenbush J. Seeded bayesian networks: Constructing genetic networks from microarray data. *BMC Syst Biol*. 2008;2:57-0509-2-57. doi: 10.1186/1752-0509-2-57; 10.1186/1752-0509-2-57.

68. Herskovits EH, Cooper GF. Algorithms for bayesian belief-network precomputation. *Methods Inf Med*. 1991;30(2):81-89.

69. Friedman N, Linial M, Nachman I, Pe'er D. Using bayesian networks to analyze expression data. *J Comput Biol*. 2000;7(3-4):601-620. doi: 10.1089/106652700750050961.

70. Imoto S, Tamada Y, Araki H, et al. Computational strategy for discovering druggable gene networks from genome-wide RNA expression profiles. *Pac Symp Biocomput*. 2006:559-571.

71. Shmulevich I, Dougherty ER, Zhang W. Gene perturbation and intervention in probabilistic boolean networks. *Bioinformatics*. 2002;18(10):1319-1331.

72. Akutsu T, Kuhara S, Maruyama O, Miyano S. A system for identifying genetic networks from gene expression patterns produced by gene disruptions and overexpressions. *Genome Inform Ser Workshop Genome Inform*. 1998;9:151-160.

73. Moore JH, Boczko EM, Summar ML. Connecting the dots between genes, biochemistry, and disease susceptibility: Systems biology modeling in human genetics. *Mol Genet Metab*. 2005;84(2):104-111. doi: 10.1016/j.ymgme.2004.10.006.

74. Peleg M, Rubin D, Altman RB. Using petri net tools to study properties and dynamics of biological systems. *J Am Med Inform Assoc*. 2005;12(2):181-199. doi: 10.1197/jamia.M1637.

75. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics*. 2004;20(1):93-99.

76. Dinu I, Potter JD, Mueller T, et al. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*. 2007;8:242. doi: 10.1186/1471-2105-8-242.

77. Bild AH, Yao G, Chang JT, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*. 2006;439(7074):353-357. doi: 10.1038/nature04296.

78. Liu D, Ghosh D, Lin X. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*. 2008;9:292. doi: 10.1186/1471-2105-9-292.

79. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550. doi: 10.1073/pnas.0506580102.

80. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003;34(3):267-273. doi: 10.1038/ng1180.

81. Backes C, Keller A, Kuentzer J, et al. GeneTrail--advanced gene set enrichment analysis. *Nucleic Acids Res*. 2007;35(Web Server issue):W186-92. doi: 10.1093/nar/gkm323.

82. Soh D, Dong D, Guo Y, Wong L. Finding consistent disease subnetworks across microarray datasets. *BMC Bioinformatics*. 2011;12 Suppl 13:S15-2105-12-S13-S15. Epub 2011 Nov 30. doi: 10.1186/1471-2105-12-S13-S15; 10.1186/1471-2105-12-S13-S15.

83. Rahnenfuhrer J, Domingues FS, Maydt J, Lengauer T. Calculating the statistical significance of changes in pathway activity from gene expression data. *Stat Appl Genet Mol Biol*. 2004;3:Article16. doi: 10.2202/1544-6115.1055.

84. Yang R, Daigle BJ,Jr, Petzold LR, Doyle FJ,3rd. Core module biomarker identification with network exploration for breast cancer metastasis. *BMC Bioinformatics*. 2012;13:12-2105-13-12. doi: 10.1186/1471-2105-13-12; 10.1186/1471-2105-13-12.

85. Shojaie A, Michailidis G. Network enrichment analysis in complex experiments. *Stat Appl Genet Mol Biol*. 2010;9(1):Article22. doi: 10.2202/1544-6115.1483.

86. Shojaie A, Michailidis G. Analysis of gene sets based on the underlying regulatory network. *J Comput Biol*. 2009;16(3):407-426. doi: 10.1089/cmb.2008.0081.

87. Segal E, Friedman N, Kaminski N, Regev A, Koller D. From signatures to models: Understanding cancer using microarrays. *Nat Genet*. 2005;37 Suppl:S38-45. doi: 10.1038/ng1561.

88. Ben-Hamo R, Efroni S. Biomarker robustness reveals the PDGF network as driving disease outcome in ovarian cancer patients in multiple studies. *BMC Syst Biol*. 2012;6:3-0509-6-3. doi: 10.1186/1752-0509-6-3; 10.1186/1752-0509-6-3.

89. Tuck DP, Kluger HM, Kluger Y. Characterizing disease states from topological properties of transcriptional regulatory networks. *BMC Bioinformatics*. 2006;7:236. doi: 10.1186/1471-2105-7-236.

90. Fox AD, Hescott BJ, Blumer AC, Slonim DK. Connectedness of PPI network neighborhoods identifies regulatory hub proteins. *Bioinformatics*. 2011;27(8):1135-1142. doi: 10.1093/bioinformatics/btr099; 10.1093/bioinformatics/btr099.

91. Garcia-Alonso L, Alonso R, Vidal E, et al. Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments. *Nucleic Acids Res*. 2012;40(20):e158. doi: 10.1093/nar/gks699; 10.1093/nar/gks699.

92. Guimera R, Sales-Pardo M, Amaral LA. A network-based method for target selection in metabolic networks. *Bioinformatics*. 2007;23(13):1616-1622. doi: 10.1093/bioinformatics/btm150.

93. Horvath S, Dong J. Geometric interpretation of gene coexpression network analysis. *PLoS Comput Biol*. 2008;4(8):e1000117. doi: 10.1371/journal.pcbi.1000117.

94. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. *Bioinformatics*. 2006;22(18):2291-2297. doi: 10.1093/bioinformatics/btl390.

95. Jonsson PF, Cavanna T, Zicha D, Bates PA. Cluster analysis of networks generated through homology: Automatic identification of important protein communities involved in cancer metastasis. *BMC Bioinformatics*. 2006;7:2. doi: 10.1186/1471-2105-7-2.

96. Sun J, Zhao Z. A comparative study of cancer proteins in the human protein-protein interaction network. *BMC Genomics*. 2010;11 Suppl 3:S5. doi: 10.1186/1471-2164-11-S3-S5.

97. Cai JJ, Borenstein E, Petrov DA. Broker genes in human disease. *Genome Biol Evol*. 2010;2:815-825. doi: 10.1093/gbe/evq064.

98. Rahmani H, Blockeel H, Bender A. Predicting genes involved in human cancer using network contextual information. *J Integr Bioinform*. 2012;9(1):210-jib-2012-210. doi: 10.2390/biecoll-jib-2012-210; 10.2390/biecoll-jib-2012-210.

99. Wang J, Chen G, Li M, Pan Y. Integration of breast cancer gene signatures based on graph centrality. *BMC Syst Biol*. 2011;5 Suppl 3:S10-0509-5-S3-S10. Epub 2011 Dec 23. doi: 10.1186/1752-0509-5-S3-S10; 10.1186/1752-0509-5-S3-S10.

100. Wachi S, Yoneda K, Wu R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics*. 2005;21(23):4205-4208. doi: 10.1093/bioinformatics/bti688.

101. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. *Proc Natl Acad Sci U S A*. 2007;104(21):8685-8690. doi: 10.1073/pnas.0701361104.

102. Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F, Furlong LI. Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PLoS One*. 2011;6(6):e20284. doi: 10.1371/journal.pone.0020284; 10.1371/journal.pone.0020284.

103. Janjic V, Przulj N. The core diseasome. *Mol Biosyst*. 2012;8(10):2614-2625. doi: 10.1039/c2mb25230a; 10.1039/c2mb25230a.

104. Jiang R, Gan M, He P. Constructing a gene semantic similarity network for the inference of disease genes. *BMC Syst Biol*. 2011;5 Suppl 2:S2-0509-5-S2-S2. Epub 2011 Dec 14. doi: 10.1186/1752-0509-5-S2-S2; 10.1186/1752-0509-5-S2-S2.

105. Xiao Y, Xu C, Guan J, et al. Discovering dysfunction of multiple microRNAs cooperation in disease by a conserved microRNA co-expression network. *PLoS One*. 2012;7(2):e32201. doi: 10.1371/journal.pone.0032201; 10.1371/journal.pone.0032201.

106. Satoh J, Tabunoki H. Comprehensive analysis of human microRNA target networks. *BioData Min*. 2011;4:17-0381-4-17. doi: 10.1186/1756-0381-4-17; 10.1186/1756-0381-4-17.

107. Xu J, Li Y. Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics*. 2006;22(22):2800-2805. doi: 10.1093/bioinformatics/btl467.

108. Zhang SH, Wu C, Li X, et al. From phenotype to gene: Detecting disease-specific gene functional modules via a text-based human disease phenotype network construction. *FEBS Lett*. 2010;584(16):3635-3643. doi: 10.1016/j.febslet.2010.07.038; 10.1016/j.febslet.2010.07.038.

109. Xu M, Kao MC, Nunez-Iglesias J, Nevins JR, West M, Zhou XJ. An integrative approach to characterize disease-specific pathways and their coordination: A case study in cancer. *BMC Genomics*. 2008;9 Suppl 1:S12-2164-9-S1-S12. doi: 10.1186/1471-2164-9-S1-S12; 10.1186/1471-2164-9-S1-S12.

110. Ruan XG, Wang JL, Li JG. A network partition algorithm for mining gene functional modules of colon cancer from DNA microarray data. *Genomics Proteomics Bioinformatics*. 2006;4(4):245-252. doi: 10.1016/S1672-0229(07)60005-9.

111. Barrenas F, Chavali S, Alves AC, et al. Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms. *Genome Biol*. 2012;13(6):R46-2012-13-6-r46. doi: 10.1186/gb-2012-13-6-r46; 10.1186/gb-2012-13-6-r46.

112. Kohler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet*. 2008;82(4):949-958. doi: 10.1016/j.ajhg.2008.02.013.

113. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. *PLoS One*. 2010;5(2):e8918. doi: 10.1371/journal.pone.0008918; 10.1371/journal.pone.0008918.

114. Kim YA, Wuchty S, Przytycka TM. Identifying causal genes and dysregulated pathways in complex diseases. *PLoS Comput Biol*. 2011;7(3):e1001095. doi: 10.1371/journal.pcbi.1001095; 10.1371/journal.pcbi.1001095.

115. Lavi O, Dror G, Shamir R. Network-induced classification kernels for gene expression profile analysis. *J Comput Biol*. 2012;19(6):694-709. doi: 10.1089/cmb.2012.0065; 10.1089/cmb.2012.0065.

116. Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. *Mol Syst Biol*. 2008;4:189. doi: 10.1038/msb.2008.27; 10.1038/msb.2008.27.

117. Li Y, Patra JC. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*. 2010;26(9):1219-1224. doi: 10.1093/bioinformatics/btq108; 10.1093/bioinformatics/btq108.

118. Wu G, Feng X, Stein L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol*. 2010;11(5):R53-2010-11-5-r53. Epub 2010 May 19. doi: 10.1186/gb-2010-11-5-r53; 10.1186/gb-2010-11-5-r53.

119. Shi M, Beauchamp RD, Zhang B. A network-based gene expression signature informs prognosis and treatment for colorectal cancer patients. *PLoS One*. 2012;7(7):e41292. doi: 10.1371/journal.pone.0041292; 10.1371/journal.pone.0041292.

120. Tu Z, Argmann C, Wong KK, et al. Integrating siRNA and protein-protein interaction data to identify an expanded insulin signaling network. *Genome Res*. 2009;19(6):1057-1067. doi: 10.1101/gr.087890.108.

121. Essaghir A, Demoulin JB. A minimal connected network of transcription factors regulated in human tumors and its application to the quest for universal cancer biomarkers. *PLoS One*. 2012;7(6):e39666. doi: 10.1371/journal.pone.0039666; 10.1371/journal.pone.0039666.

122. Lee TL, Raygada MJ, Rennert OM. Integrative gene network analysis provides novel regulatory relationships, genetic contributions and susceptible targets in autism spectrum disorders. *Gene*. 2012;496(2):88-96. doi: 10.1016/j.gene.2012.01.020; 10.1016/j.gene.2012.01.020.

123. Nibbe RK, Koyuturk M, Chance MR. An integrative -omics approach to identify functional sub-networks in human colorectal cancer. *PLoS Comput Biol*. 2010;6(1):e1000639. doi: 10.1371/journal.pcbi.1000639.

124. Pujana MA, Han JD, Starita LM, et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nat Genet*. 2007;39(11):1338-1349. doi: 10.1038/ng.2007.2.

125. Ochagavia ME, Miranda J, Nazabal M, et al. A methodology based on molecular interactions and pathways to find candidate genes associated to diseases: Its application to schizophrenia and alzheimer's disease. *J Bioinform Comput Biol*. 2011;9(4):541-557.

126. Chen Y, Wang W, Zhou Y, et al. In silico gene prioritization by integrating multiple data sources. *PLoS One*. 2011;6(6):e21137. doi: 10.1371/journal.pone.0021137; 10.1371/journal.pone.0021137.

127. Xiao Y, Xu C, Xu L, et al. Systematic identification of common functional modules related to heart failure with different etiologies. *Gene*. 2012;499(2):332-338. doi: 10.1016/j.gene.2012.03.039; 10.1016/j.gene.2012.03.039.

128. Chen L, Li W, Zhang L, et al. Disease gene interaction pathways: A potential framework for how disease genes associate by disease-risk modules. *PLoS One*. 2011;6(9):e24495. doi: 10.1371/journal.pone.0024495; 10.1371/journal.pone.0024495.

129. Rende D, Baysal N, Kirdar B. A novel integrative network approach to understand the interplay between cardiovascular disease and other complex disorders. *Mol Biosyst*. 2011;7(7):2205-2219. doi: 10.1039/c1mb05064h; 10.1039/c1mb05064h.

130. He D, Liu ZP, Chen L. Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach. *BMC Genomics*. 2011;12:592-2164-12-592. doi: 10.1186/1471-2164-12-592; 10.1186/1471-2164-12-592.

131. Berchtold LA, Storling ZM, Ortis F, et al. Huntingtin-interacting protein 14 is a type 1 diabetes candidate protein regulating insulin secretion and beta-cell apoptosis. *Proc Natl Acad Sci U S A*. 2011;108(37):E681-8. doi: 10.1073/pnas.1104384108; 10.1073/pnas.1104384108.

132. Califano A, Butte AJ, Friend S, Ideker T, Schadt E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat Genet*. 2012;44(8):841-847. doi: 10.1038/ng.2355; 10.1038/ng.2355.

133. Liu H, Su J, Li J, et al. Prioritizing cancer-related genes with aberrant methylation based on a weighted protein-protein interaction network. *BMC Syst Biol*. 2011;5:158-0509-5-158. doi: 10.1186/1752-0509-5-158; 10.1186/1752-0509-5-158.

134. Tu Z, Wang L, Arbeitman MN, Chen T, Sun F. An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics*. 2006;22(14):e489-96. doi: 10.1093/bioinformatics/btl234.

135. Essaghir A, Demoulin JB. A minimal connected network of transcription factors regulated in human tumors and its application to the quest for universal cancer biomarkers. *PLoS One*. 2012;7(6):e39666. doi: 10.1371/journal.pone.0039666; 10.1371/journal.pone.0039666.

136. Maulik U, Bhattacharyya M, Mukhopadhyay A, Bandyopadhyay S. Identifying the immunodeficiency gateway proteins in humans and their involvement in microRNA regulation. *Mol Biosyst*. 2011;7(6):1842-1851. doi: 10.1039/c1mb05026e; 10.1039/c1mb05026e.

137. Bandyopadhyay S, Mitra R, Maulik U, Zhang MQ. Development of the human cancer microRNA network. *Silence*. 2010;1(1):6-907X-1-6. doi: 10.1186/1758-907X-1-6; 10.1186/1758-907X-1-6.

138. Ooi CH, Oh HK, Wang HZ, et al. A densely interconnected genome-wide network of microRNAs and oncogenic pathways revealed using gene expression signatures. *PLoS Genet*. 2011;7(12):e1002415. doi: 10.1371/journal.pgen.1002415; 10.1371/journal.pgen.1002415.

139. Sun J, Gong X, Purow B, Zhao Z. Uncovering MicroRNA and transcription factor mediated regulatory networks in glioblastoma. *PLoS Comput Biol*. 2012;8(7):e1002488. doi: 10.1371/journal.pcbi.1002488; 10.1371/journal.pcbi.1002488.

140. Li W, Dai C, Liu CC, Zhou XJ. Algorithm to identify frequent coupled modules from two-layered network series: Application to study transcription and splicing coupling. *J Comput Biol*. 2012;19(6):710-730. doi: 10.1089/cmb.2012.0025; 10.1089/cmb.2012.0025.

141. Kosti I, Radivojac P, Mandel-Gutfreund Y. An integrated regulatory network reveals pervasive cross-regulation among transcription and splicing factors. *PLoS Comput Biol*. 2012;8(7):e1002603. doi: 10.1371/journal.pcbi.1002603; 10.1371/journal.pcbi.1002603.

142. Li W, Wang R, Bai L, Yan Z, Sun Z. Cancer core modules identification through genomic and transcriptomic changes correlation detection at network level. *BMC Syst Biol*. 2012;6:64-0509-6-64. doi: 10.1186/1752-0509-6-64; 10.1186/1752-0509-6-64.

143. Mani KM, Lefebvre C, Wang K, et al. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Mol Syst Biol*. 2008;4:169. doi: 10.1038/msb.2008.2; 10.1038/msb.2008.2.

144. Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res*. 2012;22(2):398-406. doi: 10.1101/gr.125567.111; 10.1101/gr.125567.111.

145. Gu Y, Wang H, Qin Y, et al. Network analysis of genomic alteration profiles reveals co-altered functional modules and driver genes for glioblastoma. *Mol Biosyst*. 2013;9(3):467-477. doi: 10.1039/c2mb25528f; 10.1039/c2mb25528f.

146. Zhang B, Shi Z, Duncan DT, Prodduturi N, Marnett LJ, Liebler DC. Relating protein adduction to gene expression changes: A systems approach. *Mol Biosyst*. 2011;7(7):2118-2127. doi: 10.1039/c1mb05014a; 10.1039/c1mb05014a.

147. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. *Nature*. 1999;402(6761 Suppl):C47-52. doi: 10.1038/35011540.

148. Huang Y, Li H, Hu H, et al. Systematic discovery of functional modules and context-specific functional annotation of human genome. *Bioinformatics*. 2007;23(13):i222-9. doi: 10.1093/bioinformatics/btm222.

149. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*. 2002;18 Suppl 1:S233-40.

150. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Mol Syst Biol*. 2007;3:140. doi: 10.1038/msb4100180.

151. Ulitsky I, Shamir R. Identification of functional modules using network topology and high-throughput data. *BMC Syst Biol*. 2007;1:8. doi: 10.1186/1752-0509-1-8.

152. Girvan M, Newman ME. Community structure in social and biological networks. *Proc Natl Acad Sci U S A*. 2002;99(12):7821-7826. doi: 10.1073/pnas.122653799.

153. Orman GK, Labatut V. Relative evaluation of partition algorithms for complex networks. *Networked Digital Technologies, 2009 NDT '09 First International Conference on*. 2009:20-25.

154. Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. *Nat Genet*. 2004;36(10):1090-1098. doi: 10.1038/ng1434.

155. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T. Identifying functional modules in protein-protein interaction networks: An integrated exact approach. *Bioinformatics*. 2008;24(13):i223-31. doi: 10.1093/bioinformatics/btn161.

156. Zhao XM, Wang RS, Chen L, Aihara K. Uncovering signal transduction networks from high-throughput data by integer linear programming. *Nucleic Acids Res*. 2008;36(9):e48. doi: 10.1093/nar/gkn145; 10.1093/nar/gkn145.

157. Backes C, Rurainski A, Klau GW, et al. An integer linear programming approach for finding deregulated subgraphs in regulatory networks. *Nucleic Acids Res*. 2012;40(6):e43. doi: 10.1093/nar/gkr1227; 10.1093/nar/gkr1227.

158. Chuang HY, Rassenti L, Salcedo M, et al. Subnetwork-based analysis of chronic lymphocytic leukemia identifies pathways that associate with disease progression. *Blood*. 2012;120(13):2639-2649. doi: 10.1182/blood-2012-03-416461.

159. Alcaraz N, Friedrich T, Kotzing T, et al. Efficient key pathway mining: Combining networks and OMICS data. *Integr Biol (Camb)*. 2012;4(7):756-764. doi: 10.1039/c2ib00133k; 10.1039/c2ib00133k.

160. Diez D, Goto S, Fahy JV, et al. Network analysis identifies a putative role for the PPAR and type 1 interferon pathways in glucocorticoid actions in asthmatics. *BMC Med Genomics*. 2012;5:27-8794-5-27. doi: 10.1186/1755-8794-5-27; 10.1186/1755-8794-5-27.

161. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol*. 2011;18(3):507-522. doi: 10.1089/cmb.2010.0265; 10.1089/cmb.2010.0265.

162. Nicolau M, Levine AJ, Carlsson G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc Natl Acad Sci U S A*. 2011;108(17):7265-7270. doi: 10.1073/pnas.1102826108; 10.1073/pnas.1102826108.

163. Liu X, Liu ZP, Zhao XM, Chen L. Identifying disease genes and module biomarkers by differential interactions. *J Am Med Inform Assoc*. 2012;19(2):241-248. doi: 10.1136/amiajnl-2011-000658; 10.1136/amiajnl-2011-000658.

164. Navlakha S, Kingsford C. The power of protein interaction networks for associating genes with diseases. *Bioinformatics*. 2010;26(8):1057-1063. doi: 10.1093/bioinformatics/btq076; 10.1093/bioinformatics/btq076.

165. Li Y, Patra JC. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*. 2010;26(9):1219-1224. doi: 10.1093/bioinformatics/btq108; 10.1093/bioinformatics/btq108.

166. Yao X, Hao H, Li Y, Li S. Modularity-based credible prediction of disease genes and detection of disease subtypes on the phenotype-gene heterogeneous network. *BMC Syst Biol*. 2011;5:79-0509-5-79. doi: 10.1186/1752-0509-5-79; 10.1186/1752-0509-5-79.

167. Tu Z, Wang L, Arbeitman MN, Chen T, Sun F. An integrative approach for causal gene identification and gene regulatory pathway inference. *Bioinformatics*. 2006;22(14):e489-96. doi: 10.1093/bioinformatics/btl234.

168. Komurov K, White MA, Ram PT. Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data. *PLoS Comput Biol*. 2010;6(8):e1000889. doi: 10.1371/journal.pcbi.1000889.

169. Komurov K, Dursun S, Erdin S, Ram PT. NetWalker: A contextual network analysis tool for functional genomics. *BMC Genomics*. 2012;13:282-2164-13-282. doi: 10.1186/1471-2164-13-282; 10.1186/1471-2164-13-282.

170. Csardi G, Nepusz. T. The igraph software package for complex network research. *InterJournal*. 2006;Complex Systems:1695.

171. Sun S, Dong X, Fu Y, Tian W. An iterative network partition algorithm for accurate identification of dense network modules. *Nucleic Acids Res*. 2012;40(3):e18. doi: 10.1093/nar/gkr1103; 10.1093/nar/gkr1103.

172. Keller A, Backes C, Gerasch A, et al. A novel algorithm for detecting differentially regulated paths based on gene set enrichment analysis. *Bioinformatics*. 2009;25(21):2787-2794. doi: 10.1093/bioinformatics/btp510.

173. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550. doi: 10.1073/pnas.0506580102.

174. Cui Q, Ma Y, Jaramillo M, et al. A map of human cancer signaling. *Mol Syst Biol*. 2007;3:152. doi: 10.1038/msb4100200.

175. Taylor IW, Linding R, Warde-Farley D, et al. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol*. 2009;27(2):199-204. doi: 10.1038/nbt.1522.

176. Li L, Zhang K, Lee J, Cordes S, Davis DP, Tang Z. Discovering cancer genes by integrating network and functional properties. *BMC Med Genomics*. 2009;2:61. doi: 10.1186/1755-8794-2-61.

177. Tu Z, Argmann C, Wong KK, et al. Integrating siRNA and protein-protein interaction data to identify an expanded insulin signaling network. *Genome Res*. 2009. doi: 10.1101/gr.087890.108.

178. del Rio G, Koschutzki D, Coello G. How to identify essential genes from molecular networks? *BMC Syst Biol*. 2009;3:102. doi: 10.1186/1752-0509-3-102.

179. Petrochilos D, Abernethy N. Assessing network characteristics of cancer associated genes in metabolic and signaling networks. *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on*. 2012:290-297.

180. Rodrigues FA, Ferraz de Arruda G, da Fontoura Costa L. A complex networks approach for data clustering. *ArXiv e-prints*. 2011. http://adsabs.harvard.edu/abs/2011arXiv1101.5141R.

181. Barrett T, Suzek TO, Troup DB, et al. NCBI GEO: Mining millions of expression profiles--database and tools. *Nucleic Acids Res*. 2005;33(Database issue):D562-6. doi: 10.1093/nar/gki022.

182. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *JSTOR*. 1995;57(1):289-300.

183. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: Pathway editing for the people. *PLoS Biol*. 2008;6(7):e184. doi: 10.1371/journal.pbio.0060184.

184. Wang H, Bauzon F, Ji P, et al. Skp2 is required for survival of aberrantly proliferating Rb1-deficient cells and for tumorigenesis in Rb1+/- mice. *Nat Genet*. 2010;42(1):83-88. doi: 10.1038/ng.498.

185. Kubota N, Tobe K, Terauchi Y, et al. Disruption of insulin receptor substrate 2 causes type 2 diabetes because of liver insulin resistance and lack of compensatory beta-cell hyperplasia. *Diabetes*. 2000;49(11):1880-1889.

186. Shaoul R, Eliahu L, Sher I, et al. Elevated expression of FGF7 protein is common in human gastric diseases. *Biochem Biophys Res Commun*. 2006;350(4):825-833. doi: 10.1016/j.bbrc.2006.08.198.

187. Huang SP, Bao BY, Hour TC, et al. Genetic variants in CASP3, BMP5, and IRS2 genes may influence survival in prostate cancer patients receiving androgen-deprivation therapy. *PLoS One*. 2012;7(7):e41219. doi: 10.1371/journal.pone.0041219.

188. Bonte D, Lindvall C, Liu H, Dykema K, Furge K, Weinreich M. Cdc7-Dbf4 kinase overexpression in multiple cancers and tumor cell lines is correlated with p53 inactivation. *Neoplasia*. 2008;10(9):920-931.

189. Mincheva A, Todorov I, Werner D, Fink TM, Lichter P. The human gene for nuclear protein BM28 (CDCL1), a new member of the early S-phase family of proteins, maps to chromosome band 3q21. *Cytogenet Cell Genet*. 1994;65(4):276-277.

190. Hankinson SE, Willett WC, Colditz GA, et al. Circulating concentrations of insulin-like growth factor-I and risk of breast cancer. *Lancet*. 1998;351(9113):1393-1396. doi: 10.1016/S0140-6736(97)10384-1.

191. Hauge C, Frodin M. RSK and MSK in MAP kinase signalling. *J Cell Sci*. 2006;119(Pt 15):3021-3023. doi: 10.1242/jcs.02950.

192. Bignone PA, Lee KY, Liu Y, et al. RPS6KA2, a putative tumour suppressor gene at 6q27 in sporadic epithelial ovarian cancer. *Oncogene*. 2007;26(5):683-700. doi: 10.1038/sj.onc.1209827.

193. Carro MS, Lim WK, Alvarez MJ, et al. The transcriptional network for mesenchymal transformation of brain tumours. *Nature*. 2010;463(7279):318-325. doi: 10.1038/nature08712; 10.1038/nature08712.

194. Tanaka T, Akira S, Yoshida K, et al. Targeted disruption of the NF-IL6 gene discloses its essential role in bacteria killing and tumor cytotoxicity by macrophages. *Cell*. 1995;80(2):353-361.

195. Hong D, Gupta R, Ancliff P, et al. Initiating and cancer-propagating cells in TEL-AML1-associated childhood leukemia. *Science*. 2008;319(5861):336-339. doi: 10.1126/science.1150648.

196. Rosenbauer F, Owens BM, Yu L, et al. Lymphoid cell growth and transformation are suppressed by a key regulatory element of the gene encoding PU.1. *Nat Genet*. 2006;38(1):27-37. doi: 10.1038/ng1679.

197. Melkonyan HS, Chang WC, Shapiro JP, et al. SARPs: A family of secreted apoptosis-related proteins. *Proc Natl Acad Sci U S A*. 1997;94(25):13636-13641.

198. Satoh J. Molecular network analysis of human microRNA targetome: From cancers to alzheimer's disease. *BioData Min*. 2012;5(1):17-0381-5-17. doi: 10.1186/1756-0381-5-17; 10.1186/1756-0381-5-17.

199. Ryan BM, Robles AI, Harris CC. Genetic variation in microRNA networks: The implications for cancer research. *Nat Rev Cancer*. 2010;10(6):389-402. doi: 10.1038/nrc2867; 10.1038/nrc2867.

200. Kozomara A, Griffiths-Jones S. miRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*. 2011;39(Database issue):D152-7. doi: 10.1093/nar/gkq1027; 10.1093/nar/gkq1027.

201. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*. 2009;19(1):92-105. doi: 10.1101/gr.082701.108; 10.1101/gr.082701.108.

202. O'Day E, Lal A. MicroRNAs and their target gene networks in breast cancer. *Breast Cancer Res*. 2010;12(2):201. doi: 10.1186/bcr2484; 10.1186/bcr2484.

203. Laios A, O'Toole S, Flavin R, et al. Potential role of miR-9 and miR-223 in recurrent ovarian cancer. *Mol Cancer*. 2008;7:35-4598-7-35. doi: 10.1186/1476-4598-7-35; 10.1186/1476-4598-7-35.

204. Gennarino VA, D'Angelo G, Dharmalingam G, et al. Identification of microRNA-regulated gene networks by expression analysis of target genes. *Genome Res*. 2012;22(6):1163-1172. doi: 10.1101/gr.130435.111; 10.1101/gr.130435.111.

205. Nam S, Long X, Kwon C, Kim S, Nephew KP. An integrative analysis of cellular contexts, miRNAs and mRNAs reveals network clusters associated with antiestrogen-resistant breast cancer cells. *BMC Genomics*. 2012;13(1):732. doi: 10.1186/1471-2164-13-732.

206. Zhang W, Edwards A, Fan W, Flemington EK, Zhang K. miRNA-mRNA correlation-network modules in human prostate cancer and the differences between primary and metastatic tumor subtypes. *PLoS One*. 2012;7(6):e40130. doi: 10.1371/journal.pone.0040130; 10.1371/journal.pone.0040130.

207. Gu Z, Zhang C, Wang J. Gene regulation is governed by a core network in hepatocellular carcinoma. *BMC Syst Biol*. 2012;6:32-0509-6-32. doi: 10.1186/1752-0509-6-32; 10.1186/1752-0509-6-32.

208. Hansen BB, Klopfer SO. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics Journal of Computational and Graphical Statistics*. 2006;15(3):609-627.

209. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*. 2000;16(5):412-424.

210. Murphy M, Pykett MJ, Harnish P, Zang KD, George DL. Identification and characterization of genes differentially expressed in meningiomas. *Cell Growth Differ*. 1993;4(9):715-722.

211. Butt AJ, Dickson KA, McDougall F, Baxter RC. Insulin-like growth factor-binding protein-5 inhibits the growth of human breast cancer cells in vitro and in vivo. *J Biol Chem*. 2003;278(32):29676-29685. doi: 10.1074/jbc.M301965200.

212. Pazaitou-Panayiotou K, Kelesidis T, Kelesidis I, et al. Growth hormone-binding protein is directly and IGFBP-3 is inversely associated with risk of female breast cancer. *Eur J Endocrinol*. 2007;156(2):187-194. doi: 10.1530/EJE-06-0611.

213. Rinaldi S, Peeters PH, Berrino F, et al. IGF-I, IGFBP-3 and breast cancer risk in women: The european prospective investigation into cancer and nutrition (EPIC). *Endocr Relat Cancer*. 2006;13(2):593-605. doi: 10.1677/erc.1.01150.

214. Bilal E, Alexe G, Yao M, et al. Identification of the YES1 kinase as a therapeutic target in basal-like breast cancers. *Genes Cancer*. 2010;1(10):1063-1073. doi: 10.1177/1947601910395583; 10.1177/1947601910395583.

215. Tian Q, Frierson HF,Jr, Krystal GW, Moskaluk CA. Activating c-kit gene mutations in human germ cell tumors. *Am J Pathol*. 1999;154(6):1643-1647. doi: 10.1016/S0002-9440(10)65419-3.

216. Regan JL, Kendrick H, Magnay FA, Vafaizadeh V, Groner B, Smalley MJ. C-kit is required for growth and survival of the cells of origin of Brca1-mutation-associated breast cancer. *Oncogene*. 2012;31(7):869-883. doi: 10.1038/onc.2011.289; 10.1038/onc.2011.289.

217. Tamimi RM, Brugge JS, Freedman ML, et al. Circulating colony stimulating factor-1 and breast cancer risk. *Cancer Res*. 2008;68(1):18-21. doi: 10.1158/0008-5472.CAN-07-3234; 10.1158/0008-5472.CAN-07-3234.

218. Zrihan-Licht S, Lim J, Keydar I, Sliwkowski MX, Groopman JE, Avraham H. Association of csk-homologous kinase (CHK) (formerly MATK) with HER-2/ErbB-2 in breast cancer cells. *J Biol Chem*. 1997;272(3):1856-1863.

219. Fallon L, Belanger CM, Corera AT, et al. A regulated interaction with the UIM protein Eps15 implicates parkin in EGF receptor trafficking and PI(3)K-akt signalling. *Nat Cell Biol*. 2006;8(8):834-842. doi: 10.1038/ncb1441.

220. Gao M, Labuda T, Xia Y, et al. Jun turnover is controlled through JNK-dependent phosphorylation of the E3 ligase itch. *Science*. 2004;306(5694):271-275. doi: 10.1126/science.1099414.

221. Wang LC, Swat W, Fujiwara Y, et al. The TEL/ETV6 gene is required specifically for hematopoiesis in the bone marrow. *Genes Dev*. 1998;12(15):2392-2402.

222. Guidez F, Parks S, Wong H, et al. RARalpha-PLZF overcomes PLZF-mediated repression of CRABPI, contributing to retinoid resistance in t(11;17) acute promyelocytic leukemia. *Proc Natl Acad Sci U S A*. 2007;104(47):18694-18699. doi: 10.1073/pnas.0704433104.

223. Erickson P, Gao J, Chang KS, et al. Identification of breakpoints in t(8;21) acute myelogenous leukemia and isolation of a fusion transcript, AML1/ETO, with similarity to drosophila segmentation gene, runt. *Blood*. 1992;80(7):1825-1831.

224. Yamamoto Y, Tsuzuki S, Tsuzuki M, Handa K, Inaguma Y, Emi N. BCOR as a novel fusion partner of retinoic acid receptor alpha in a t(X;17)(p11;q12) variant of acute

promyelocytic leukemia. *Blood*. 2010;116(20):4274-4283. doi: 10.1182/blood-2010-01-264432; 10.1182/blood-2010-01-264432.

225. Pierron G, Tirode F, Lucchesi C, et al. A new subtype of bone sarcoma defined by BCOR-CCNB3 gene fusion. *Nat Genet*. 2012;44(4):461-466. doi: 10.1038/ng.1107; 10.1038/ng.1107.

226. Cattoretti G, Chang CC, Cechova K, et al. BCL-6 protein is expressed in germinal-center B cells. *Blood*. 1995;86(1):45-53.

227. Schick N, Oakeley EJ, Hynes NE, Badache A. TEL/ETV6 is a signal transducer and activator of transcription 3 (Stat3)-induced repressor of Stat3 activity. *J Biol Chem*. 2004;279(37):38787-38796. doi: 10.1074/jbc.M312581200.

228. Stegmaier K, Pendse S, Barker GF, et al. Frequent loss of heterozygosity at the TEL gene locus in acute lymphoblastic leukemia of childhood. *Blood*. 1995;86(1):38-44.

229. Anderson K, Lutz C, van Delft FW, et al. Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature*. 2011;469(7330):356-361. doi: 10.1038/nature09650; 10.1038/nature09650.

230. Fleischer TC, Yun UJ, Ayer DE. Identification and characterization of three new components of the mSin3A corepressor complex. *Mol Cell Biol*. 2003;23(10):3456-3467.

231. Shinagawa T, Nomura T, Colmenares C, Ohira M, Nakagawara A, Ishii S. Increased susceptibility to tumorigenesis of ski-deficient heterozygous mice. *Oncogene*. 2001;20(56):8100-8108. doi: 10.1038/sj.onc.1204987.

232. Cundell DR, Gerard NP, Gerard C, Idanpaan-Heikkila I, Tuomanen EI. Streptococcus pneumoniae anchor to activated human cells by the receptor for platelet-activating factor. *Nature*. 1995;377(6548):435-438. doi: 10.1038/377435a0.

233. Van Raamsdonk CD, Griewank KG, Crosby MB, et al. Mutations in GNA11 in uveal melanoma. *N Engl J Med*. 2010;363(23):2191-2199. doi: 10.1056/NEJMoa1000584; 10.1056/NEJMoa1000584.

234. Allen LF, Lefkowitz RJ, Caron MG, Cotecchia S. G-protein-coupled receptor genes as protooncogenes: Constitutively activating mutation of the alpha 1B-adrenergic receptor enhances mitogenesis and tumorigenicity. *Proc Natl Acad Sci U S A*. 1991;88(24):11354-11358.

235. Boire A, Covic L, Agarwal A, Jacques S, Sherifi S, Kuliopulos A. PAR1 is a matrix metalloprotease-1 receptor that promotes invasion and tumorigenesis of breast cancer cells. *Cell*. 2005;120(3):303-313. doi: 10.1016/j.cell.2004.12.018.

236. Buckanovich RJ, Facciabene A, Kim S, et al. Endothelin B receptor mediates the endothelial barrier to T cell homing to tumors and disables immune therapy. *Nat Med*. 2008;14(1):28-36. doi: 10.1038/nm1699.

237. Stoyanov B, Volinia S, Hanck T, et al. Cloning and characterization of a G protein-activated human phosphoinositide-3 kinase. *Science*. 1995;269(5224):690-693.

238. Ali K, Bilancio A, Thomas M, et al. Essential role for the p110delta phosphoinositide 3-kinase in the allergic response. *Nature*. 2004;431(7011):1007-1011. doi: 10.1038/nature02991.

239. Weng L, Brown J, Eng C. PTEN induces apoptosis and cell cycle arrest through phosphoinositol-3-kinase/akt-dependent and -independent pathways. *Hum Mol Genet*. 2001;10(3):237-242.

240. Yadav V, Denning MF. Fyn is induced by ras/PI3K/akt signaling and is required for enhanced invasion/migration. *Mol Carcinog*. 2011;50(5):346-352. doi: 10.1002/mc.20716; 10.1002/mc.20716.

241. Berger AH, Niki M, Morotti A, et al. Identification of DOK genes as lung tumor suppressors. *Nat Genet*. 2010;42(3):216-223. doi: 10.1038/ng.527; 10.1038/ng.527.

242. Bustelo XR. Regulatory and signaling properties of the vav family. *Mol Cell Biol*. 2000;20(5):1461-1477.

243. Citterio C, Menacho-Marquez M, Garcia-Escudero R, et al. The rho exchange factors vav2 and vav3 control a lung metastasis-specific transcriptional program in breast cancer cells. *Sci Signal*. 2012;5(244):ra71. doi: 10.1126/scisignal.2002962; 10.1126/scisignal.2002962.

244. Pappu R, Cheng AM, Li B, et al. Requirement for B cell linker protein (BLNK) in B cell development. *Science*. 1999;286(5446):1949-1954.

245. Flemming A, Brummer T, Reth M, Jumaa H. The adaptor protein SLP-65 acts as a tumor suppressor that limits pre-B cell expansion. *Nat Immunol*. 2003;4(1):38-43. doi: 10.1038/ni862.

246. Hu MC, Qiu WR, Wang X, Meyer CF, Tan TH. Human HPK1, a novel human hematopoietic progenitor kinase that activates the JNK/SAPK kinase cascade. *Genes Dev*. 1996;10(18):2251-2264.

247. Kiefer F, Tibbles LA, Anafi M, et al. HPK1, a hematopoietic protein kinase activating the SAPK/JNK pathway. *EMBO J*. 1996;15(24):7013-7025.

248. de Fraipont F, El Atifi M, Gicquel C, Bertagna X, Chambaz EM, Feige JJ. Expression of the angiogenesis markers vascular endothelial growth factor-A, thrombospondin-1, and platelet-derived endothelial cell growth factor in human sporadic adrenocortical tumors: Correlation with genotypic alterations. *J Clin Endocrinol Metab*. 2000;85(12):4734-4741.

249. Rodriguez-Manzaneque JC, Lane TF, Ortega MA, Hynes RO, Lawler J, Iruela-Arispe ML. Thrombospondin-1 suppresses spontaneous tumor growth and inhibits activation of matrix metalloproteinase-9 and mobilization of vascular endothelial growth factor. *Proc Natl Acad Sci U S A*. 2001;98(22):12485-12490. doi: 10.1073/pnas.171460498.

250. Matsunaga T, Takemoto N, Sato T, et al. Interaction between leukemic-cell VLA-4 and stromal fibronectin is a decisive factor for minimal residual disease of acute myelogenous leukemia. *Nat Med*. 2003;9(9):1158-1165. doi: 10.1038/nm909.

251. Minn AJ, Gupta GP, Siegel PM, et al. Genes that mediate breast cancer metastasis to lung. *Nature*. 2005;436(7050):518-524. doi: 10.1038/nature03799.

252. Simon MP, Pedeutour F, Sirvent N, et al. Deregulation of the platelet-derived growth factor B-chain gene via fusion with collagen gene COL1A1 in dermatofibrosarcoma protuberans and giant-cell fibroblastoma. *Nat Genet*. 1997;15(1):95-98. doi: 10.1038/ng0197-95.

253. Abeysinghe HR, Cao Q, Xu J, et al. THY1 expression is associated with tumor suppression of human ovarian cancer. *Cancer Genet Cytogenet*. 2003;143(2):125-132.

254. Qu S, Yao Y, Shang C, et al. MicroRNA-330 is an oncogenic factor in glioblastoma cells by regulating SH3GL2 gene. *PLoS One*. 2012;7(9):e46010. doi: 10.1371/journal.pone.0046010; 10.1371/journal.pone.0046010.

255. Zou C, Ma J, Wang X, et al. Lack of fas antagonism by met in human fatty liver disease. *Nat Med*. 2007;13(9):1078-1085. doi: 10.1038/nm1625.

256. Silber J, Jacobsen A, Ozawa T, et al. miR-34a repression in proneural malignant gliomas upregulates expression of its target PDGFRA and promotes tumorigenesis. *PLoS One*. 2012;7(3):e33844. doi: 10.1371/journal.pone.0033844; 10.1371/journal.pone.0033844.

257. Park SM, Gaur AB, Lengyel E, Peter ME. The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. *Genes Dev*. 2008;22(7):894-907. doi: 10.1101/gad.1640608; 10.1101/gad.1640608.

258. Deng Y, Deng H, Bi F, et al. MicroRNA-137 targets carboxyl-terminal binding protein 1 in melanoma cell lines. *Int J Biol Sci*. 2011;7(1):133-137.

259. Yan J, Zhu J, Zhong H, Lu Q, Huang C, Ye Q. BRCA1 interacts with FHL2 and enhances FHL2 transactivation function. *FEBS Lett*. 2003;553(1-2):183-189.

260. Ng CF, Ng PK, Lui VW, et al. FHL2 exhibits anti-proliferative and anti-apoptotic activities in liver cancer cells. *Cancer Lett*. 2011;304(2):97-106. doi: 10.1016/j.canlet.2011.02.001; 10.1016/j.canlet.2011.02.001.

261. Park JH, Lin ML, Nishidate T, Nakamura Y, Katagiri T. PDZ-binding kinase/T-LAK cell-originated protein kinase, a putative cancer/testis antigen with an oncogenic activity in breast cancer. *Cancer Res*. 2006;66(18):9186-9195. doi: 10.1158/0008-5472.CAN-06-1601.

262. Boxall S, Stanton T, Hirai K, et al. Disease associations and altered immune function in CD45 138G variant carriers. *Hum Mol Genet*. 2004;13(20):2377-2384. doi: 10.1093/hmg/ddh276.

263. Dawes R, Hennig B, Irving W, et al. Altered CD45 expression in C77G carriers influences immune function and outcome of hepatitis C infection. *J Med Genet*. 2006;43(8):678-684. doi: 10.1136/jmg.2005.040485.

264. Irie-Sasaki J, Sasaki T, Matsumoto W, et al. CD45 is a JAK phosphatase and negatively regulates cytokine receptor signalling. *Nature*. 2001;409(6818):349-354. doi: 10.1038/35053086.

265. Marie-Cardine A, Bruyns E, Eckerskorn C, Kirchgessner H, Meuer SC, Schraven B. Molecular cloning of SKAP55, a novel protein that associates with the protein tyrosine kinase p59fyn in human T-lymphocytes. *J Biol Chem*. 1997;272(26):16077-16080.

266. Raab M, Smith X, Matthess Y, Strebhardt K, Rudd CE. SKAP1 protein PH domain determines RapL membrane localization and Rap1 protein complex formation for T cell receptor (TCR) activation of LFA-1. *J Biol Chem*. 2011;286(34):29663-29670. doi: 10.1074/jbc.M111.222661; 10.1074/jbc.M111.222661.

267. Geng L, Pfister S, Kraeft SK, Rudd CE. Adaptor FYB (fyn-binding protein) regulates integrin-mediated adhesion and mediator release: Differential involvement of the FYB SH3 domain. *Proc Natl Acad Sci U S A*. 2001;98(20):11527-11532. doi: 10.1073/pnas.191378198.

268. Papetti M, Augenlicht LH. MYBL2, a link between proliferation and differentiation in maturing colon epithelial cells. *J Cell Physiol*. 2011;226(3):785-791. doi: 10.1002/jcp.22399; 10.1002/jcp.22399.

269. Mailand N, Falck J, Lukas C, et al. Rapid destruction of human Cdc25A in response to DNA damage. *Science*. 2000;288(5470):1425-1429.

270. Latres E, Chiarle R, Schulman BA, et al. Role of the F-box protein Skp2 in lymphomagenesis. *Proc Natl Acad Sci U S A*. 2001;98(5):2515-2520. doi: 10.1073/pnas.041475098.

271. Lin HK, Chen Z, Wang G, et al. Skp2 targeting suppresses tumorigenesis by arf-p53-independent cellular senescence. *Nature*. 2010;464(7287):374-379. doi: 10.1038/nature08815; 10.1038/nature08815.

272. Frau M, Ladu S, Calvisi DF, et al. Mybl2 expression is under genetic control and contributes to determine a hepatocellular carcinoma susceptible phenotype. *J Hepatol*. 2011;55(1):111-119. doi: 10.1016/j.jhep.2010.10.031; 10.1016/j.jhep.2010.10.031.

273. Calvisi DF, Simile MM, Ladu S, et al. Activation of v-myb avian myeloblastosis viral oncogene homolog-like2 (MYBL2)-LIN9 complex contributes to human hepatocarcinogenesis and identifies a subset of hepatocellular carcinoma with mutant p53. *Hepatology*. 2011;53(4):1226-1236. doi: 10.1002/hep.24174; 10.1002/hep.24174.

274. Nakajima T, Yasui K, Zen K, et al. Activation of B-myb by E2F1 in hepatocellular carcinoma. *Hepatol Res*. 2008;38(9):886-895. doi: 10.1111/j.1872-034X.2008.00324.x; 10.1111/j.1872-034X.2008.00324.x.

275. Wang X, Quail E, Hung NJ, Tan Y, Ye H, Costa RH. Increased levels of forkhead box M1B transcription factor in transgenic mouse hepatocytes prevent age-related proliferation defects in regenerating liver. *Proc Natl Acad Sci U S A*. 2001;98(20):11468-11473. doi: 10.1073/pnas.201360898.

276. Kalinichenko VV, Major ML, Wang X, et al. Foxm1b transcription factor is essential for development of hepatocellular carcinomas and is negatively regulated by the p19ARF tumor suppressor. *Genes Dev*. 2004;18(7):830-850. doi: 10.1101/gad.1200704.

277. Palmer CN, Richardson TH, Griffin KJ, et al. Characterization of a cDNA encoding a human kidney, cytochrome P-450 4A fatty acid omega-hydroxylase and the cognate enzyme expressed in escherichia coli. *Biochim Biophys Acta*. 1993;1172(1-2):161-166.

278. Powell PK, Wolf I, Lasker JM. Identification of CYP4A11 as the major lauric acid omega-hydroxylase in human liver microsomes. *Arch Biochem Biophys*. 1996;335(1):219-226. doi: 10.1006/abbi.1996.0501.

279. Chen X, Wang H, Xie W, et al. Association of CYP1A2 genetic polymorphisms with hepatocellular carcinoma susceptibility: A case-control study in a high-risk region of china. *Pharmacogenet Genomics*. 2006;16(3):219-227. doi: 10.1097/01.fpc.0000194424.20393.c6.

280. Mihelich BL, Khramtsova EA, Arva N, et al. miR-183-96-182 cluster is overexpressed in prostate tissue and regulates zinc homeostasis in prostate cells. *J Biol Chem*. 2011;286(52):44503-44511. doi: 10.1074/jbc.M111.262915; 10.1074/jbc.M111.262915.

281. Loukola A, Chadha M, Penn SG, et al. Comprehensive evaluation of the association between prostate cancer and genotypes/haplotypes in CYP17A1, CYP3A4, and SRD5A2. *Eur J Hum Genet*. 2004;12(4):321-332. doi: 10.1038/sj.ejhg.5201101.

282. Kotani M, Tanaka I, Ogawa Y, et al. Structural organization of the human prostaglandin EP3 receptor subtype gene (PTGER3). *Genomics*. 1997;40(3):425-434. doi: 10.1006/geno.1996.4585.

283. Lesurtel M, Graf R, Aleil B, et al. Platelet-derived serotonin mediates liver regeneration. *Science*. 2006;312(5770):104-107. doi: 10.1126/science.1123842.

284. Kanaoka Y, Maekawa A, Penrose JF, Austen KF, Lam BK. Attenuated zymosan-induced peritoneal vascular permeability and IgE-dependent passive cutaneous anaphylaxis in mice lacking leukotriene C4 synthase. *J Biol Chem*. 2001;276(25):22608-22613. doi: 10.1074/jbc.M103562200.

285. Gearing DP, Druck T, Huebner K, et al. The leukemia inhibitory factor receptor (LIFR) gene is located within a cluster of cytokine receptor loci on mouse chromosome 15 and human chromosome 5p12-p13. *Genomics*. 1993;18(1):148-150. doi: 10.1006/geno.1993.1441.

286. Arita K, South AP, Hans-Filho G, et al. Oncostatin M receptor-beta mutations underlie familial primary localized cutaneous amyloidosis. *Am J Hum Genet*. 2008;82(1):73-80. doi: 10.1016/j.ajhg.2007.09.002; 10.1016/j.ajhg.2007.09.002.

287. Song HY, Jeon ES, Jung JS, Kim JH. Oncostatin M induces proliferation of human adipose tissue-derived mesenchymal stem cells. *Int J Biochem Cell Biol*. 2005;37(11):2357-2365. doi: 10.1016/j.biocel.2005.05.007.

288. Sia CL, Traniello S, Pontremoli S, Horecker BL. Studies on the subunit structure of rabbit liver fructose diphosphatase. *Arch Biochem Biophys*. 1969;132(1):325-330.

289. Chesney J, Mitchell R, Benigni F, et al. An inducible gene product for 6-phosphofructo-2-kinase with an AU-rich instability element: Role in tumor cell glycolysis and the warburg effect. *Proc Natl Acad Sci U S A*. 1999;96(6):3047-3052.

290. Yi W, Clark PM, Mason DE, et al. Phosphofructokinase 1 glycosylation regulates cell growth and metabolism. *Science*. 2012;337(6097):975-980. doi: 10.1126/science.1222278; 10.1126/science.1222278.

291. Passino MA, Adams RA, Sikorski SL, Akassoglou K. Regulation of hepatic stellate cell differentiation by the neurotrophin receptor p75NTR. *Science*. 2007;315(5820):1853-1856. doi: 10.1126/science.1137603.

292. Hao K, Luk JM, Lee NP, et al. Predicting prognosis in hepatocellular carcinoma after curative surgery with common clinicopathologic parameters. *BMC Cancer*. 2009;9:389-2407-9-389. doi: 10.1186/1471-2407-9-389; 10.1186/1471-2407-9-389.

293. Tsai WC, Hsu PW, Lai TC, et al. MicroRNA-122, a tumor suppressor microRNA that regulates intrahepatic metastasis of hepatocellular carcinoma. *Hepatology*. 2009;49(5):1571-1582. doi: 10.1002/hep.22806; 10.1002/hep.22806.

294. Coulouarn C, Factor VM, Andersen JB, Durkin ME, Thorgeirsson SS. Loss of miR-122 expression in liver cancer correlates with suppression of the hepatic phenotype and gain of metastatic properties. *Oncogene*. 2009;28(40):3526-3536. doi: 10.1038/onc.2009.211; 10.1038/onc.2009.211.

295. Brandon M, Baldi P, Wallace DC. Mitochondrial mutations in cancer. *Oncogene*. 2006;25(34):4647-4662. doi: 10.1038/sj.onc.1209607.

296. Liu AM, Yao TJ, Wang W, et al. Circulating miR-15b and miR-130b in serum as potential markers for detecting hepatocellular carcinoma: A retrospective cohort study. *BMJ Open*. 2012;2(2):e000825-2012-000825. Print 2012. doi: 10.1136/bmjopen-2012-000825; 10.1136/bmjopen-2012-000825.

297. Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. *Nat Rev Cancer*. 2004;4(3):177-183. doi: 10.1038/nrc1299.

298. Huret JL, Minor SL, Dorkeld F, Dessen P, Bernheim A. Atlas of genetics and cytogenetics in oncology and haematology, an interactive database. *Nucleic Acids Res*. 2000;28(1):349-351.

299. Waldman Lab U. Http://Waldman.ucsf.edu/GENES/completechroms.html.

300. Higgins ME, Claremont M, Major JE, Sander C, Lash AE. CancerGenes: A gene selection resource for cancer genome projects. *Nucleic Acids Res*. 2007;35(Database issue):D721-6. doi: 10.1093/nar/gkl811.

301. D'Antonio M, Pendino V, Sinha S, Ciccarelli FD. Network of cancer genes (NCG 3.0): Integration and analysis of genetic and network properties of cancer genes. *Nucleic Acids Res*. 2012;40(Database issue):D978-83. doi: 10.1093/nar/gkr952; 10.1093/nar/gkr952.

Appendix A: Supplementry Code

A.1 Chapter 3 Workflow

```
1.KEGG PARSER
2.
3.reaction.list <- c()
4.relation.list <- c()
5.entry.list<- c()
6.attrs.list<- c()
7.
8.# manually download kgml file(s) from ftp://ftp.genome.jp/pub/kegg/xml/kgml/ to local directory
9.file.vector<- c(#ENTER FILE NAMEs#)
10.
11.    #loop through kgml files
12.    for (j in file.vector) {
13.        entID2Name <- c()
14.        #parse xml and store as list
15.        TreeMet <-xmlParse(j)
16.        KGMLList <- xmlToList(TreeMet)
17.            #loop through xml elements
18.            for (i in 1:length(KGMLList)) {
19.            group.list <- c("(compound)")
20.            nodeType <- names(KGMLList[i])
21.                # store pathway attributes
22.                if (nodeType == ".attrs") {
23.                    attrs.list <- as.list(c(attrs.list, KGMLList[i]$.attrs[["name"]], KGMLList[i]$.attrs[["org"]]
24.                    , KGMLList[i]$.attrs[["number"]], KGMLList[i]$.attrs[["title"]], KGMLList[i]$.attrs[["image"]],
25.                    KGMLList[i]$.attrs[["link"]]
26.                    ))
27.                    print(KGMLList[i]$.attrs[["title"]])
28.                }
29.                #parse entry type including reaction attributes
30.                if (nodeType == "entry" && length(KGMLList[i]$entry$.attrs) == 5) {
31.                    entry.list <- as.list(c(entry.list, KGMLList$.attrs[["name"]], KGMLList[i]$entry$graphics[["name"]],
32.                    KGMLList[i]$entry$.attrs[["id"]],
33.                    KGMLList[i]$entry$.attrs[["name"]],KGMLList[i]$entry$.attrs[["type"]],
34.                    KGMLList[i]$entry$.attrs[["link"]], KGMLList[i]$entry$.attrs[["reaction"]]
35.                    ))
36.                    entID2Name[KGMLList[i]$entry$.attrs[["id"]]] <-
       strsplit(gsub('  ', '',KGMLList[i]$entry$graphics[["name"]]), ",")
37.                }
```

```
38.             #parse entry type without reaction attributes
39.             if (nodeType == "entry" && length(KGMLList[i]$entry$.attrs) == 4) {
40.                 entry.list <- as.list(c(entry.list, KGMLList$.attrs[["name"]], KGMLList[i]$entry$graphics[["name"]],
41.                 KGMLList[i]$entry$.attrs[["id"]],
42.                 KGMLList[i]$entry$.attrs[["name"]],KGMLList[i]$entry$.attrs[["type"]],
43.                 KGMLList[i]$entry$.attrs[["link"]], "NA"
44.                 ))
45.                 entID2Name[KGMLList[i]$entry$.attrs[["id"]]] <-
    strsplit(gsub('  ', '',KGMLList[i]$entry$graphics[["name"]]), ",")
46.             }
47.             #parse entry type corresponding to group lists
48.             if (nodeType == "entry" && length(KGMLList[i]$entry$.attrs) == 3 && KGMLList[i]$entry$.attrs[["type"]] ==
49.                 "group") {
50.                     #retrieve elements in group list
51.                     for (z in 2:(length(KGMLList[i]$entry) - 1)) {
52.                         group.list <- c(group.list, " ", entID2Name[KGMLList[i]$entry[[z]]])
53.                     }
54.                 entID2Name[KGMLList[i]$entry$.attrs[["id"]]] <- group.list
55.                 entry.list <- as.list(c(entry.list, KGMLList$.attrs[["name"]], "NA",
56.                 KGMLList[i]$entry$.attrs[["id"]],
57.                 KGMLList[i]$entry$.attrs[["name"]],KGMLList[i]$entry$.attrs[["type"]],
58.                 "NA", "NA"
59.                 ))
60.             }
61.             #parse relation type
62.             if (nodeType == "relation") {
63.                 if (names(KGMLList[i]$relation[1]) == "text") {
64.                     subtype_name <- "na"
65.                     subtype_value <- "na"
66.                 }
67.                 else {
68.                     subtype_name <- KGMLList[i]$relation$subtype[["name"]]
69.                     subtype_value <- KGMLList[i]$relation$subtype[["value"]]
70.                 }
71.                 for (k in 1:length(entID2Name[KGMLList[i]$relation$.attrs[["entry1"]]][[1]])) {
72.                     for (l in 1: length(entID2Name[KGMLList[i]$relation$.attrs[["entry2"]]][[1]])) {
73.                         relation.list <- as.list(c(relation.list, KGMLList$.attrs[["name"]],
74.                         subtype_name,
75.                         subtype_value,
76.                         KGMLList[i]$relation$.attrs[["entry1"]],
77.                         KGMLList[i]$relation$.attrs[["entry2"]],
78.                         gsub("\\.", "", entID2Name[KGMLList[i]$relation$.attrs[["entry1"]]][[1]][k]),
79.                         gsub("\\.", "", entID2Name[KGMLList[i]$relation$.attrs[["entry2"]]][[1]][l]),
80.                         KGMLList[i]$relation$.attrs[["type"]]
```

```
                        ))
                    }
                }
            }
            #parse reaction type
            if (nodeType == "reaction" && length(KGMLList[i]$reaction) >= 3) {
                substrateList <- c()
                productList <- c()
                    for (s in 1:length(KGMLList[i]$reaction)) {
                        if (names(KGMLList[i]$reaction[s]) == "substrate") {
                            substrateList <- paste(substrateList, KGMLList[i]$reaction[s]$substrate[[2]], sep=" ")
                        }
                        if (names(KGMLList[i]$reaction[s]) == "product") {
                            productList <- paste(productList, KGMLList[i]$reaction[s]$product[[2]], sep=" ")
                        }
                    }
                reaction.list <- as.list(c(reaction.list, KGMLList$.attrs[["name"]],
                productList, substrateList, KGMLList[i]$reaction$.attrs[["name"]],
                KGMLList[i]$reaction$.attrs[["type"]]
                ))
            }
        }
    }

#create dataframes
entry.df <- as.data.frame(matrix(entry.list, ncol=7, byrow=TRUE))
relation.df <- as.data.frame(matrix(relation.list, ncol=8,byrow=TRUE))
reaction.df <- as.data.frame(matrix(reaction.list, ncol=5, byrow=TRUE)) #Only for metabolic pathways
attrs.dft <- as.data.frame(matrix(attrs.list, ncol=6, byrow=TRUE))
entry.names <- as.list(c("pathway.name", "graphics.name", "entry.id", "entry.name", "entry.type", "entry.link",
"entry.reaction"))
#add names to dataframes
reaction.names <- as.list(c("pathway.name", "product", "substrate", "reaction.name", "reaction.type"))
relation.names <- as.list(c("pathway.name", "subtype.name", "subtype.value", "entry1", "entry2", "entry1_name",
"entry2_name", "relation.type"))
attrs.names <- as.list(c("name", "org", "number", "title", "image", "link"))
names(reaction.df) <- reaction.names
names(relation.df)<- relation.names
names(entry.df) <- entry.names
names(attrs.df) <- attrs.names
#extract list of relation pairs for graph analysis
relationPairs <- as.data.frame(relation.df[,c(6,7)])


```

```
125.   LOGISTIC REGRESSION SAMPLE TEST
126.
127.   lrMetDeg15 <- lrm(metResults$Cancer ~ metResults$Degree15)
128.
129.   vv <- diag(lrMetDeg15$var)
130.   cof <- lrMetDeg15$coef
131.   secof <- sqrt(vv)
132.   z <- cof/sqrt(vv)
133.   pv <- 1 - pchisq(z^2, 1)
134.   ap <- anova(lrMetDeg15)["metResults", "P"]
135.
136.   lrResults[which(lrResults[,1] == "lrMetDeg15"),2] <- pv[[2]]
137.   lrResults[which(lrResults[,1] == "lrMetDeg15"),3] <- ap
138.   lrResults[which(lrResults[,1] == "lrMetDeg15"),4] <- cof[[2]]
139.   lrResults[which(lrResults[,1] == "lrMetDeg15"),5] <- secof[[2]]
140.   lrResults[which(lrResults[,1] == "lrMetDeg15"),6] <- exp(cof[[2]])
141.   lrResults[which(lrResults[,1] == "lrMetDeg15"),7] <- exp(secof[[2]])
142.   lrResults[which(lrResults[,1] == "lrMetDeg15"),8] <- z[[2]]
143.
144.   GLOBAL NETWORK PROPERTIES
145.   metICC <- transitivity(metNet, type="local")
146.   metBetween <- betweenness(metNet)
147.   metDegree <- degree(metNet)
148.   metClose <- closeness(metNet)
```

A.2 Chapter 4 Workflow

```
INSTALL AND LOAD IGRAPH PACKAGE
0.install.packages("igraph", lib="/my/own/R-packages/")
1.library("igraph", lib.loc="/my/own/R-packages/")
2.
3.PARSE GENE EXPRESSION DATA
4.gse14520 <- getGEO("GSE14520")
5.show(gse14520)
6.
7.EXPONENTIATE DATA
8.exp14520 <- exprs(gse14520[[1]])
9.
10.    PARSE PHENOTYPE DATA
11.    pheno14520.df <- pData(phenoData(gse14520[[1]]))
12.    TumorStatus14520 <- c()
13.    NonTumor14520 <- grep("Liver Non-Tumor", pheno14520.df$characteristics_ch1)
14.    Tumor14520 <- grep("Liver Tumor", pheno14520.df$characteristics_ch1)
15.    TumorStatus14520[c(Tumor14520)] <- 1
16.    TumorStatus14520[c(NonTumor14520)] <- 0
17.    design14520 = model.matrix(~ -1+factor(c(TumorStatus14520)))
18.    colnames(design14520) = c("Normal", "Tumor")
19.    contrast.matrix14520 <- makeContrasts(Tumor-Normal, levels=design14520)
20.
21.    CALCULATE P-VALUES AND FOLD CHANGE
22.    fit14520 <- lmFit(exp14520, design14520)
23.    fit.contrast.14520 <- contrasts.fit(fit14520, contrast.matrix14520)
24.    fit.ebayes.14520 <- eBayes(fit.contrast.14520)
25.    names(fit.contrast.14520)
26.    names(fit.ebayes.14520)
27.    top14520.pval.1 <- topTable(fit.ebayes.14520, n=Inf, p.value=.1, sort.by="logFC", adjust.method="BH")
28.
29.    MAP HUGO IDS TO GENE SYMBOLS
30.    pval1.14520.IDs <- top14520.pval.1[,1]
31.
32.    x <- hgu133plus2SYMBOL
33.    mapped_probes <- mappedkeys(x)
34.    xx <- as.list(x[mapped_probes])
35.    if(length(xx) > 0) {
36.        # Get the SYMBOL for the first five probes
37.        xx[1:5]
38.        # Get the first one
```

```
        xx[[1]]
    }

    count = 0
    for (i in pval1.14520.IDs) {
        count = count + 1
        if (length(xx[[i]]) > 0) {
            top14520.pval.1[count, 8] <- xx[[i]]
        }
    }

    BUILD INTERACTION NETWORK
    globalNet <- graph.data.frame(allKeggHPRD, directed=FALSE)]
    globalNet.bak <- globalNet
    globalNet <- set.edge.attribute(globalNet, "source", index=E(globalNet), labelEdges)
    sGlobalNet <- simplify(globalNet)
    sGlobalNet <- delete.vertices(sGlobalNet, V(sGlobalNet)[ degree(sGlobalNet)==0 ])
    summary(globalNet)
    summary(sGlobalNet)
    globalClusters <- clusters(sGlobalNet)

    FIND CONNECTED CLUSTERS
    globalClusters$csize
    cluster1 <- which(globalClusters$membership == 0)
    length(cluster1)

    EXTRACT CONNECTED CLUSTER
    cGlobalNet <- subgraph(sGlobalNet, cluster1 - 1)
    summary(cGlobalNet)

    SAVE EDGELIST AND VERTEX LIST
    globalEdges <- get.edgelist(cGlobalNet, names=TRUE)
    clusterNetVertices <- cGlobalNet[9][[1]][[3]][[1]]
    globalNetVertices <- globalNet[9][[1]][[3]][[1]]

    CREATE VECTOR OF EDGE WEIGHTS
    ExpWeights14520.p1 <- c()
    for (i in 1:length(globalEdges[,1])) {
        tempEdge1 <- globalEdges[i,1]
        tempEdge2 <-  globalEdges[i,2]
        if ((length(which(top14520.pval.1[,8] == tempEdge1)) >0)  || (length(which(top14520.pval.1[,8] == tempEdge2)) >0)) {
            if ((length(which(top14520.pval.1[,8] == tempEdge1)) >0)  && (length(which(top14520.pval.1[,8] == tempEdge2)) >0))
     {
```

```
                ExpWeights14520.p1[i] <-
 (max(abs(top14520.pval.1[which(top14520.pval.1[,8] == tempEdge1),][,2])) + max(abs(top14520.pval.1[which(top14520.pval.1[
,8] == tempEdge2),][,2])))/2}
        else if  ((length(which(top14520.pval.1[,8] == tempEdge1)) >0)) {
            ExpWeights14520.p1[i] <- (max(abs(top14520.pval.1[which(top14520.pval.1[,8] == tempEdge1),][,2])))/2}
        else if ((length(which(top14520.pval.1[,8] == tempEdge2)) >0)) {
            ExpWeights14520.p1[i] <- (max(abs(top14520.pval.1[which(top14520.pval.1[,8] == tempEdge2),][,2])))/2}
        else { ExpWeights14520.p1[count] <- .01}
    }
}

CREATE VECTOR OF VERTEX WEIGHTS
ExpVWeights14520.p1 <- c()
count = 0
for (i in clusterNetVertices) {
    count = count +1
    if (length(which(top14520.pval.1[,8] == i)) >0)  {
        ExpVWeights14520.p1[count] <- max(abs(top14520.pval.1[which(top14520.pval.1[,8] == i),][,2]))}
    else {ExpVWeights14520.p1[count] <- .01}
}

wtc14520.p1 <-
 walktrap.community(cGlobalNet, steps = 3, merges=TRUE, modularity = TRUE, labels = TRUE, membership = TRUE, weights = Exp
Weights14520.p1)

CREATE BOOTSTRAP DISTRIBUTION FOR CLUSTER SCORES
draws14520.p1 <- matrix (ncol = 5000, nrow = 200)
for (i in 3:200) {
    draws <- matrix(sample(ExpVWeights14520.p1, size = i * 5000, replace = TRUE), i)
        drawmeans <- apply(draws, 2, mean)
        draws14520.p1[i,] <- drawmeans
}

CANCER LIST AND CANCER VERTEX WEIGHTS
cancerVertexWeights <- c()
count = 0
for (i in clusterNetVertices) {
    count = count +1
    if (length(which(cancerList == i)) >0)  {
        cancerVertexWeights[count] <- (cancerList[which(cancerList == i),2])
    }
    else{}
}

```

```
drawsCancer <- matrix (ncol = 5000, nrow = 1000)
for (i in 3:1000) {
    draws <- matrix(sample(cancerVertexWeights, size = i * 5000, replace = TRUE), i)
    drawmeans <- apply(draws, 2, mean)
    drawsCancer[i,] <- drawmeans
}

MODULARITY CHECK
which.max(wtc14520.p1$modularity)-1

CHECK FOR BEST MODEL USING MAX SIZE <=200
stop = 0
comm.scores14520.p1 <- c(0)
step.size <- round(.20 * (length(wtc14520.p1$labels)))
increment <- round(.005 * (length(wtc14520.p1$labels)))

while (step.size <= length(wtc14520.p1$labels)) {
    comm.steps <- c(0)
    comm.memb <-
  community.to.membership(cGlobalNet, wtc14520.p1$merges, steps=step.size,   membership=TRUE, csize=TRUE)
    community.vector <-  which(comm.memb$csize > 3) -1
    if(max(comm.memb$csize) <= 200) {
        all.comm.means <- c(0)
        for (i in community.vector) {
            comm.size <- comm.memb$csize[i +1]
            comm.total.mean <- mean(ExpVWeights14520.p1[which(comm.memb$memb == i)])
            comm.total.zscore <- abs(comm.total.mean -
                                                    mean(draws14520.p1[comm.size,]))/sqrt(var(draws14520.p1[comm.size,]))

            all.comm.means <- c(all.comm.means, comm.total.zscore)
        }
    }
else {}

comm.scores14520.p1[step.size] <- max(all.comm.means)
step.size <- step.size + increment

}

CHECK FOR OPTIMAL STEP SIZE
which(comm.scores == max(comm.scores, na.rm=TRUE))

RUN CLUSTERING WITH OPTIMAL STEP SIZE

```

```
  rw14520.sigvalues <- matrix(ncol=4)
   comm.memb14520 <-  community.to.membership(cGlobalNet, wtc14520.p1$merges, steps=2393, membership=TRUE, csize=TRUE)
   community.vector <-  which(comm.memb14520$csize >= 3) -1
   if(max(comm.memb14520$csize) <= 200) {
      all.comm.means <- c(0)
       for (i in community.vector) {
           cancerScores <- cancerVertexWeights[which(comm.memb14520$memb == i)]
         cmean <- mean(cancerScores)
           comm.size <- comm.memb14520$csize[i +1]
           comm.total.mean <- mean(ExpVWeights14520.p1[which(comm.memb14520$memb == i)])
           comm.total.zscore <- (comm.total.mean -
                                                 mean(draws14520.p1[comm.size,]))/sqrt(var(draws14520.p1[comm.size,]))
         comm.total.cscore <- (cmean -
                                             mean(drawsCancer[length(cancerScores),]))/sqrt(var(drawsCancer[length(cancerScores
    ),]))
           rw14520.sigvalues <-
    rbind(rw14520.sigvalues, c(i, comm.total.zscore, comm.size,                          comm.total.cscore))
         print(comm.total.zscore)
                    all.comm.means <- c(all.comm.means, comm.total.zscore)
       }
 }
```

A.3 Chapter 5 Workflow I

```
   1. LOAD BIOCONDUCTOR GEOquery and limma packages
2.source("http://www.bioconductor.org/biocLite.R"
3.biocLite("GEOquery")
4.biocLite("limma")
5.
6.INSTALL AND LOAD IGRAPH, OPTMATCH PACKAGE
7.install.packages("igraph")
8.library("igraph")
9.install.packages("optmatch")
10.    library("optmatch")
11.
12.    gse22058 <- getGEO("GSE22058")
13.    show(gse22058)
14.    exp22058 <- exprs(gse22058[[1]])
15.    exp22058.3 <- exprs(gse22058[[3]])
16.    exp22058.2 <- exprs(gse22058[[2]])
17.    newmat <- apply(exprs(gse22058[[1]]), 2, as.numeric)
18.    exp22058.bak <- exp22058
19.    exp22058 <- newmat
20.    rownames(exp22058) <- rownames(exp22058.bak)
21.    pheno22058.df <- pData(phenoData(gse22058[[1]]))
22.    colnames(pheno22058.df)
23.    HCTumor <- c()
24.    nonTumor <- grep("adjacent", pheno22058.df[,11])
25.    tumor <- grep("liver tumor", pheno22058.df[,11])
26.    HCTumor[c(nonTumor)] <- 0
27.    HCTumor[c(tumor)] <- 1
28.    design22058 = model.matrix(~ -1+factor(c(HCTumor)))
29.    colnames(design22058) = c("nonTumor", "tumor")
30.    contrast.matrix22058 <- makeContrasts(tumor-nonTumor, levels=design22058)
31.
32.    fit22058 <- lmFit(exp22058, design22058)
33.    fit.contrast.22058 <- contrasts.fit(fit22058, contrast.matrix22058)
34.    fit.ebayes.22058 <- eBayes(fit.contrast.22058)
35.    top22058.all <- topTable(fit.ebayes.22058, n=Inf, sort.by="logFC", adjust.method="BH")
36.    top22058.pval.1 <- topTable(fit.ebayes.22058, n=Inf, p.value=.1, sort.by="logFC", adjust.method="BH")
37.    top22058.pval.05 <- topTable(fit.ebayes.22058, n=Inf, p.value=.05, sort.by="logFC", adjust.method="BH")
38.    top22058.pval.05 <- topTable(fit.ebayes.22058, n=Inf, p.value=.05, sort.by="logFC", adjust.method="BH")
39.
40.    miRNADS1.FCall <- top22058.all[,c(7,2)]
```

```
41.    names(miRNADS1.FCall) <- names(miRNAFoldChange.p1[1:2])
42.    miRNA.RNADS1.FCall <- rbind(miRNAFoldChange.p1[,1:2], miRNADS1.FCall)
43.
44.    #CREATE WEIGHT MATRICES
45.
46.    miRNAbfc <- bigFoldChange[,1]
47.    miRNAmatchesP1FC2DS1 <- miRNAtable[miRNAtable$target %in% miRNAbfc,]
48.    miRNAmatchesSort <- miRNAmatchesP1FC2DS1[with(miRNAmatchesP1FC2DS1, order(miRNA, corr)),]
49.
50.    for (i in unique(miRNAmatchesSort[,1])) {
51.        newvec <- which(miRNAmatchesSort[,1] == i)
52.        if (length(newvec) > 5) {
53.            newvec <- newvec[1:5] }
54.        best5DS1mat <- rbind(best5DS1mat, miRNAmatchesSort[newvec,])
55.    }
56.
57.    best5DS1mat <- matrix(ncol=3)
58.    colnames(best5DS1mat) <- colnames(miRNAmatchesSort)
59.
60.    #fbest5DS1mat filtered matches are parsed separately from best5DS1mat, perl code attached.
61.
62.    FDS1B5Names <- c(t(fbest5DS1mat[1:194, 1:2]))
63.    FDS1B5UNames <- unique FDS1B5Names[!(FDS1B5Names %in% clusterNetVNames)])
64.    fDS1B5Targets <- fbest5DS1mat[2:194, 2]
65.
66.    ADD MIRNA VERTICES AND EDGES
67.    fmiRNAB5DS1.net <- add.vertices(clusterGlobalNet.labels.test, c(115), name= FDS1B5UNames)
68.    fclusterNetDS1B5Names <- V(fmiRNAB5DS1.net)$name
69.
70.    FDS1B5EdgeNames <- c()
71.        count <- 0
72.        for (i in  FDS1B5Names) {
73.            count = count +1
74.        FDS1B5EdgeNames[count] <- c(which(fclusterNetDS1B5Names  == i) -1)
75.    }
76.
77.    fmiRNAB5DS1.net <- add.edges(fmiRNAB5DS1.net, FDS1B5EdgeNames, source="fmiRNAB5")
78.     (fmiRNAallDS2.net, FDS2allEdgeNames, source="fmiRNAallMatch")
79.
80.    EDGE AND VERTEX LIST
81.    fmiRNAB5DS1Vertices <- fmiRNAB5DS1.net[9][[1]][[3]][[1]]
82.    fglobalEdges.miRNAB5DS1 <- get.edgelist(fmiRNAB5DS1.net, names=TRUE)
83.
84.    CALCULATE VERTEX WEIGHTS
```

```
fmiRNA.RNADS1B5.Weights.p1 <- c()
count = 0
for (i in fmiRNAB5DS1Vertices) {
    count = count +1
    if (length(which(miRNA.RNADS1.FCall[,1] == i)) >0)  {
        fmiRNA.RNADS1B5.Weights.p1[count] <-
                                  max(abs(miRNA.RNADS1.FCall[which(miRNA.RNADS1.FCall[,1] == i),][,2]))}
    else {fmiRNA.RNADS1B5.Weights.p1[count] <- .01}
    }
}

BUILD BOOTSTRAP DISTRIBUTION FOR MODULE SCORES
fdrawsB5.miRNA.RNADS1.p1 <- matrix (ncol = 5000, nrow = 200)
for (i in 3:200) {
    drawsDS1 <- matrix(sample(fmiRNA.RNADS1B5.Weights.p1, size = i * 5000, replace = TRUE), i)
    drawmeansDS1 <- apply(drawsDS1, 2, mean)
    fdrawsB5.miRNA.RNADS1.p1[i,] <- drawmeansDS1
}

CALCULATE EDGE WEIGHTS
fExpWeightsmiRNAB5DS1.p1 <- c()
for (i in 1:length(fglobalEdges.miRNAB5DS1[,1])) {
    tempEdge1 <- fglobalEdges.miRNAB5DS1[i,1]
    tempEdge2 <-  fglobalEdges.miRNAB5DS1[i,2]
    if ((length(which(miRNA.RNADS1.FCall[,1] == tempEdge1)) >0)  ||                 (length(which(miRNA.RNADS1.FCall[,1]
== tempEdge2)) >0)) {
    if ((length(which(miRNA.RNADS1.FCall[,1] == tempEdge1)) >0)  &&      (length(which(miRNA.RNADS1.FCall[,1] == tempEdge2
)) >0)) {
        fExpWeightsmiRNAB5DS1.p1[i] <-
                                  (max(abs(miRNA.RNADS1.FCall[which(miRNA.RNADS1.FCall[,1] == tempEdge1),][,2]))
 + max(abs(miRNA.RNADS1.FCall[which(miRNA.RNADS1.FCall[,1] ==                    tempEdge2),][,2])))/2}
    else if  ((length(which(miRNA.RNADS1.FCall[,1] == tempEdge1)) >0)) {
        fExpWeightsmiRNAB5DS1.p1[i] <-
   (max(abs(miRNA.RNADS1.FCall[which(miRNA.RNADS1.FCall[,1] == tempEdge1),][,2])))/2}
    else if ((length(which(miRNA.RNADS1.FCall[,1] == tempEdge2)) >0)) {
        fExpWeightsmiRNAB5DS1.p1[i] <-
   (max(abs(miRNA.RNADS1.FCall[which(miRNA.RNADS1.FCall[,1] == tempEdge2),][,2])))/2}
    else { fExpWeightsmiRNAB5DS1.p1[i] <- .01}
 }
}

BUILD WALKTRAP COMMUNITY
```

```
fwtcmiRNAB5DS1 <-
 walktrap.community(fmiRNAB5DS1.net, steps = 3, merges=TRUE, modularity = TRUE, labels = TRUE, membership = TRUE, weights
= fExpWeightsmiRNAB5DS1.p1)

BUILD BOOTSTRAP DISTRIBUTION FOR CANCER GENES
drawsCancer <- matrix (ncol = 5000, nrow = 1000)
for (i in 3:1000) {
    draws <- matrix(sample(cancerVertexWeights, size = i * 5000, replace = TRUE), i)
    drawmeans <- apply(draws, 2, mean)
    drawsCancer[i,] <- drawmeans
}

CALCULATE VERTEX WEIGHTS FOR MIRNA ENRICHMENT
fmiRNAWeightsEAB5.DS1 <- c()
count = 0
for (i in clusterNetVertices) {
    count = count +1
    if (length(which(fDS1B5Targets == i)) >0)  {
        fmiRNAWeightsEAB5.DS1[count] <- 1}
    else {fmiRNAWeightsEAB5.DS1[count] <- 0}
}

BUILD BOOTSTRAP DISTRIBUTION FOR MODULE MIRNA ENRICHMENT SCORES
fdrawsmiRNAEAB5DS1.p1 <- matrix (ncol = 5000, nrow = 200)
for (i in 3:200) {
    drawsDS1 <- matrix(sample(fmiRNAWeightsEAB5.DS1, size = i * 5000, replace = TRUE), i)
    drawmeansDS1 <- apply(drawsDS1, 2, mean)
    fdrawsmiRNAEAB5DS1.p1[i,] <- drawmeansDS1
}

CHECK MODULARITY
which.max(fwtcmiRNAB5DS1$modularity)-1, membership=TRUE, csize=TRUE)

FIND BEST STEP SIZE
stop = 0
comm.scoresmiRNAB5DS1 <- c(0)
step.size <- round(.2 * (length(wtcmiRNAB5DS1$labels)))
increment <- round(.005 * (length(wtcmiRNAB5DS1$labels)))

while (step.size <= length(wtcmiRNAB5DS1$labels)) {
    comm.steps <- c(0)
    comm.memb <-
  community.to.membership(fmiRNAB5DS1.net, fwtcmiRNAB5DS1$merges,    steps=step.size, membership=TRUE, csize=TRUE)
    community.vector <-  which(comm.memb$csize > 3) -1
```

```
161.      all.comm.means <- c(0)
162.      if(max(comm.memb$csize) <= 200) {
163.          for (i in community.vector) {
164.              comm.size <- comm.memb$csize[i +1]
165.              comm.total.mean <-
                                              mean((fmiRNA.RNADS1B5.Weights.p1[which(comm.memb$memb == i)]^2))
166.              comm.total.zscore <- abs(comm.total.mean -
                                 mean((fdrawsB5.miRNA.RNADS1.p1[comm.memb$csize[i                        +1],])^2))/sqrt(var
      ((fdrawsB5.miRNA.RNADS1.p1[comm.memb$csize[i +1],])^2))
167.              all.comm.means <- c(all.comm.means, comm.total.zscore)
168.          }
169.      }
170.      else {}
171.
172.  comm.scoresmiRNAB5DS1[step.size] <- max(all.comm.means)
173.  step.size <- step.size + increment
174.  print(step.size)
175.  }
176.
177.  which(comm.scoresmiRNAB5DS1 == max(comm.scoresmiRNAB5DS1, na.rm=TRUE))
178.
179.  RUN MODEL AND ARCHIVE SCORES
180.  comm.miRNAB5DS1.2776 <-
       community.to.membership(fmiRNAB5DS1.net,  fwtcmiRNAB5DS1$merges, steps=2776, membership=TRUE, csize=TRUE)
181.
182.  miRNADS1.2776B5.sigvalues <-  matrix(ncol=5)
183.   community.vector <-  which(comm.miRNAB5DS1.2776$csize > 3) -1
184.   if(max(comm.miRNAB5DS1.2776$csize) <= 200) {
185.       all.comm.means <- c(0)
186.       for (i in community.vector) {
187.          cancerScores <- cancerVertexWeights[which(comm.miRNAB5DS1.2776$memb == i)]
188.          cmean <- mean(cancerScores)
189.          comm.total.cscore <- (cmean -
                                                    mean(drawsCancer[length(cancerScores),]))/sqrt(var(drawsCancer[length(can
      cerScores),]))
190.           comm.size <-comm.miRNAB5DS1.2776$csize[i +1]
191.          comm.total.mean <-
                                              mean((fmiRNA.RNADS1B5.Weights.p1[which(comm.miRNAB5DS1.2776$memb ==
       i)]^2))
192.          comm.total.zscore <- abs(comm.total.mean -
                                 mean((fdrawsB5.miRNA.RNADS1.p1[comm.miRNAB5DS1.2776$csize[i                    +1],])^2))/
      sqrt(var((fdrawsB5.miRNA.RNADS1.p1
193.          [comm.miRNAB5DS1.2776$csize[i +1],])^2))
```

```
194.          ecomm.total.mean <-
                                                mean((fmiRNAWeightsEAB5.DS1[which(comm.miRNAB5DS1.2776$memb == i)]^2))
195.          comm.total.eascore <- abs(ecomm.total.mean -
                                      mean((fdrawsmiRNAEAB5DS1.p1[comm.miRNAB5DS1.2776$csize[i                    +1],])^2))/sq
      rt(var((fdrawsmiRNAEAB5DS1.p1[comm.miRNAB5DS1.2776$csize[i             +1],])^2))
196.          miRNADS1.2776B5.sigvalues <-
       rbind(miRNADS1.2776B5.sigvalues, c(i,              comm.total.zscore, comm.miRNAB5DS1.2776$csize[i +1], comm.total.cscor
      e,                 comm.total.eascore))
197.          print(comm.total.zscore)
198.          all.comm.means <- c(all.comm.means, comm.total.zscore)
199.      }
200.  }
201.
202.  CODE FOR OPTIMAL MATCHING
203.  > miRNAtable <- read.csv("C:\\Users\\dpetroch\\downloads\\1_corrs_0.5.csv", header=TRUE, sep=",")
204.
205.  > dim(miRNAtable)
206.  [1] 40809     3
207.
208.  miRNAtable <- read.csv("C:\\Users\\dpetroch\\downloads\\1_corrs_0.5.csv", header=FALSE, sep=",")
209.  miRNAtableNames <- read.csv("C:\\Users\\dpetroch\\downloads\\1_corrs_0.5Unique.csv", header=FALSE, sep=",")
210.  colnames(miRNAtable) <- c("miRNA", "target", "corr")
211.  miRNAs <- levels(miRNAs)
212.  targets <- levels(targets)
213.  miRNAs <- miRNAs[2:156]))
214.
215.  local({pkg <- select.list(sort(.packages(all.available = TRUE)),graphics=TRUE)
216.  if(nchar(pkg)) library(pkg, character.only=TRUE)})
217.
218.  bigFoldChange <- read.csv("C:\\Users\\dpetroch\\downloads\\Burchard HCC fold abv 2.csv", header = TRUE, sep = ",")
219.  bigFoldColumns <- bigFoldChange[which(levels(bigFoldChange[,1]) %in% colnames(miRNAmatrix)),]
220.
221.  dim(miRNAmatrix[,which(colnames(miRNAmatrix) %in% bigFoldChange[,1])])
222.  miRNAmatrixFilt <- miRNAmatrix[,which(colnames(miRNAmatrix) %in% bigFoldChange[,1])]
223.
224.  miRNAmatrix <- matrix(0, nrow=155, ncol=5057, dimnames = list(levels(miRNAtable[,1]), levels(miRNAtable[,2])))
225.  miRNAmatrix[1:155,1:5057] <-2
226.  miRNAmatrix[cbind(miRNAtable$miRNA, miRNAtable$target)] <- (2- abs(miRNAtable$corr))
227.  miRNAmatrixFilt <- miRNAmatrix[,which(colnames(miRNAmatrix) %in% bigFoldChange[,1])]
228.
229.  #output: list of genes and miRNAs and their match numbers
230.  optPairMatch <- pairmatch(distance = miRNAmatrixFilt, tol = .0001, remove.unmatchables = TRUE, controls = 1)
231.  optPairMatches <- optPairMatch[where=(matched(optPairMatch))]
232.  optPairs <- matrix(ncol=2)
```

```
233.   for (i in c(67:221)) {
optPairs <-
 rbind(optPairs,c(names(optPairMatches[which(optPairMatches == optPairMatches[i])])[1], names(optPairMatches[which(optPairMatches
 == optPairMatches[i])])[2]))}
```

A.4: Chapter 5 Workflow II

Finding Queries for miRNA family memebers

```
   1. open (OUTFAM, ">C:\\myDir\\FilteredFamilyInfo.txt") || warn "can not open file for writing: FilteredFamilyInfo.txt";
2.open (MFAM, ">C:\\myDir\\familyMatches.txt") || warn "can not open file for writing:familyMatches.txt";
3.open (NOFAM, ">C:\\myDir\\noMatches.txt") || warn "can not open file for writing:familyMatches.txt";;
4.
5.
6.
7.
8.open (FAMMATCH, "C:\\myDir\\miRNAFamilies.txt") || warn "can not open file for reading:Nonconserved_Family_Info.txt";
9.
10.
11.
12.    #read in $miRNA- $miRNA_family list as hash table
13.    my %mfams;
14.    my %mmaps;
15.    while($line = <FAMMATCH>){
16.    chomp $line;
17.    ($miRNA, $mfamily) = split(/\t/, $line);
18.    $mfamily =~ s/\s*//;
19.    $mfams{$mfamily} = $miRNA;
20.    %mmaps = reverse %mfams;
21.    }
22.
23.    close (FAMMATCH);
24.
25.    while ( ($k,$v) = each %mfams ) {
26.       print "$k => $v end\n";
27.    }
28.
29.    while ( ($keym,$valuem) = each %mmaps ) {
30.       print "$keym => $valuem end\n";
31.    }
32.
33.
34.    open (INFAM, "C:\\myDir\\parsedmiRNA2s.txt") || warn "can not open file for reading: miRNAFamilies.txt";
35.
36.    while($line2 = <INFAM>){
37.        chomp $line2;
38.        ($miRFamily, $GeneSymbol, $PCT) = split("\t", $line2);
39.        #print "$line2\n";
40.        if(exists($mfams{$miRFamily})) {
```

```
41.             if (! grep(/$GeneSymbol/, @{$famTargets{$miRFamily}})){
42.             print OUTFAM "$miRFamily\t$GeneSymbol\t$PCT\n";
43.             push @{$famTargets{$miRFamily}}, $GeneSymbol;}
44.             else {}
45.         }
46.         else{print NOFAM "$miRFamily\n"};
47.     }
48.
49.     foreach $mmiRNA (keys %mmaps) {
50.         $fam = $mmaps{$mmiRNA};
51.             $mmiRNA = lc($mmiRNA);
52.         print MFAM "$mmiRNA\t$fam\t@{$famTargets{$fam}}\n";
53.         $matNum =  @{$famTargets{$fam}};
54.         print  "fam $matNum\n";
55.     }
56.
57.     close (INFAM);
58.
59.     close (OUTFAM);
60.     close (MFAM);
```

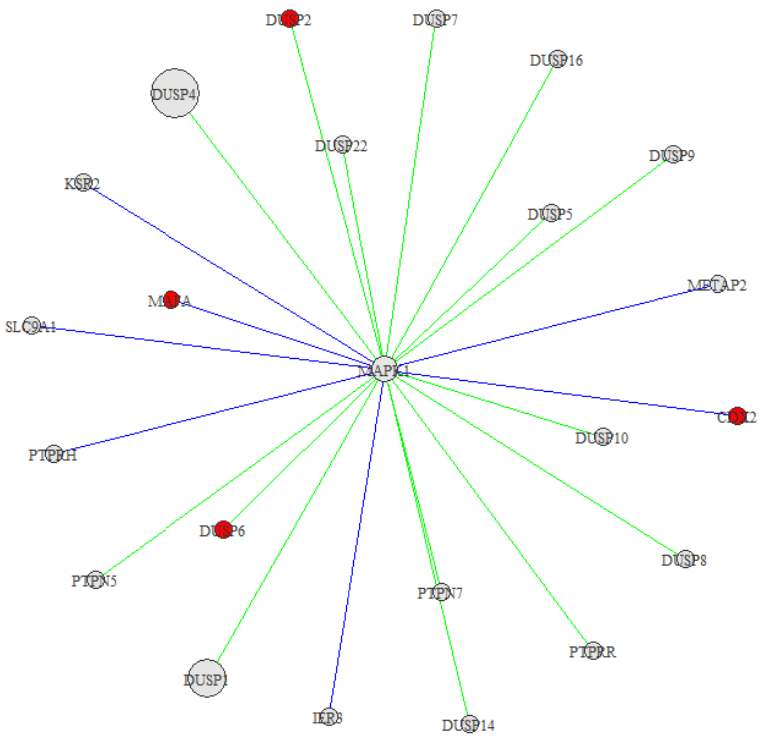
A.5: Chapter 5 Workflow III

```
   1. #matching to correlated results
2.open (CORRMIRNA, "C:\\myDir\\optPairsDS2.txt") || warn "can not open file for reading: optPairsDS2.txt";
3.open (INFAM, "C:\\myDir\\familyMatches.txt") || warn "can not open file for reading: miRNAFamilies.txt";
4.open (FILTMIRNA, ">C:\\myDir\\optPairsDS2.filtered") || warn "can not open file for writing: optPairsDS2.filtered";
5.open (EXTARGETS, ">C:\\myDir\\optPairsDS2.excluded") || warn "can not open file for writing: optPairsDS2.excluded";
6.
7.
8.#reformat DS2
9.#exclude unavailable targets
10.
11.    #read in file 2 as a hash table
12.    %famtargets;
13.    my $count = 0;
14.    while($line = <INFAM>){
15.    chomp $line;
16.    $count = ($count + 1);
17.    ($miRNA, $family, @targets) = split(/\t/, $line);
18.    $miRNA =~ s/\s*//g;
19.    $famtargets{$miRNA} = [ @targets ];
20.    #print "@{$famtargets{$miRNA}}";
21.    print "$miRNA.";
22.    }
23.
24.    close (INFAM);
25.
26.    my $count2 = 0;
27.    while($line2 = <CORRMIRNA>){
28.    chomp $line2;
29.    $count2 = $count2 + 1;
30.    $line2 =~ s/"//g;
31.    ($id, $miRNA2, $target2, $corr) = split(/\t/, $line2);
32.
33.    #debug statements
34.    #print "split $id, $miRNA2, $target2, $corr\n";
35.    #print "mirna family string $miRNA2 @{$famtargets{$miRNA2}}";
36.    #while ( ($k,$v) = each %famtargets ) {print "$k => @{$v} end\n";}
37.
38.    if (!defined @{ $famtargets{$miRNA2} }) {
39.        "$count2: $miRNA2 does  not exist in target list, continuing to next match.\n";
40.    }
41.    else {
```

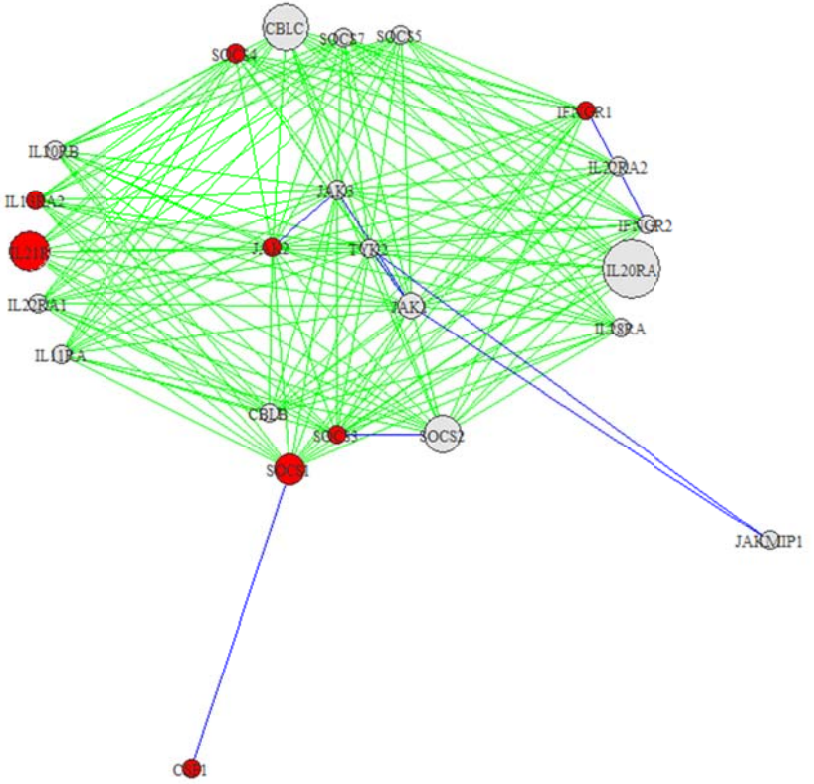
```
42.    #print "$count2: Processing $miRNA2 in target list...";
43.    }
44.
45.    if (grep /$target2/, @{ $famtargets{$miRNA2} }) {
46.    print FILTMIRNA "$miRNA2\t$target2\n";
47.    }
48.    else {print EXTARGETS "$miRNA2\t$target2\n";}
49.    }
50.
51.
52.    close (CORRMIRNA);
53.    close (FILTMIRNA);
close(EXTARGETS);
```

Appendix B: Supplemental Figures

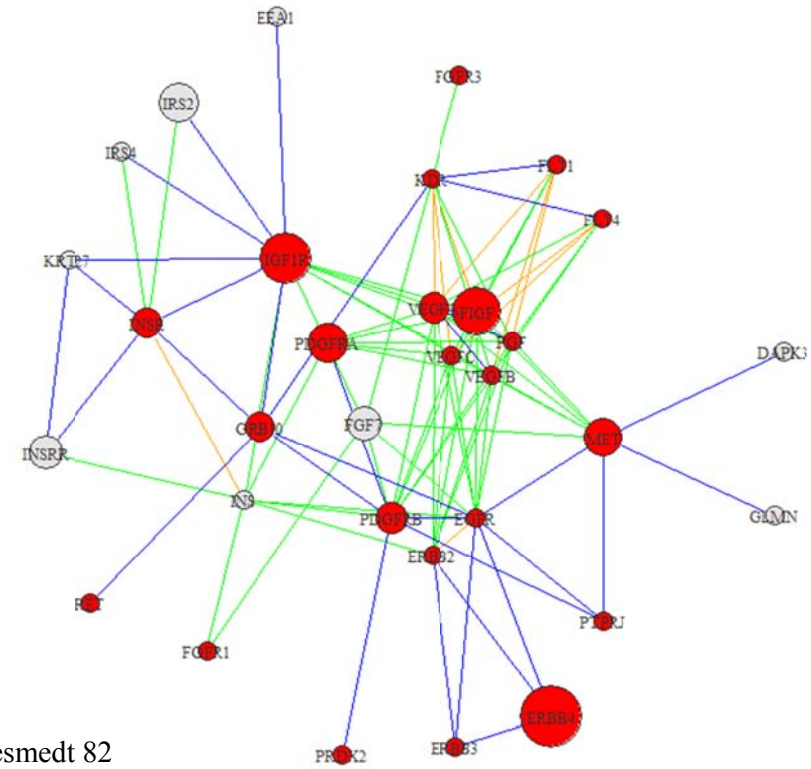
Significant Modules in Desmedt 2007 Data



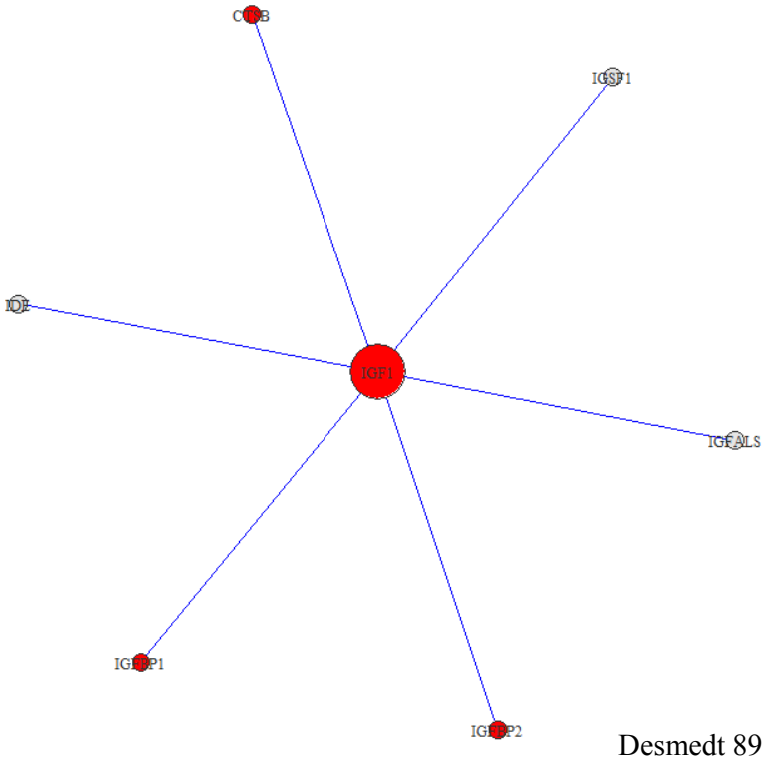
Desmedt 12



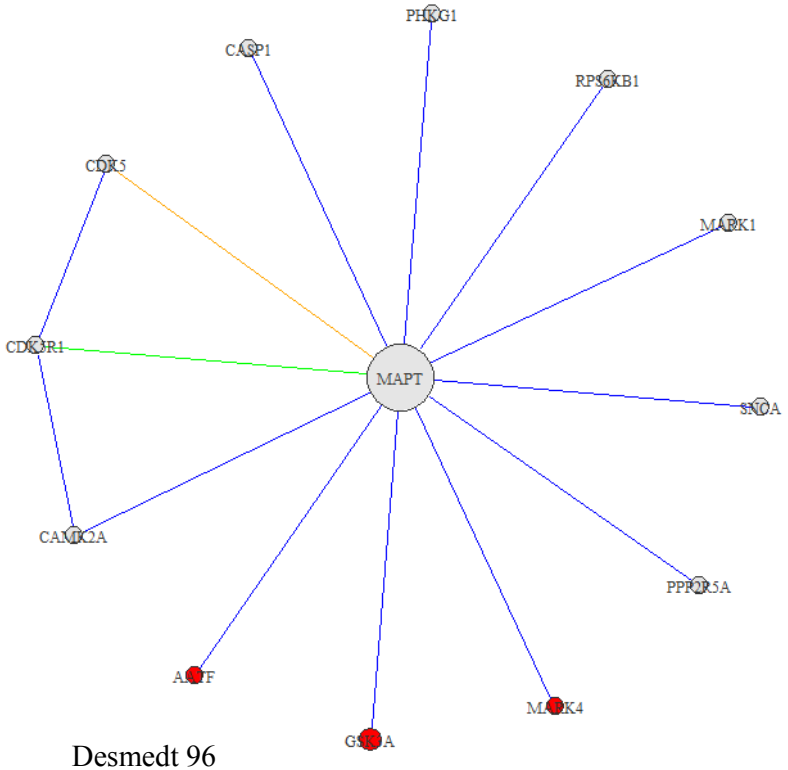
Desmedt 79



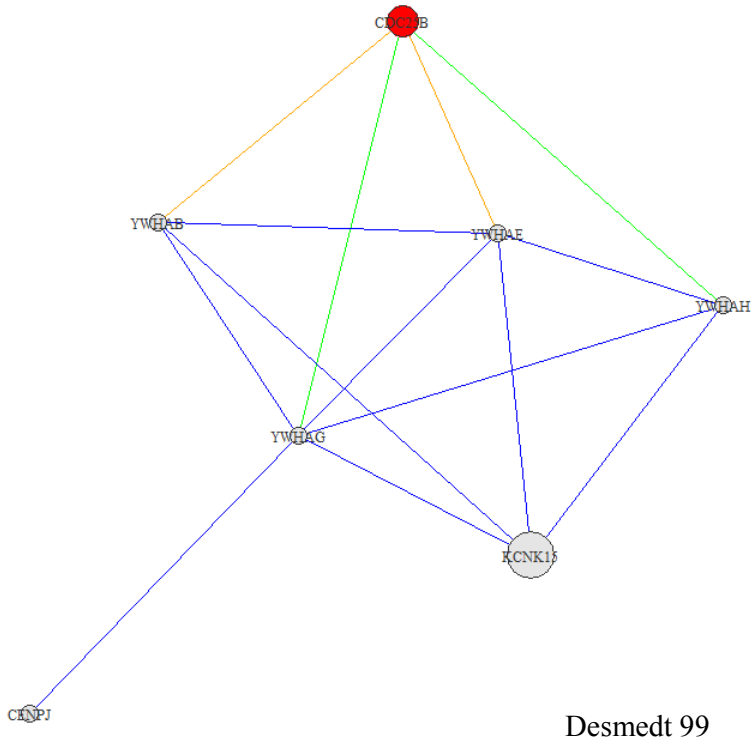
Desmedt 82



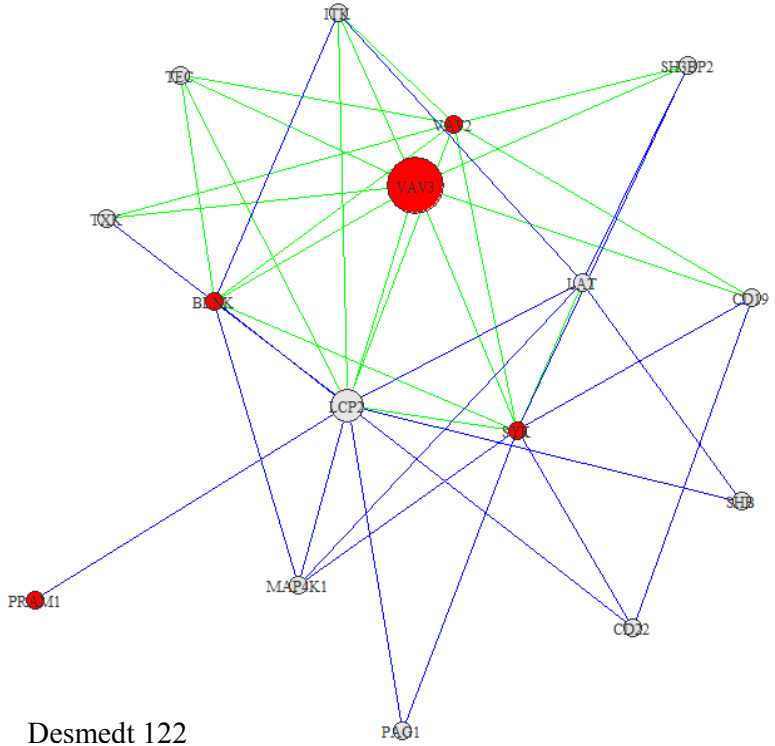
Desmedt 89



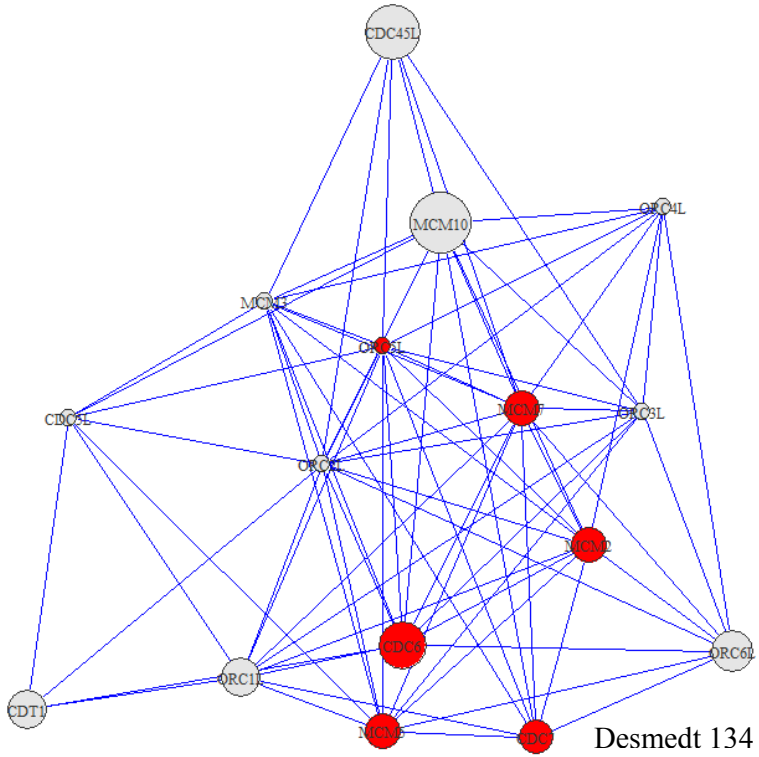
Desmedt 96



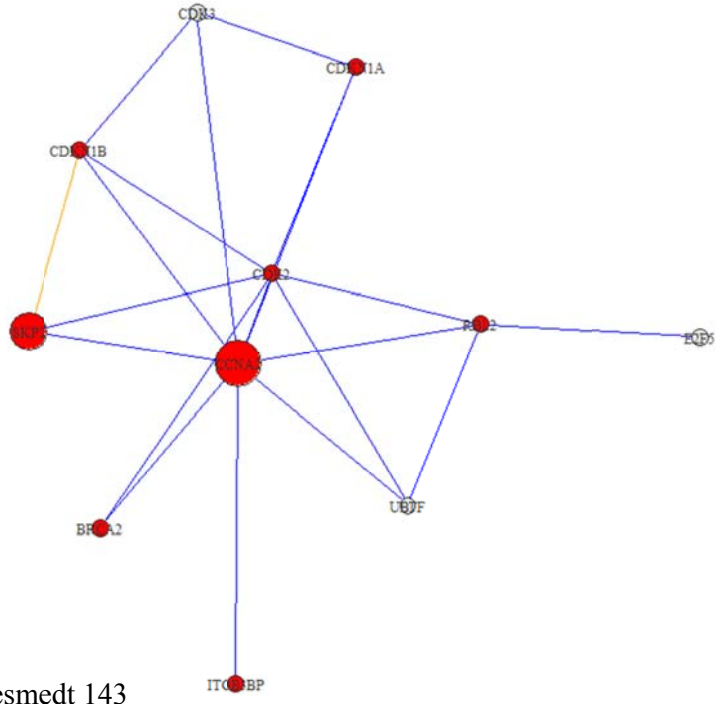
Desmedt 99



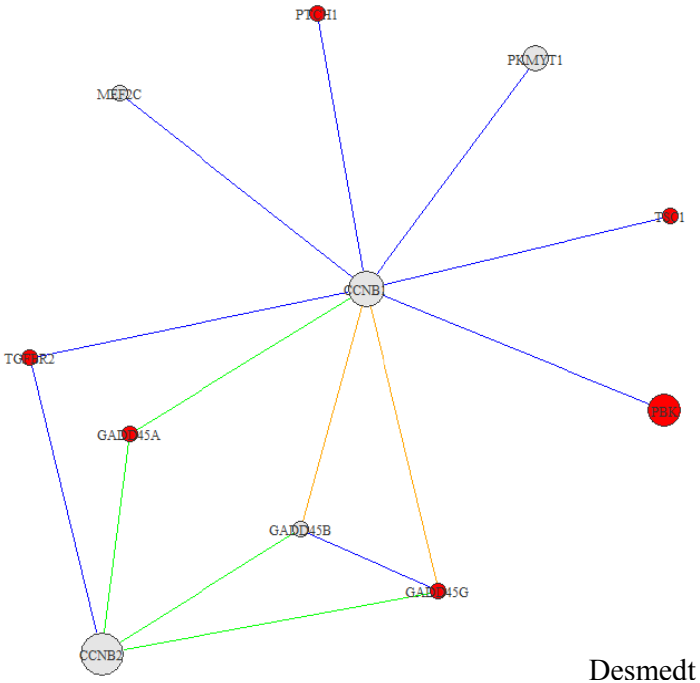
Desmedt 122



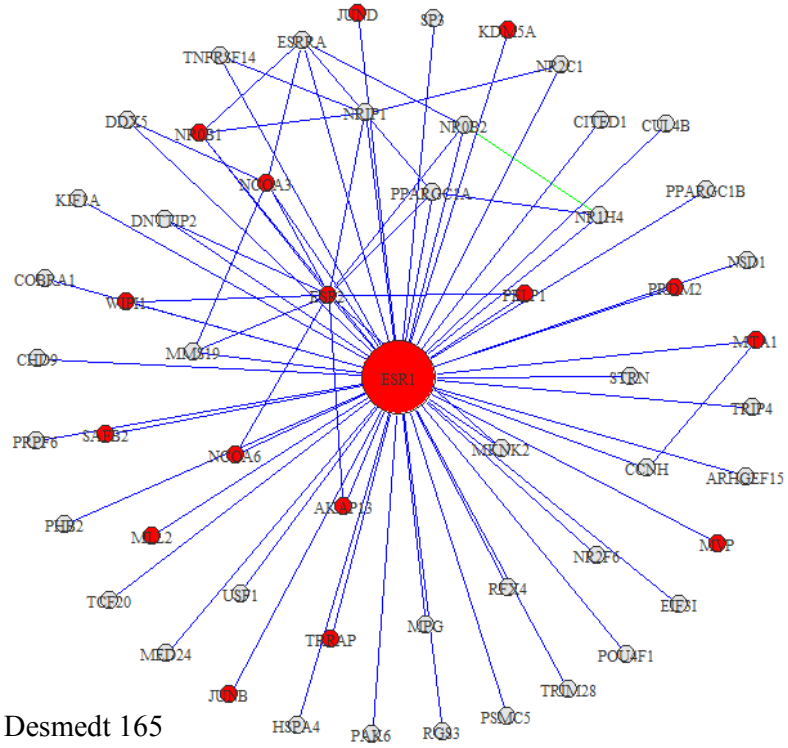
Desmedt 134



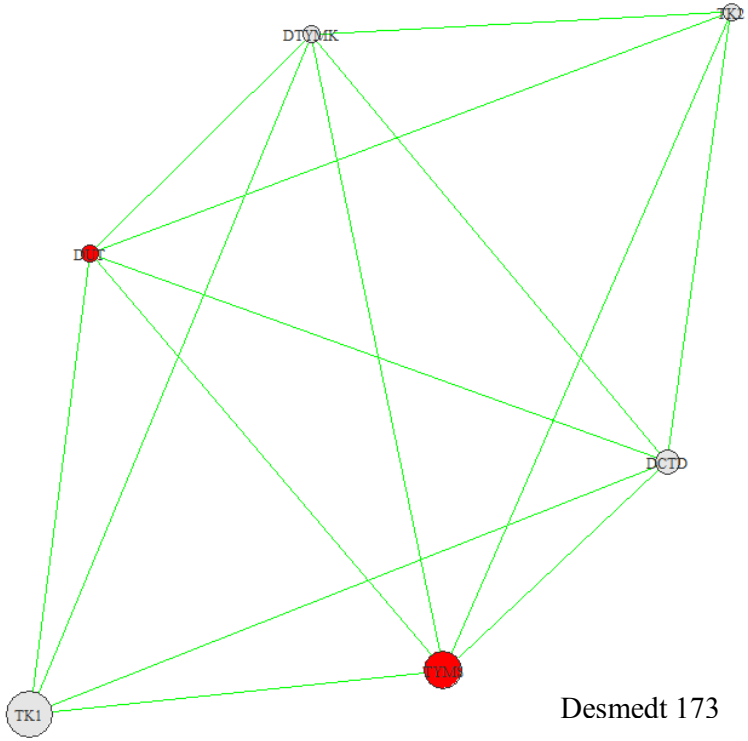
Desmedt 143



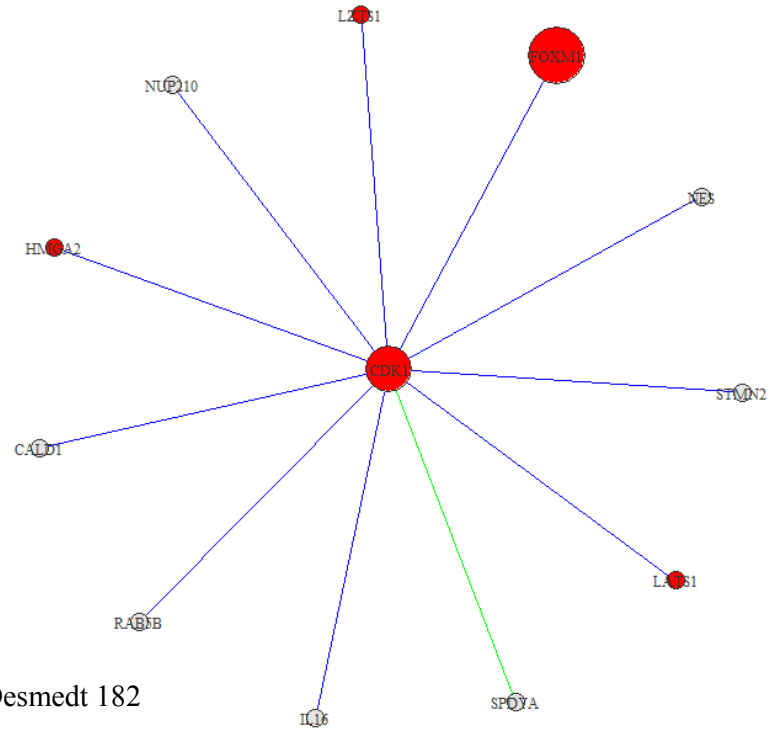
Desmedt 145



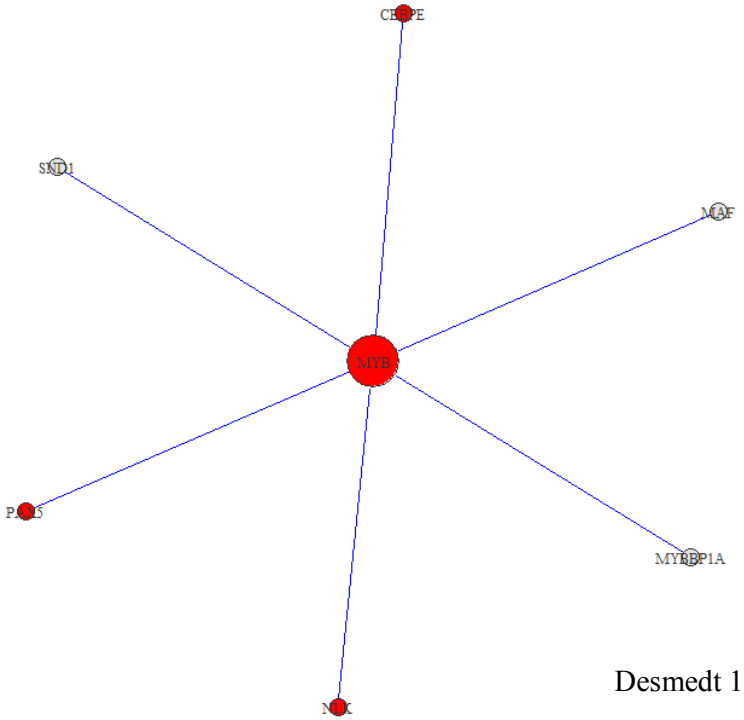
Desmedt 165



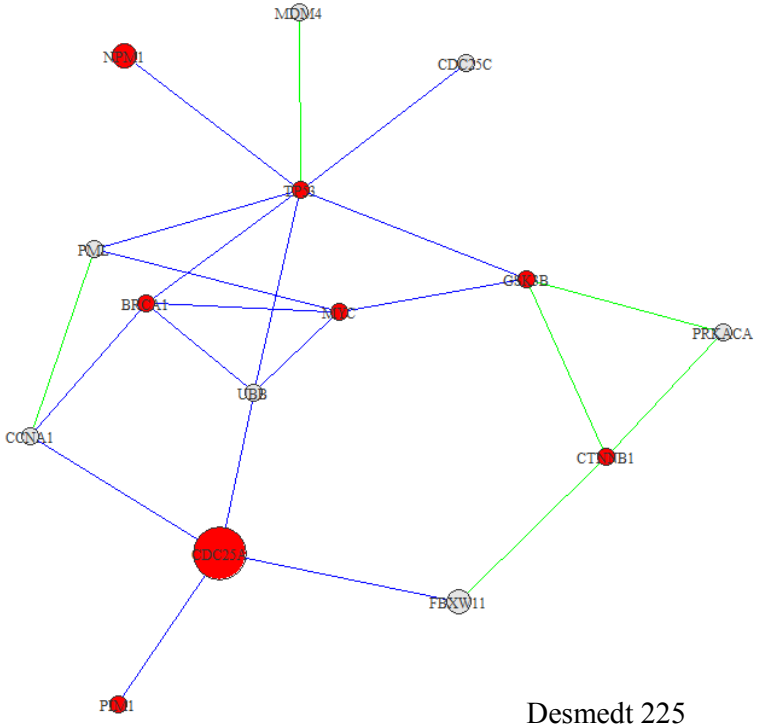
Desmedt 173



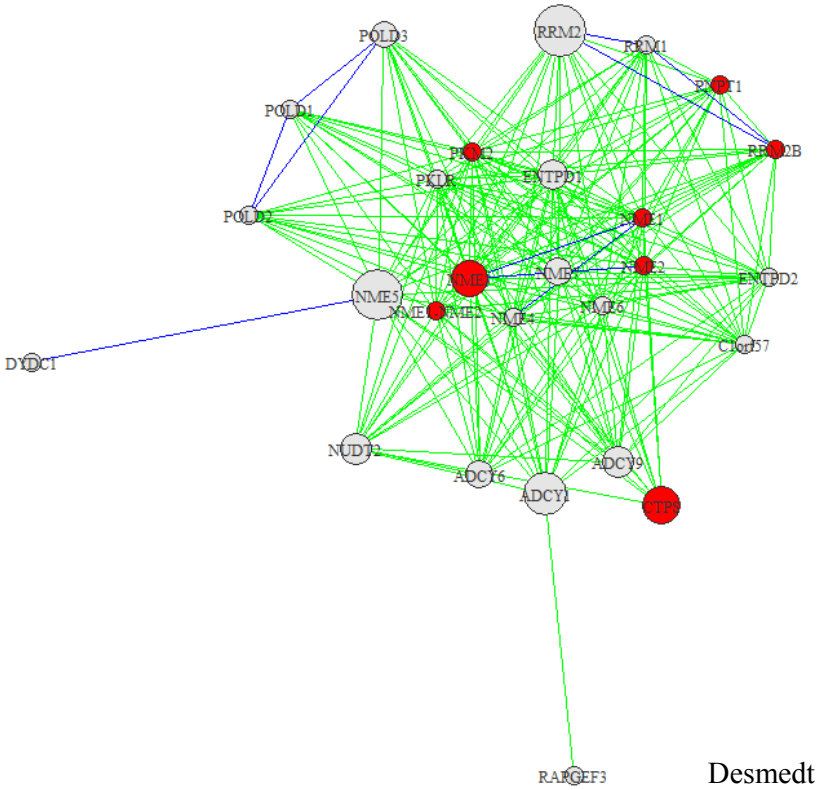
Desmedt 182



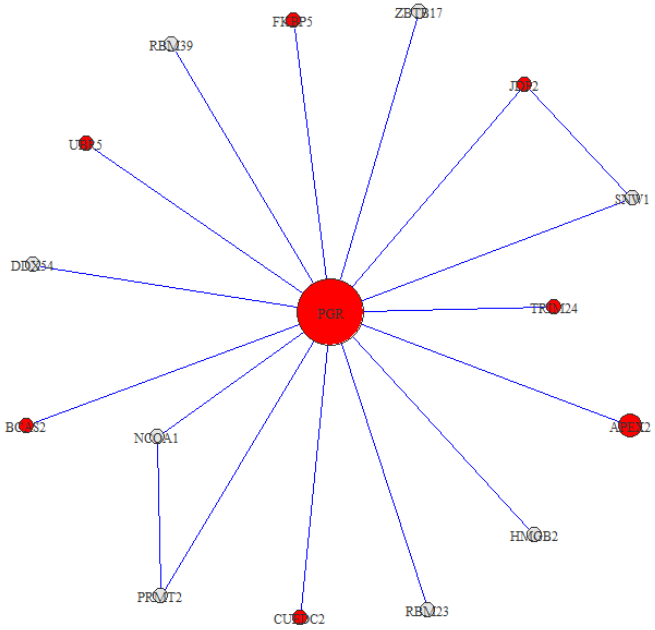
Desmedt 189



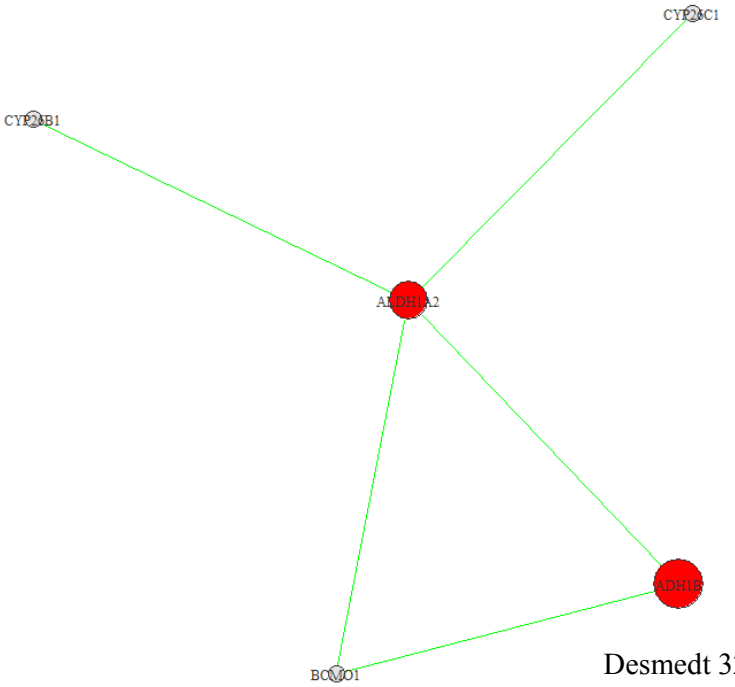
Desmedt 225



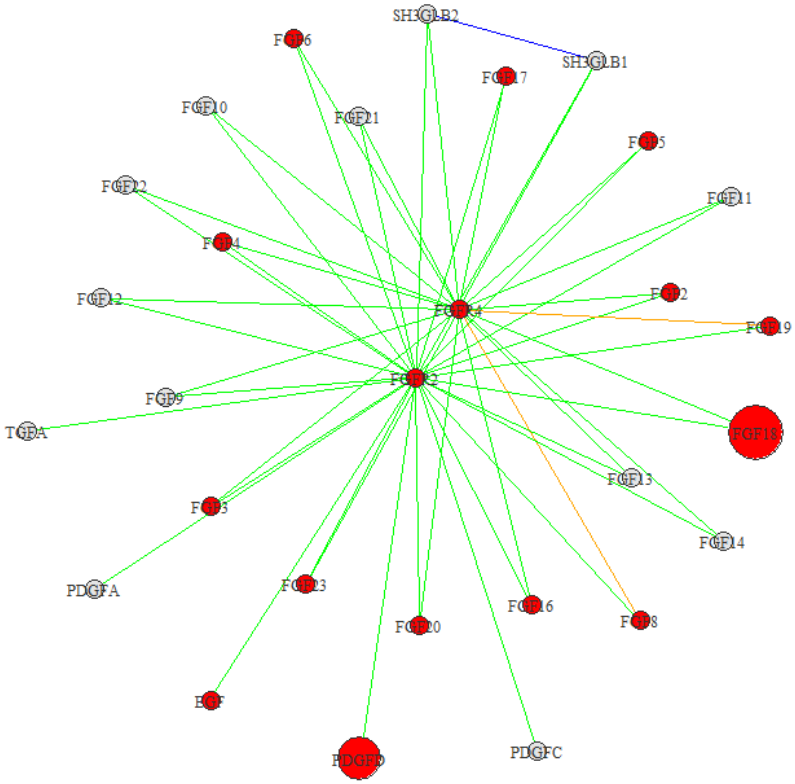
Desmedt 224



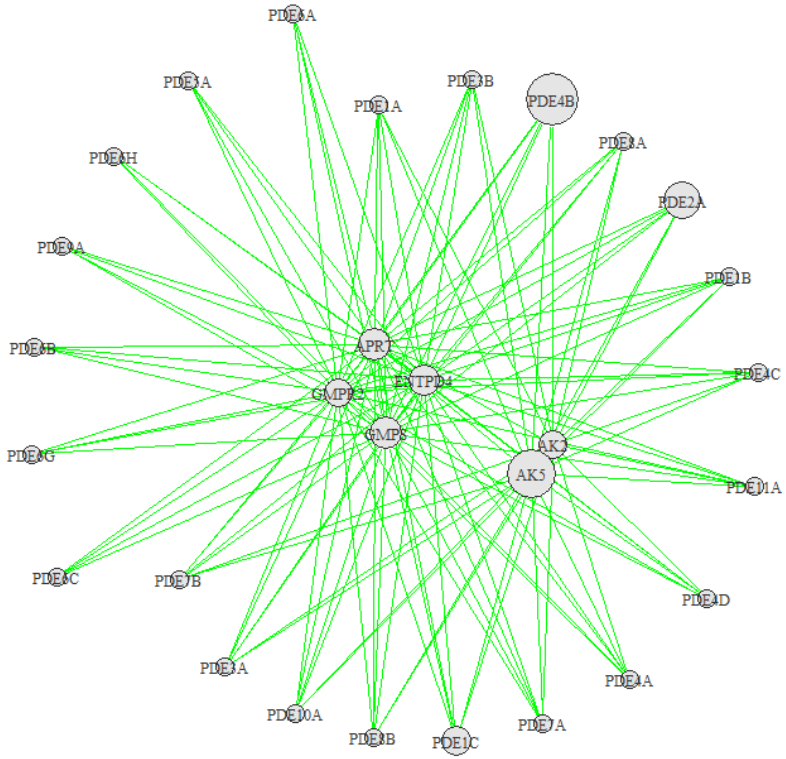
Desmedt 226



Desmedt 321

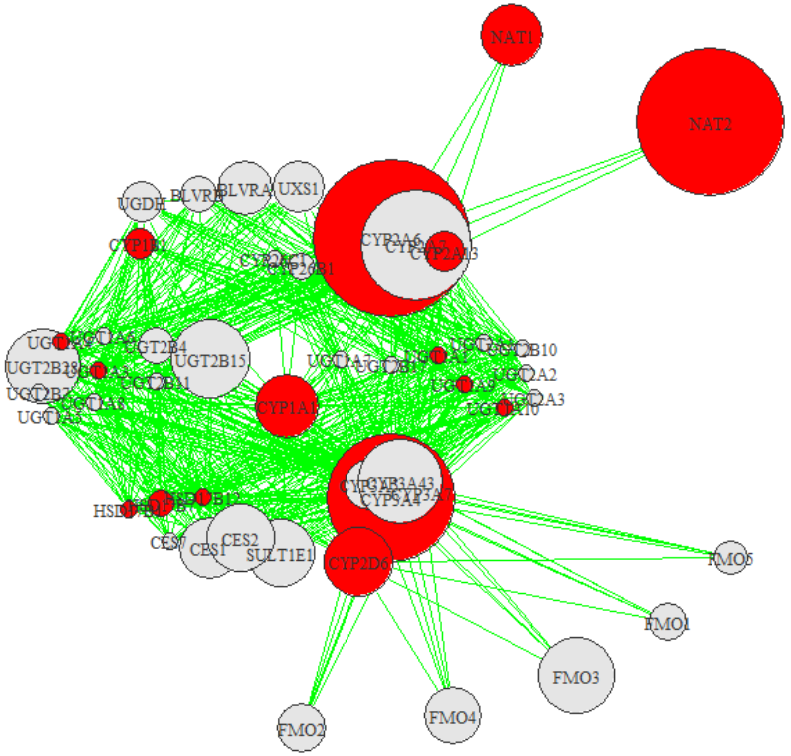


Desmedt 348

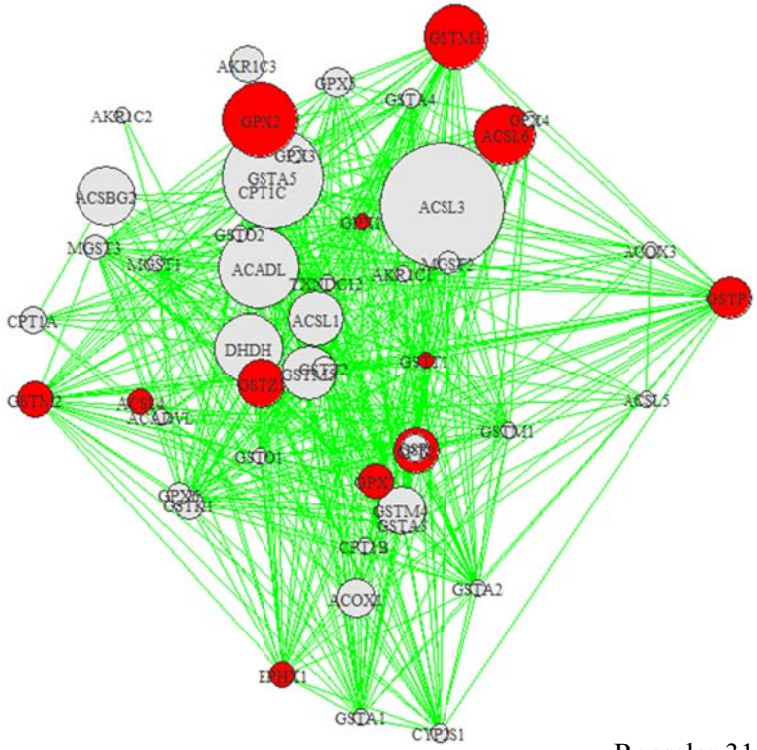


Desmedt 395

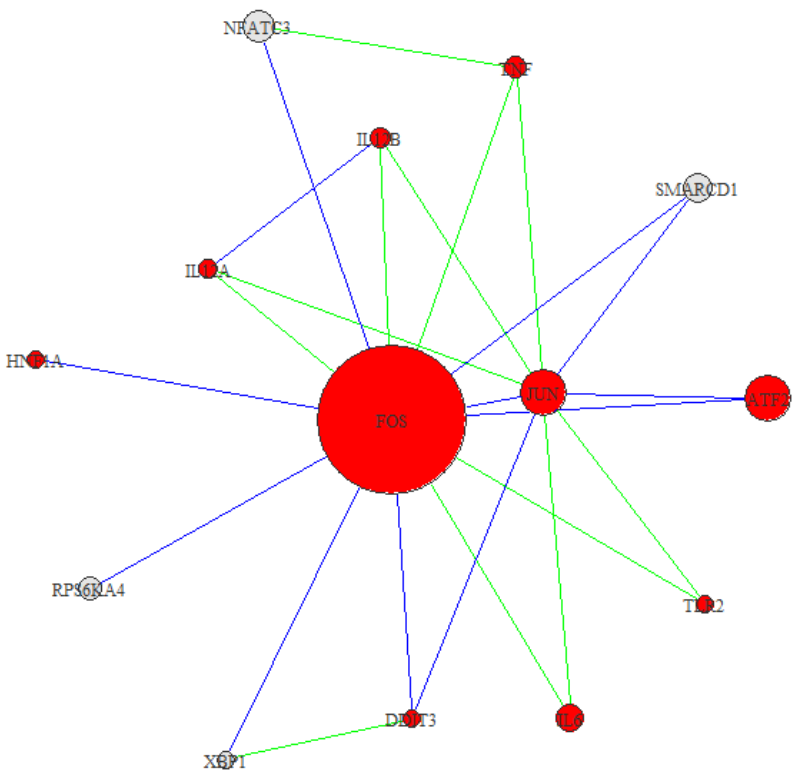
Significant Modules in Roessler 2010 Data



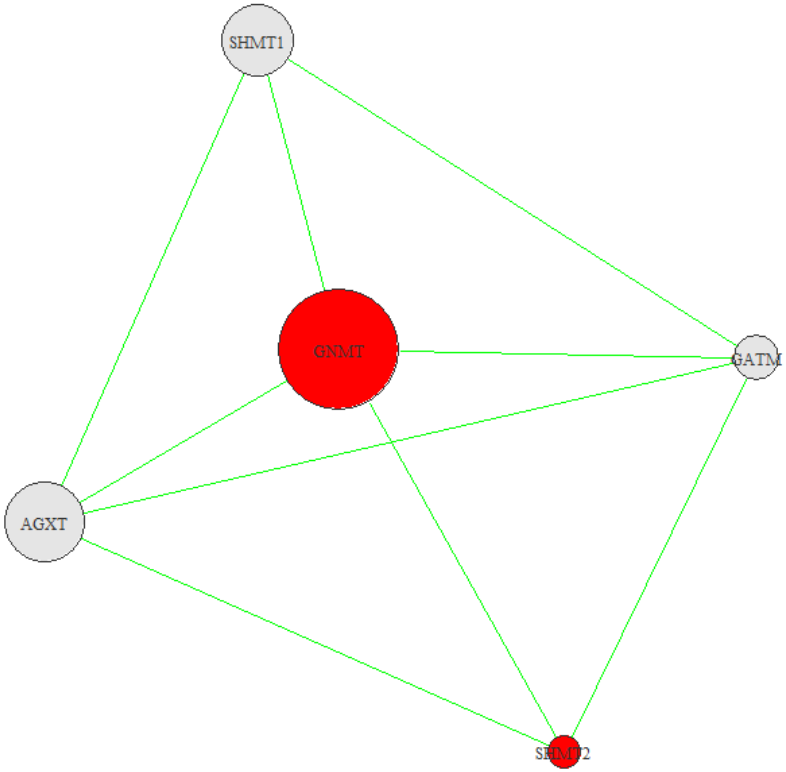
Roessler 10



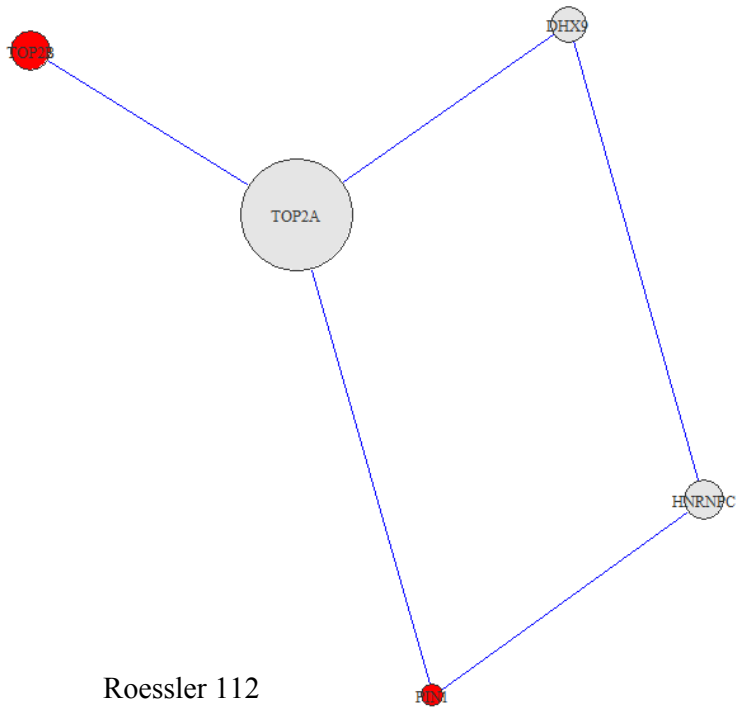
Roessler 31



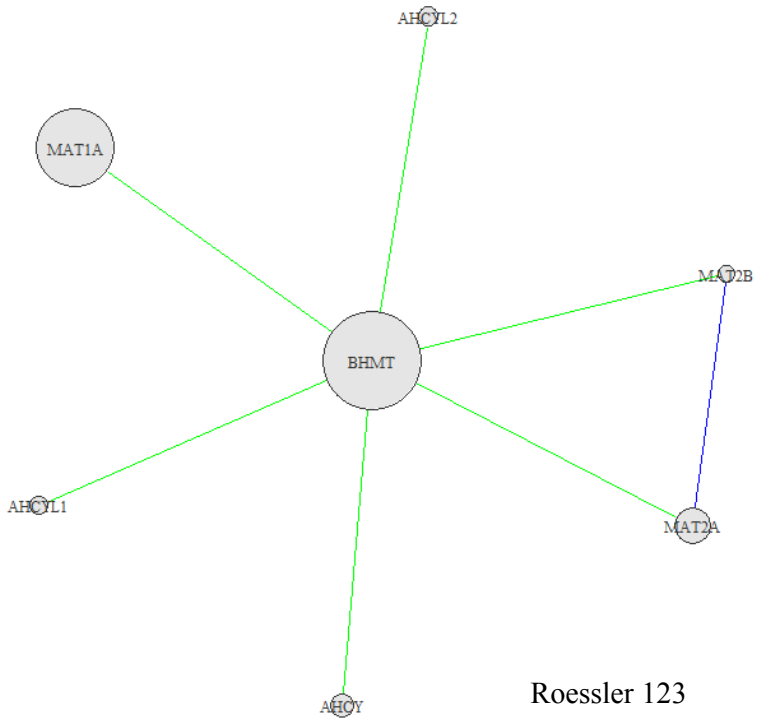
Roessler 34



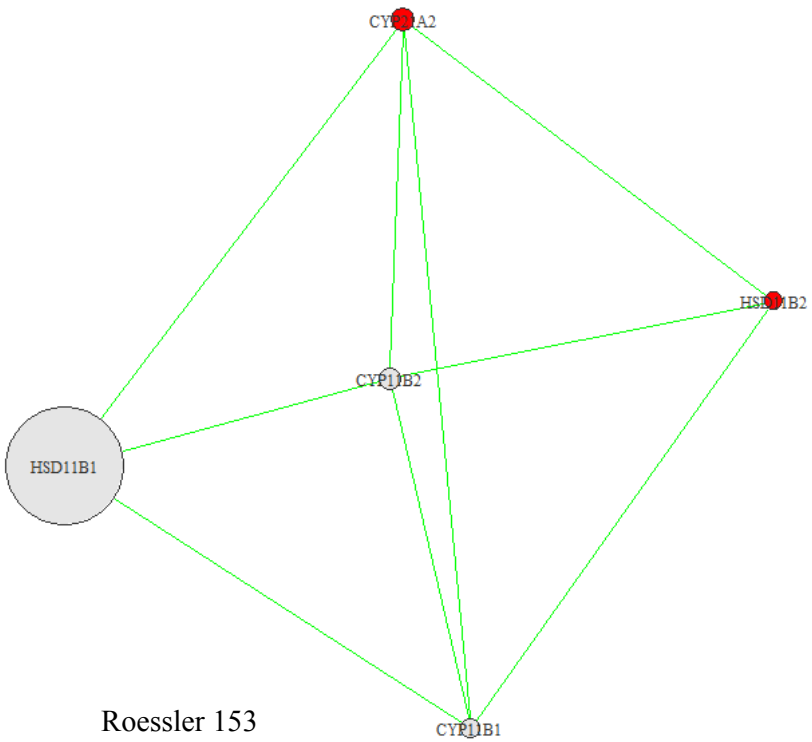
Roessler 97



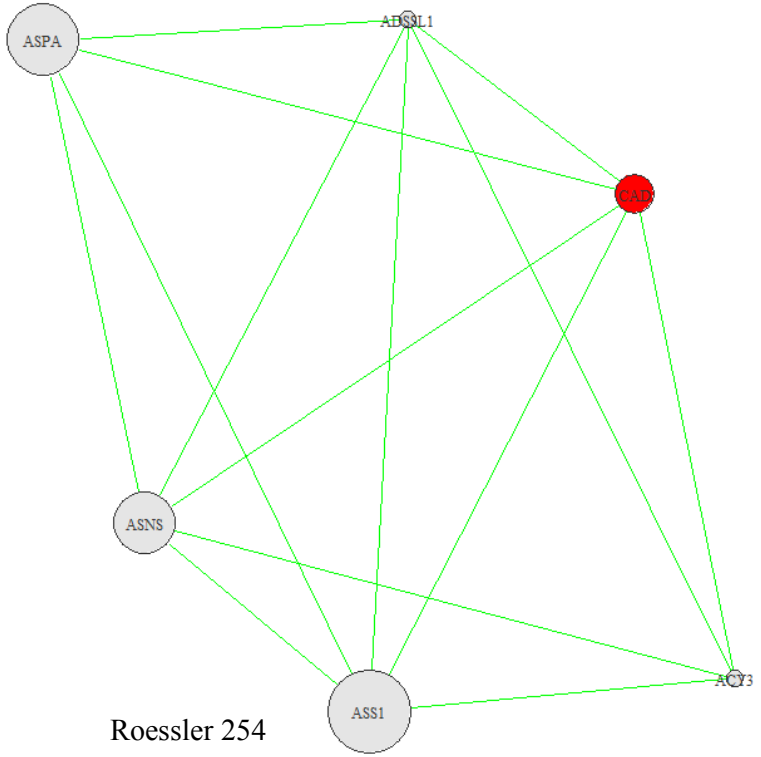
Roessler 112



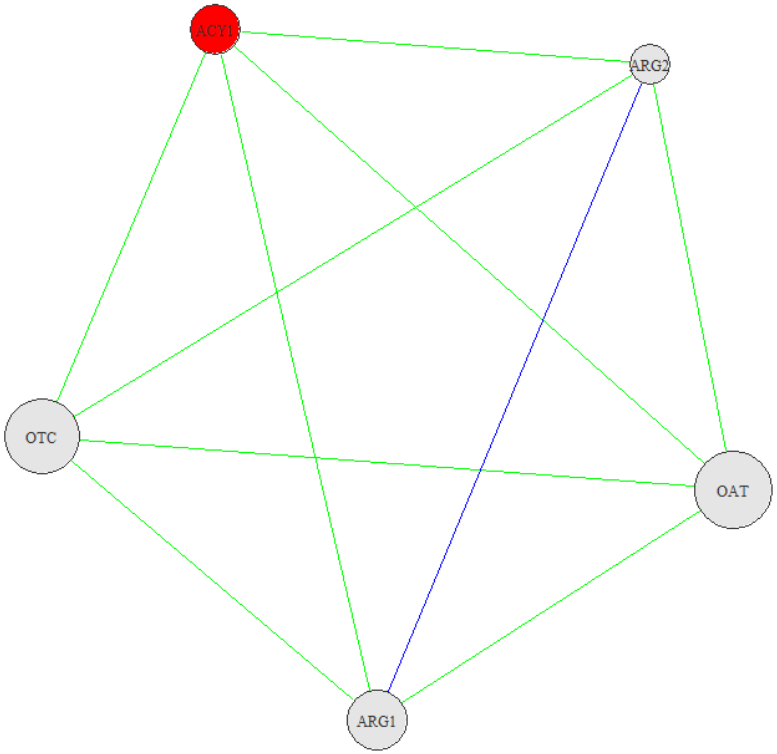
Roessler 123



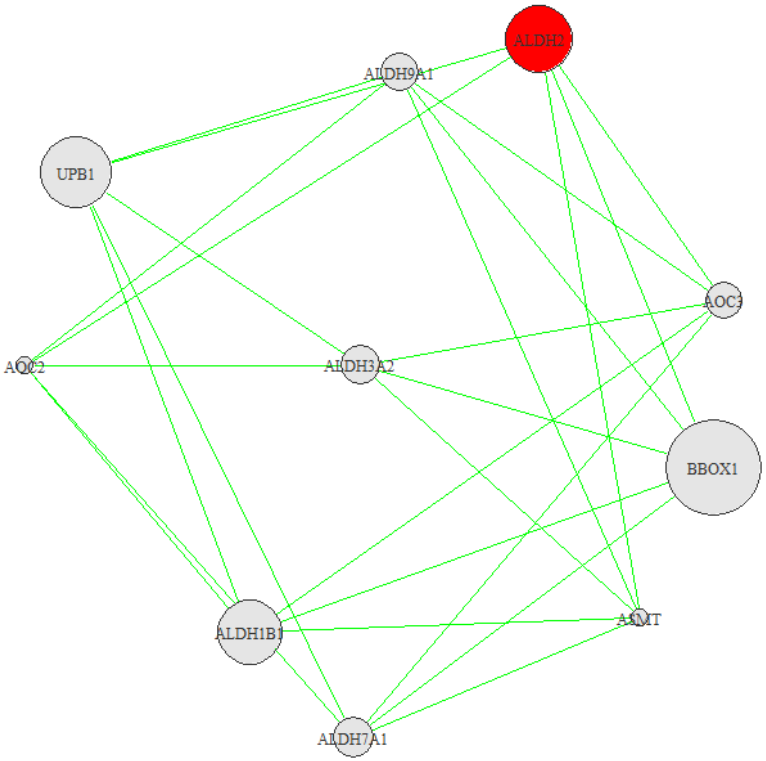
Roessler 153



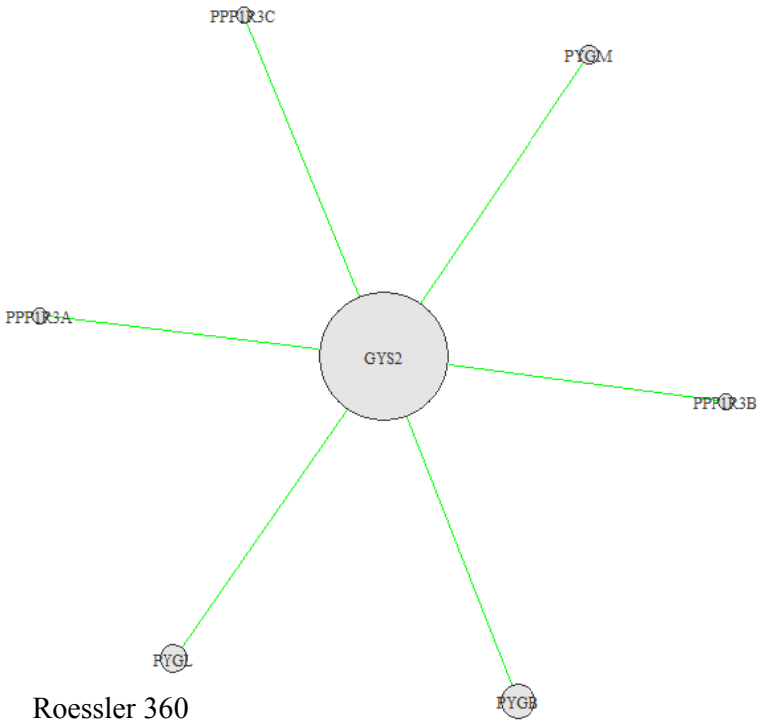
Roessler 254



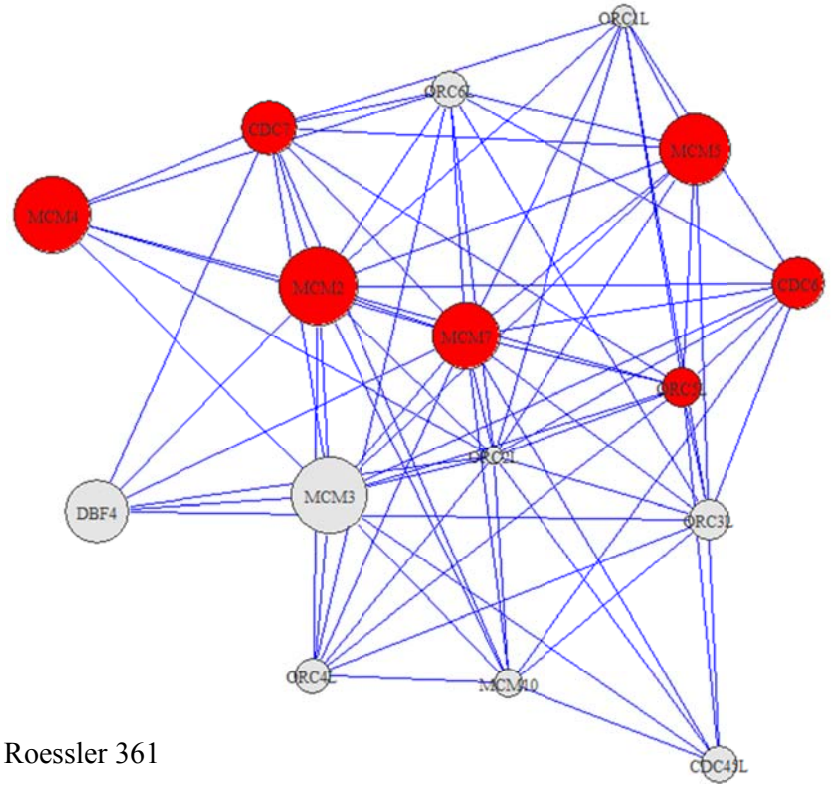
Roessler 257



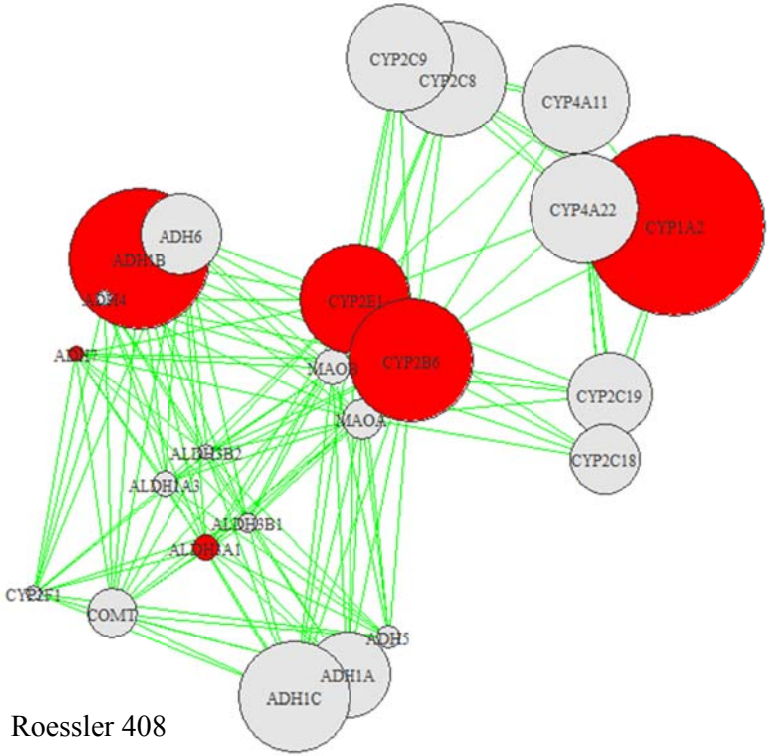
Roessler 314



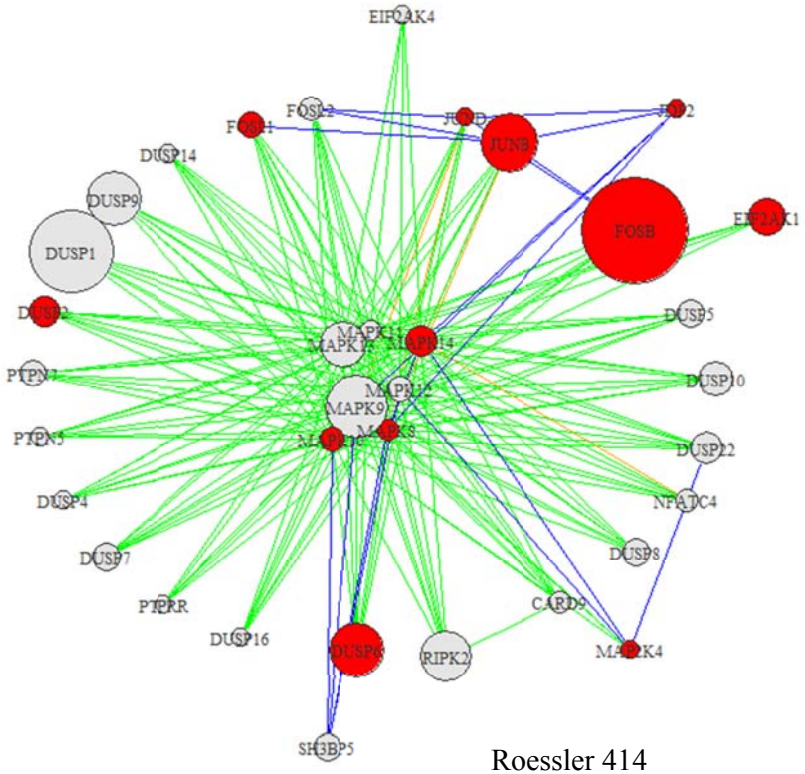
Roessler 360



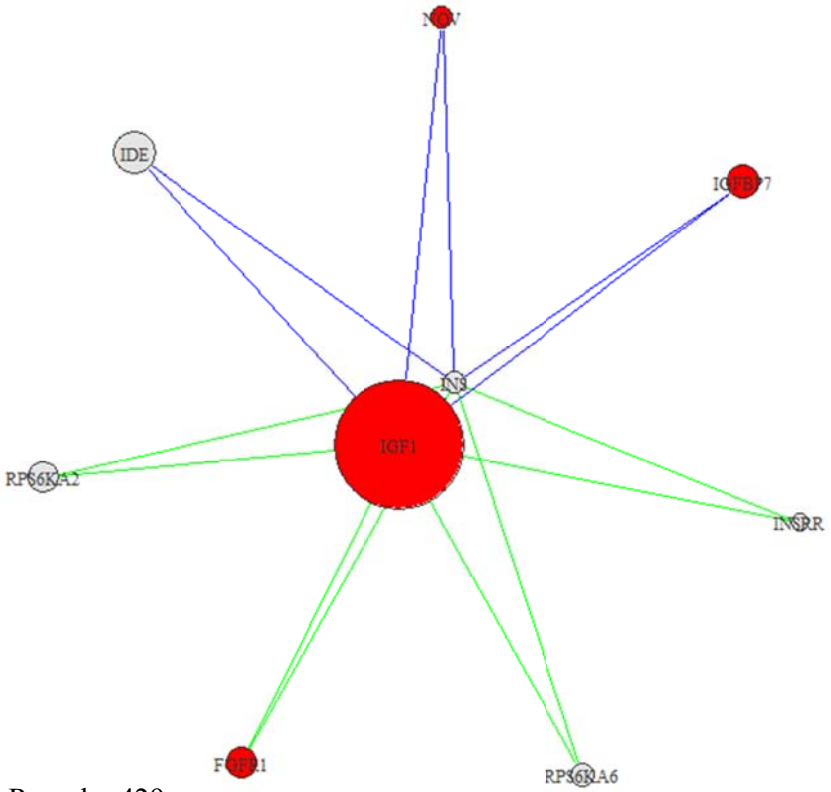
Roessler 361



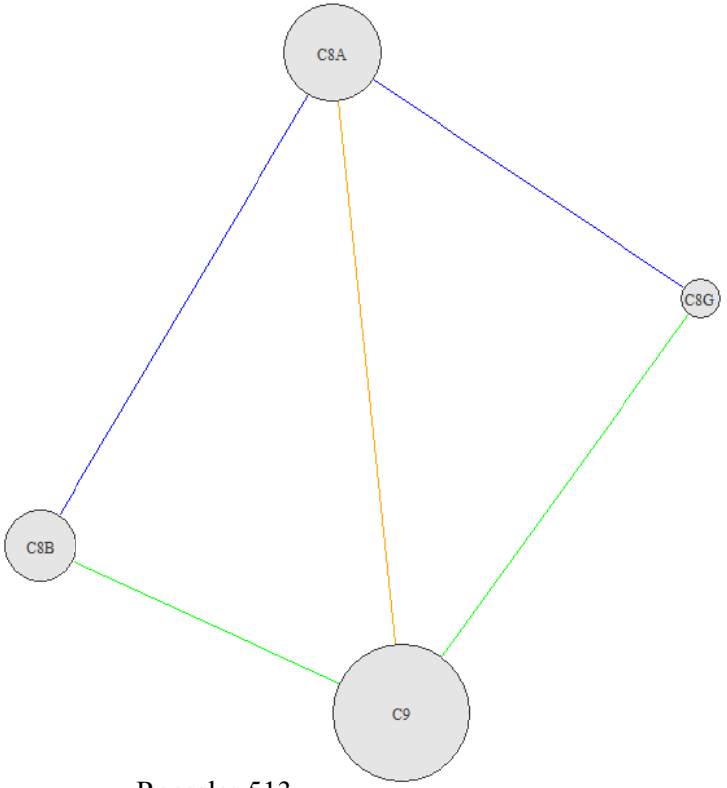
Roessler 408



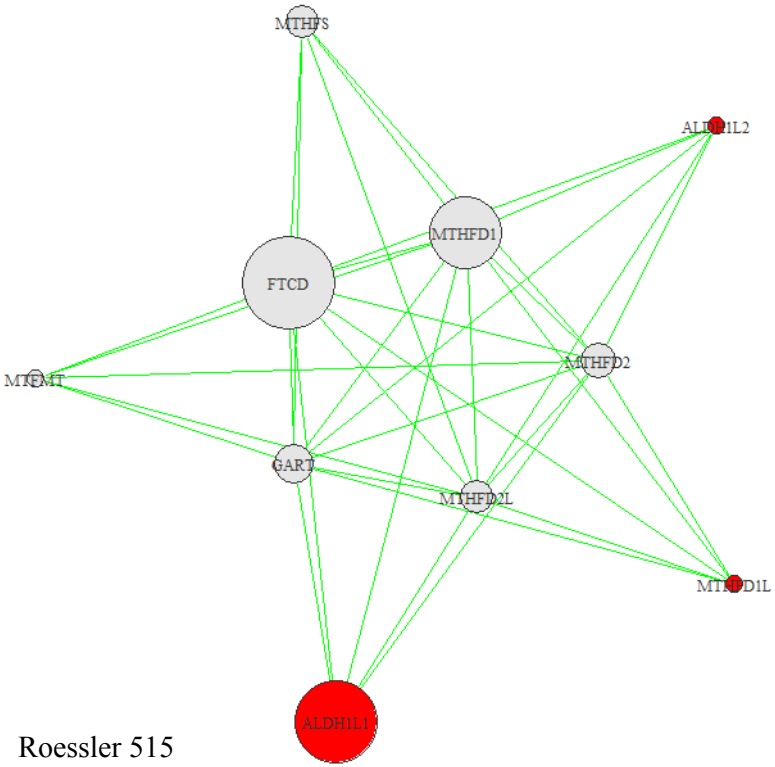
Roessler 414



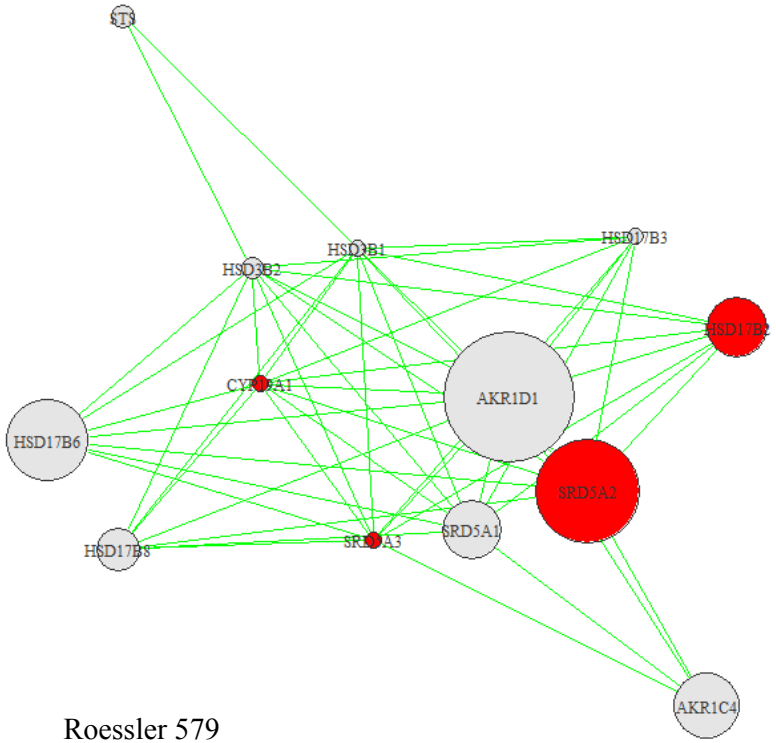
Roessler 429



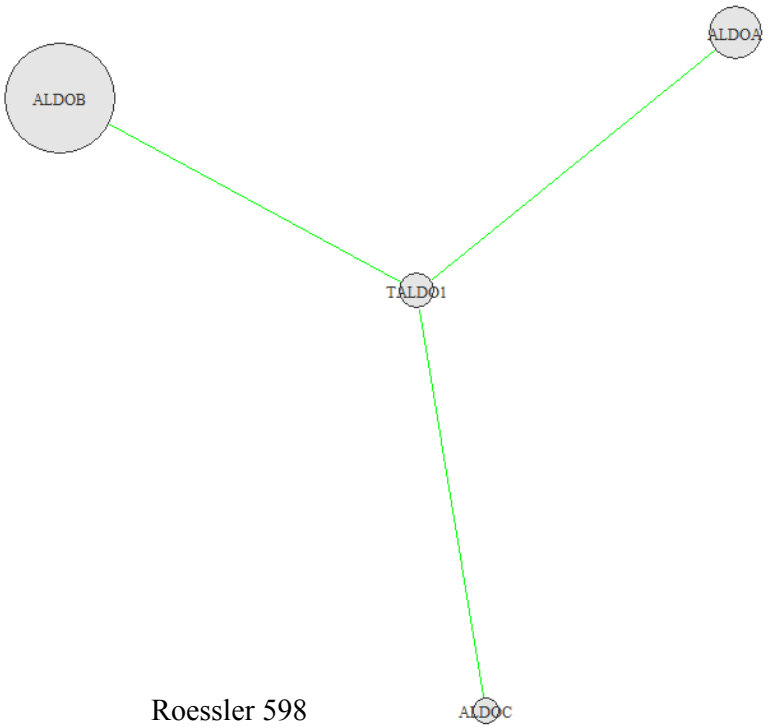
Roessler 513



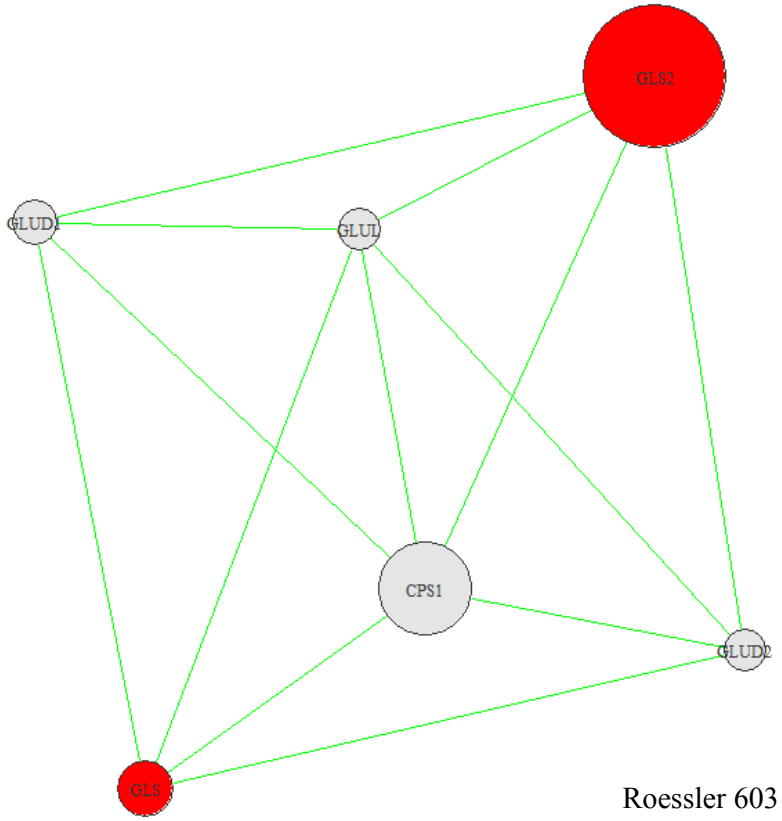
Roessler 515



Roessler 579

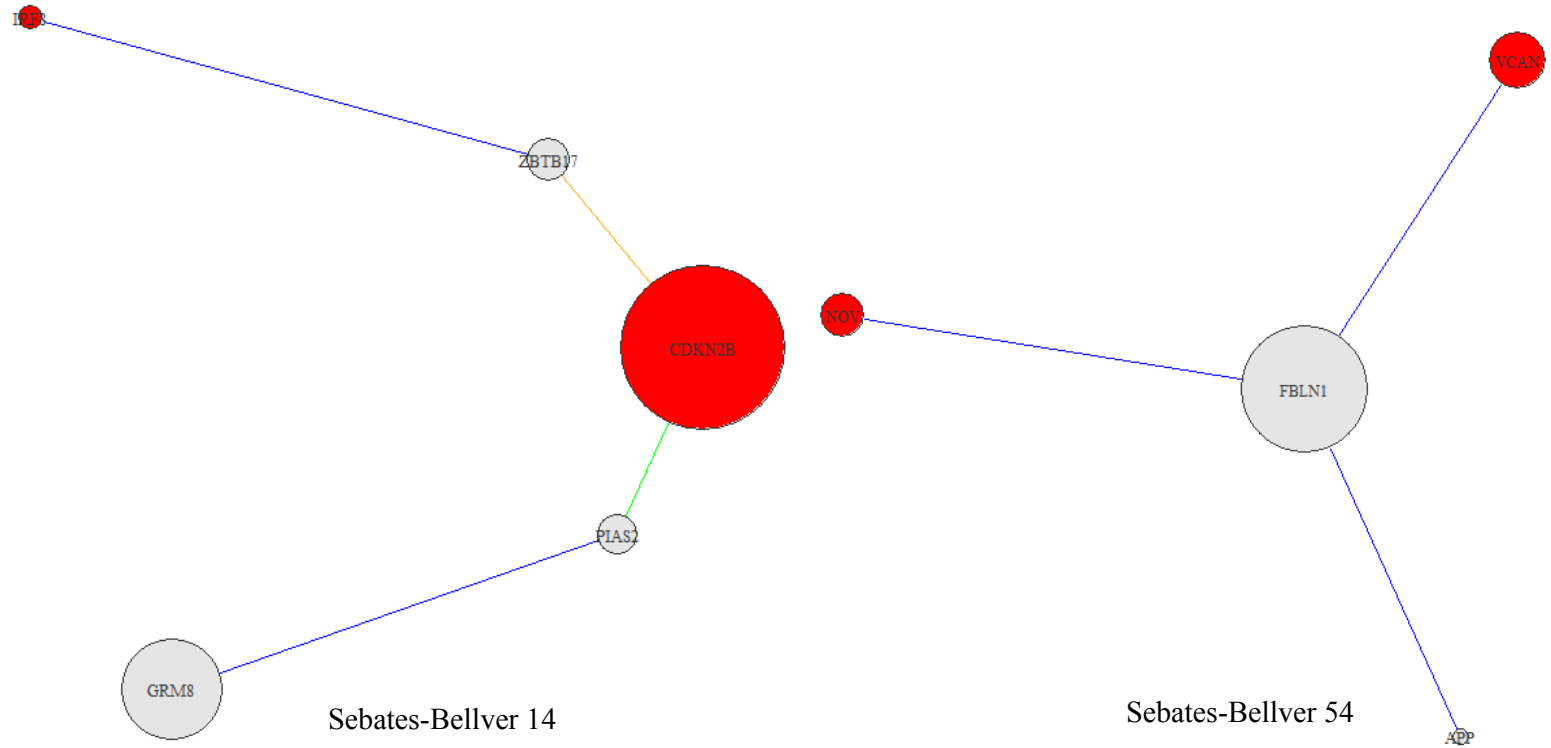


Roessler 598



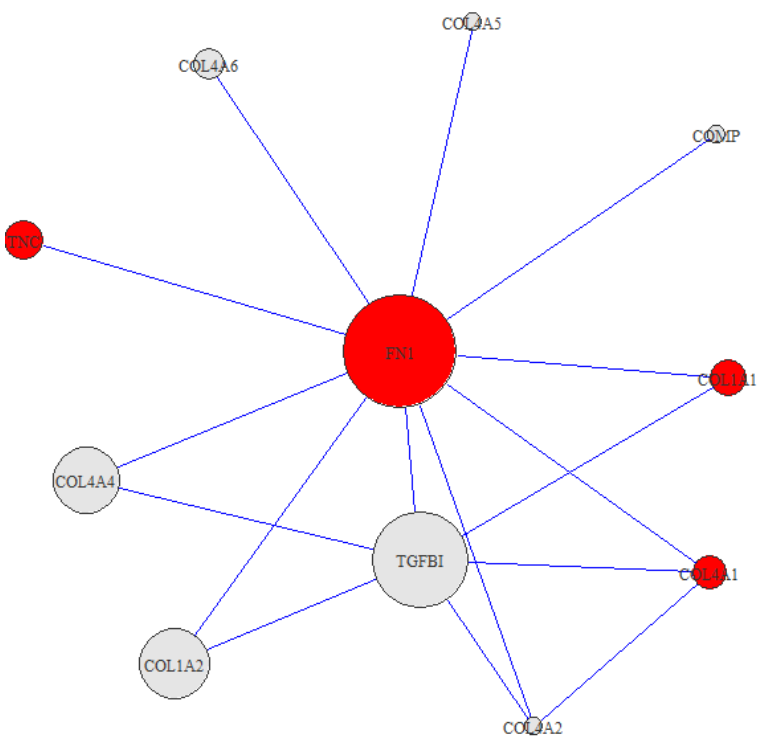
Roessler 603

Significant Modules in Sebates-Bellver 2007 Data

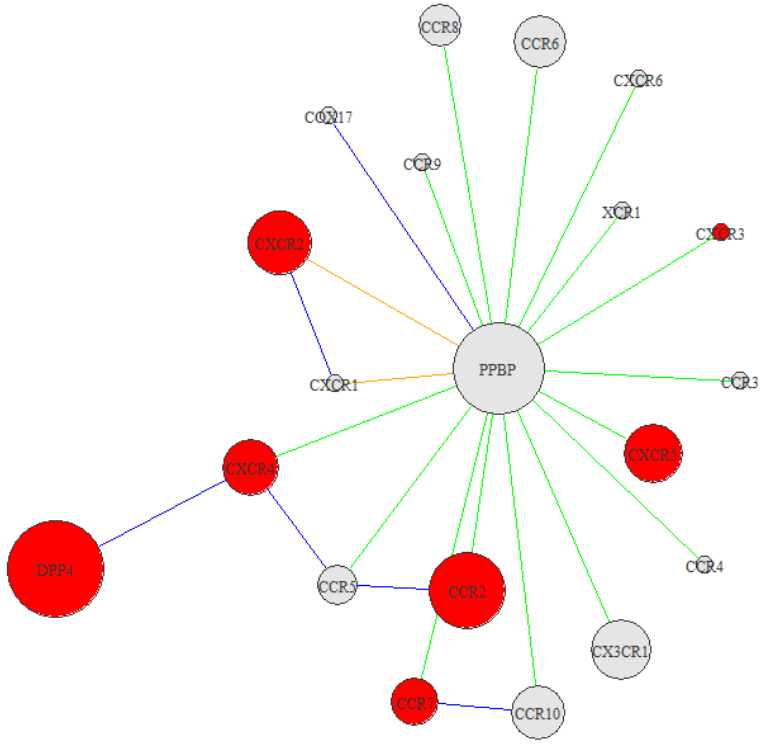


Sebates-Bellver 14

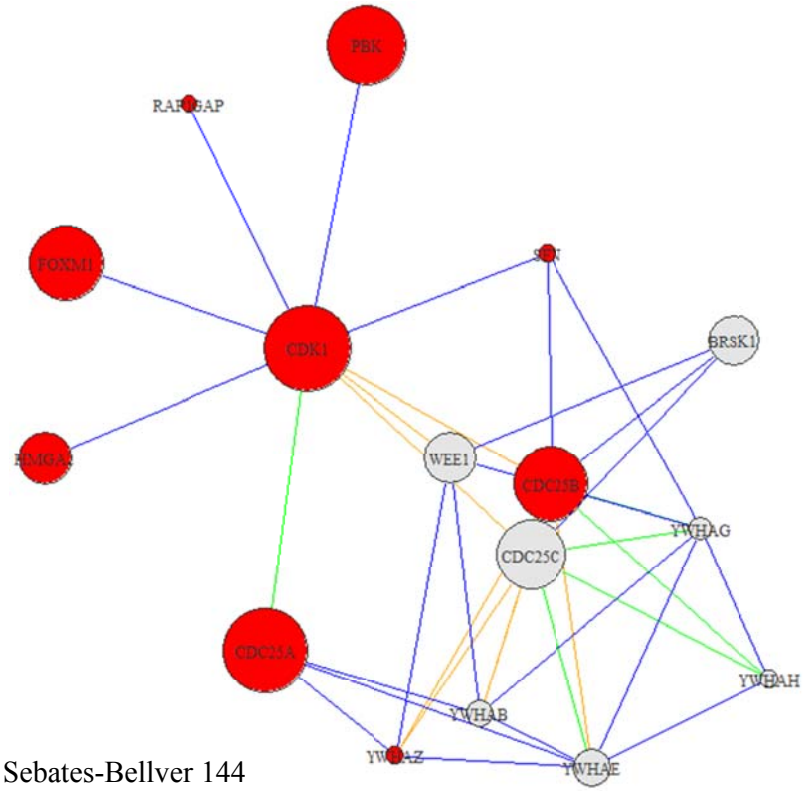
Sebates-Bellver 54



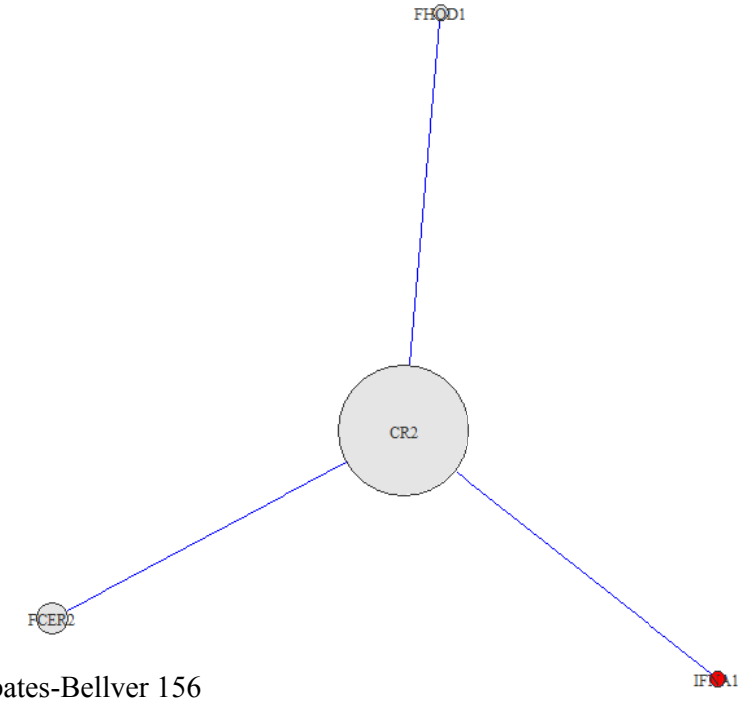
Sebates-Bellver 111



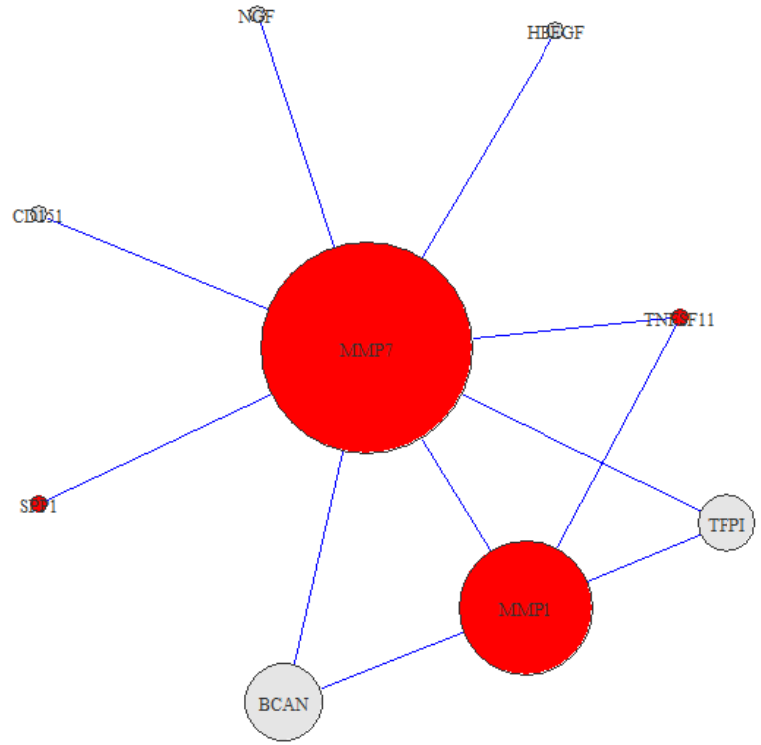
Sebates-Bellver 25



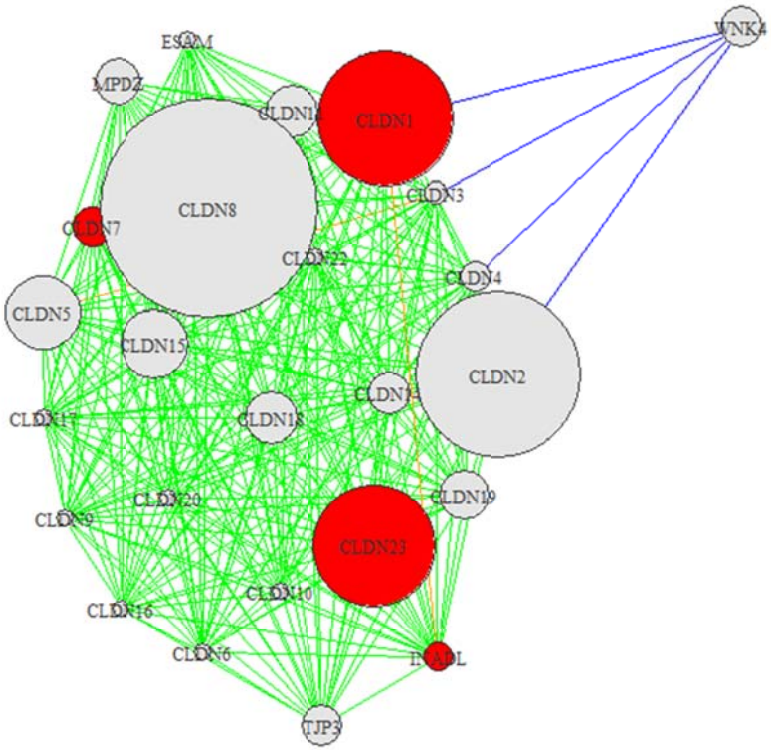
Sebates-Bellver 144



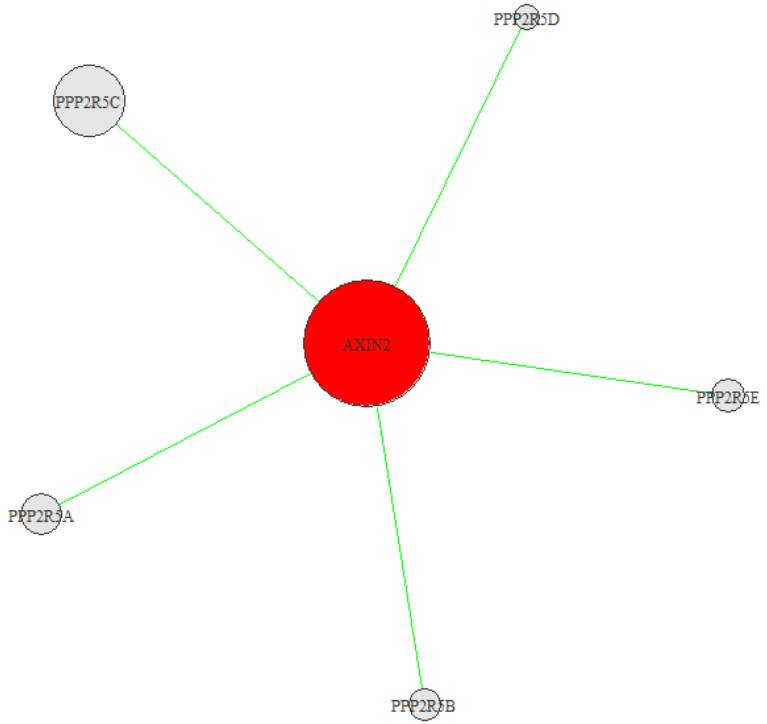
Sebates-Bellver 156



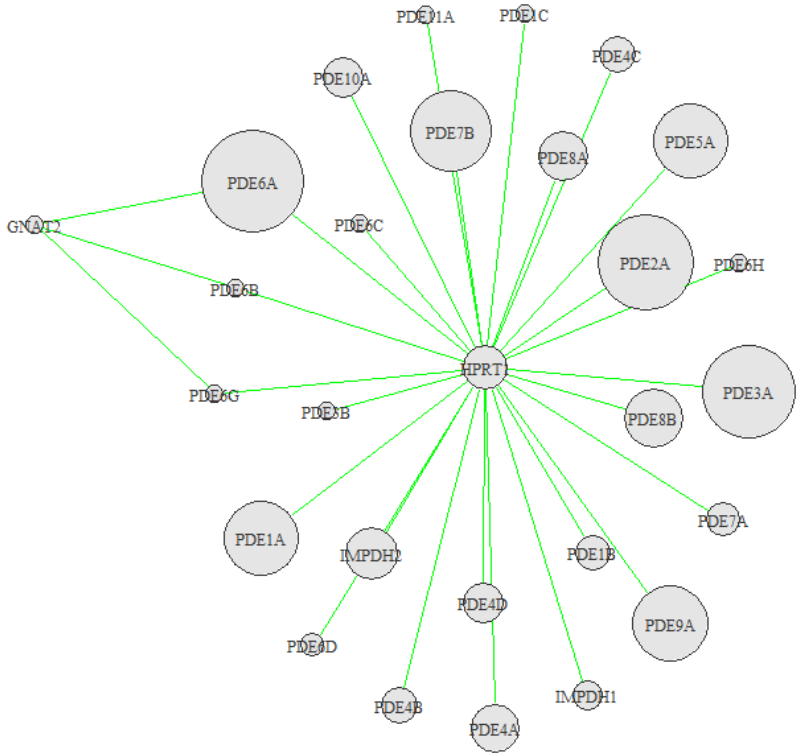
Sebates-Bellver 158



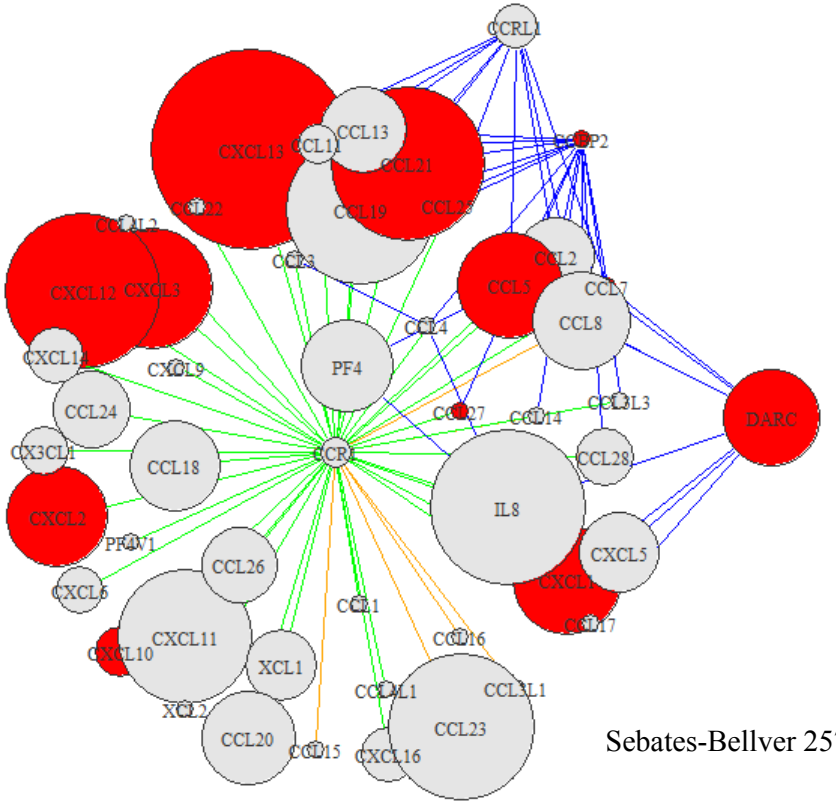
Sebates-Bellver 182



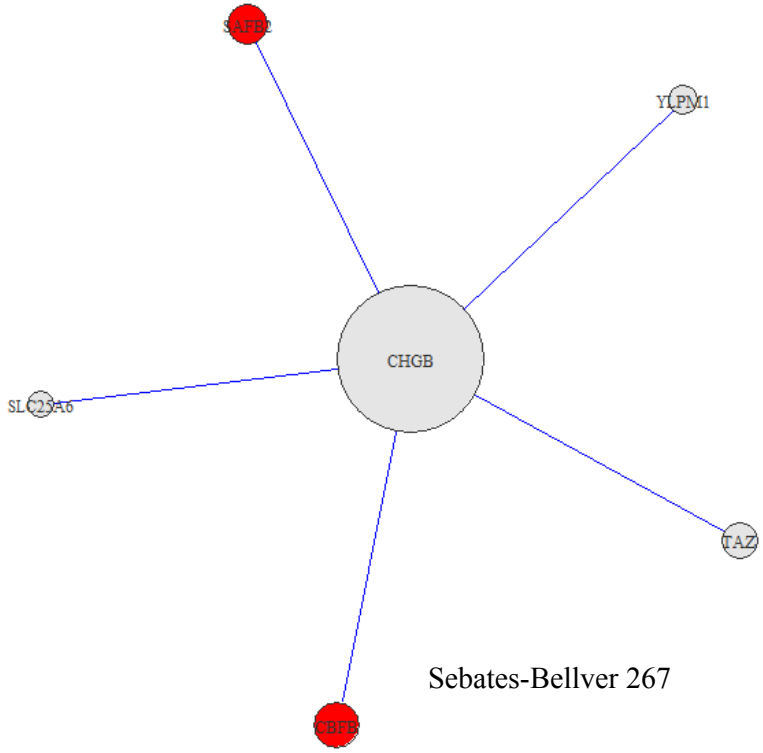
Sebates-Bellver 183



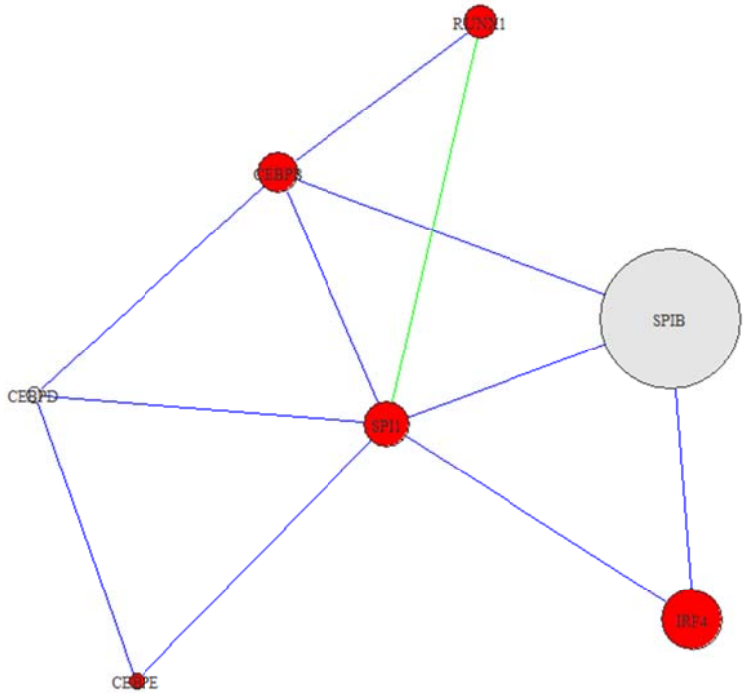
Sebates-Bellver 240



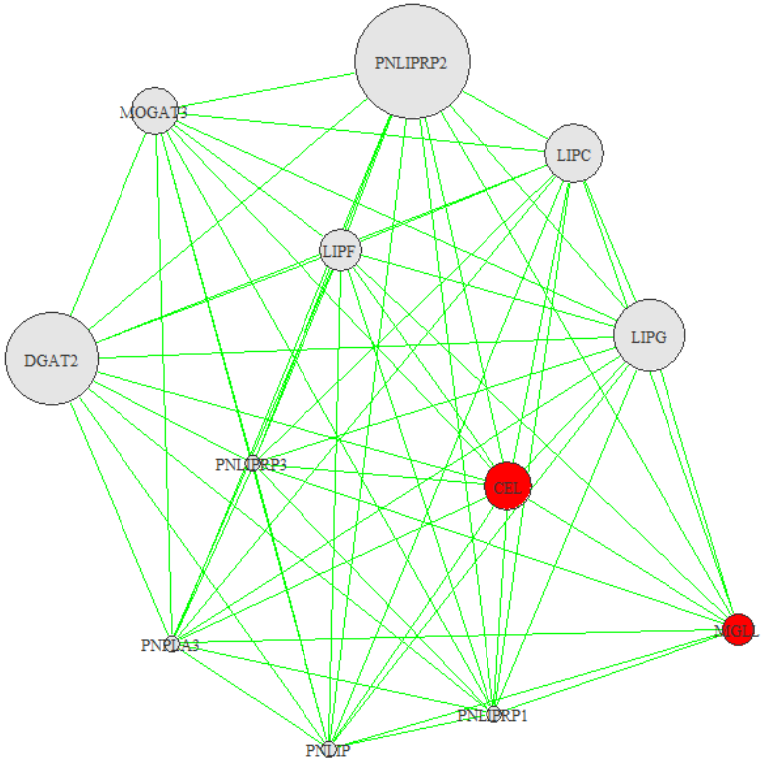
Sebates-Bellver 257



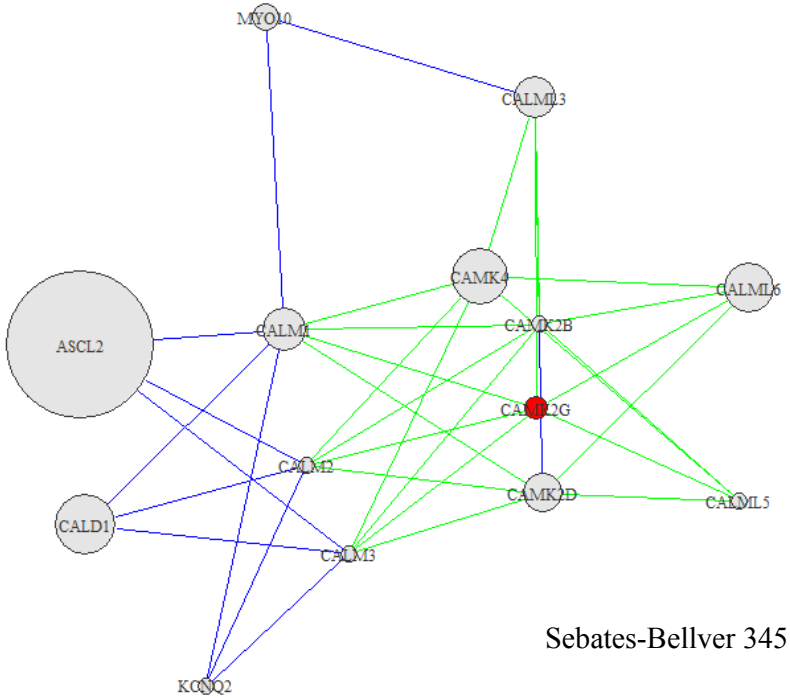
Sebates-Bellver 267



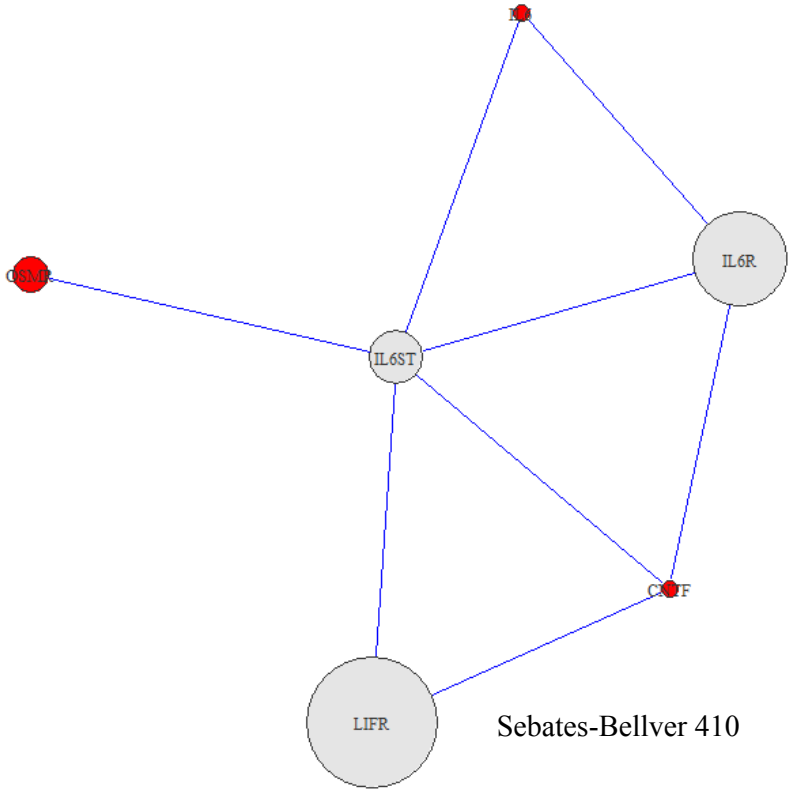
Sebates-Bellver 301



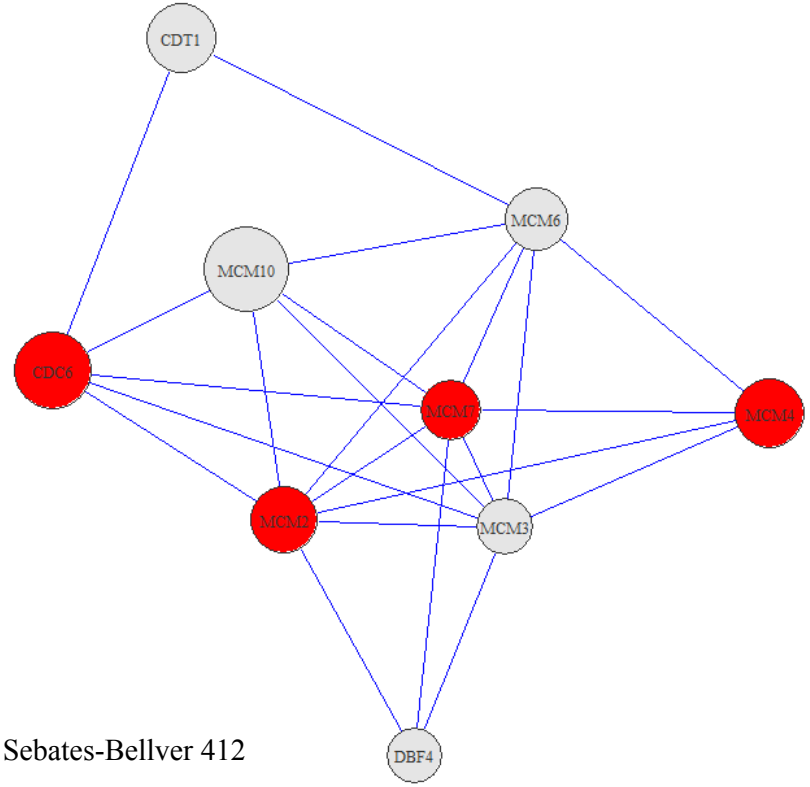
Sebates-Bellver 334



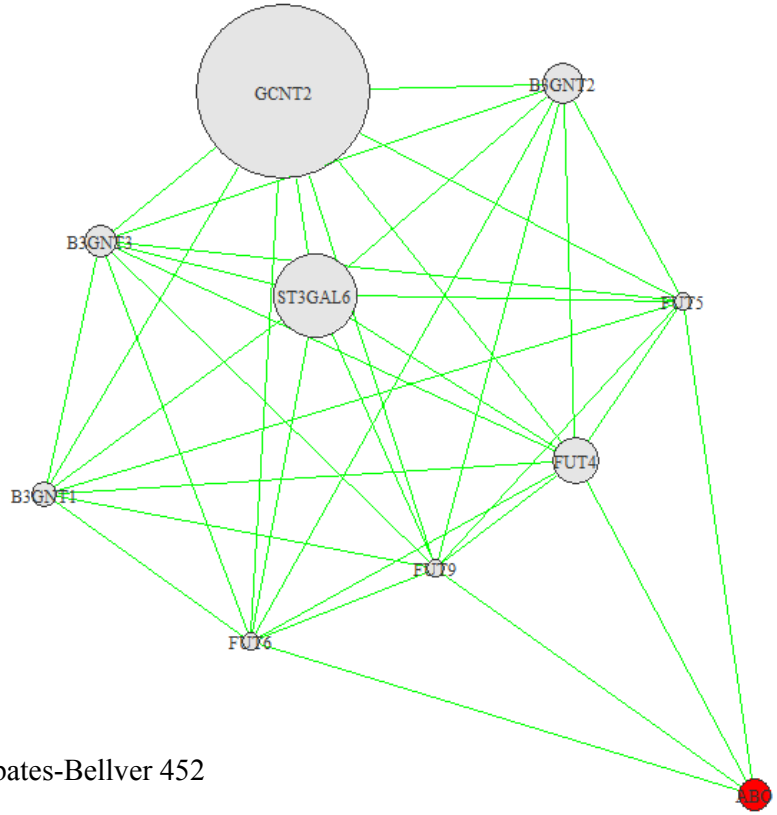
Sebates-Bellver 345



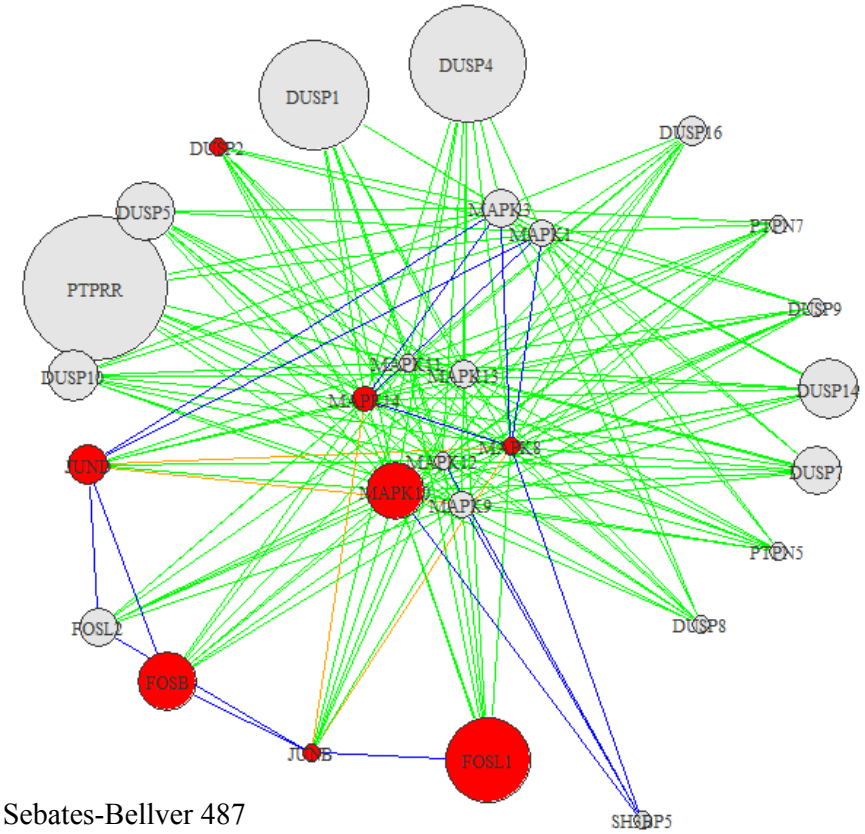
Sebates-Bellver 410



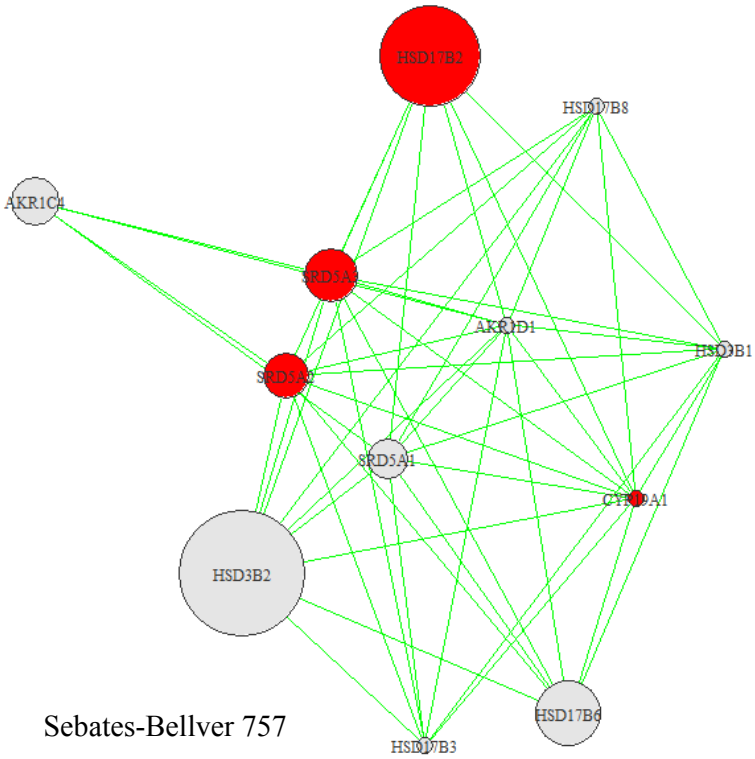
Sebates-Bellver 412



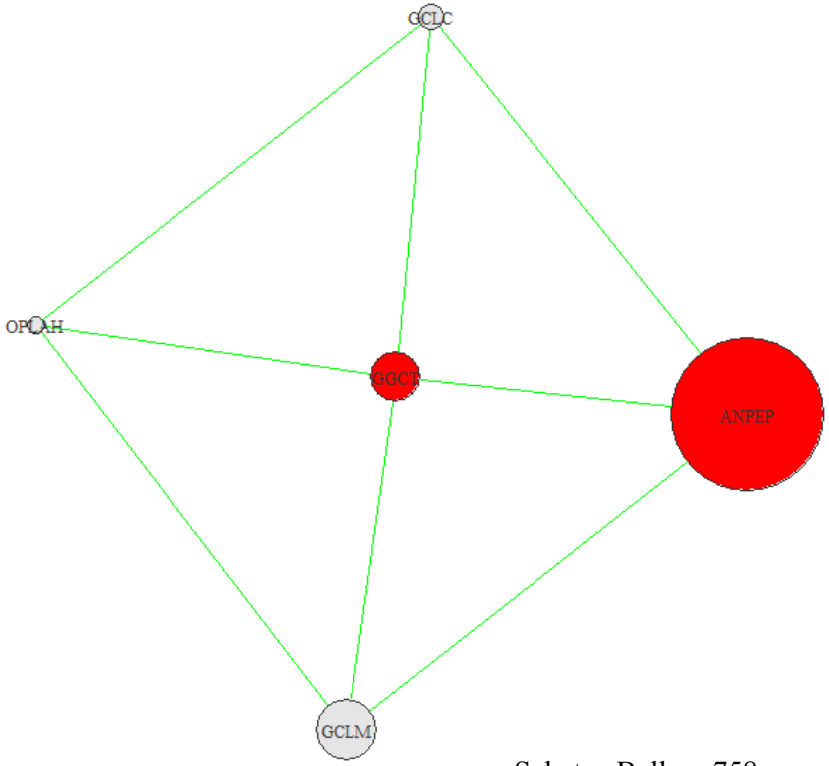
Sebates-Bellver 452



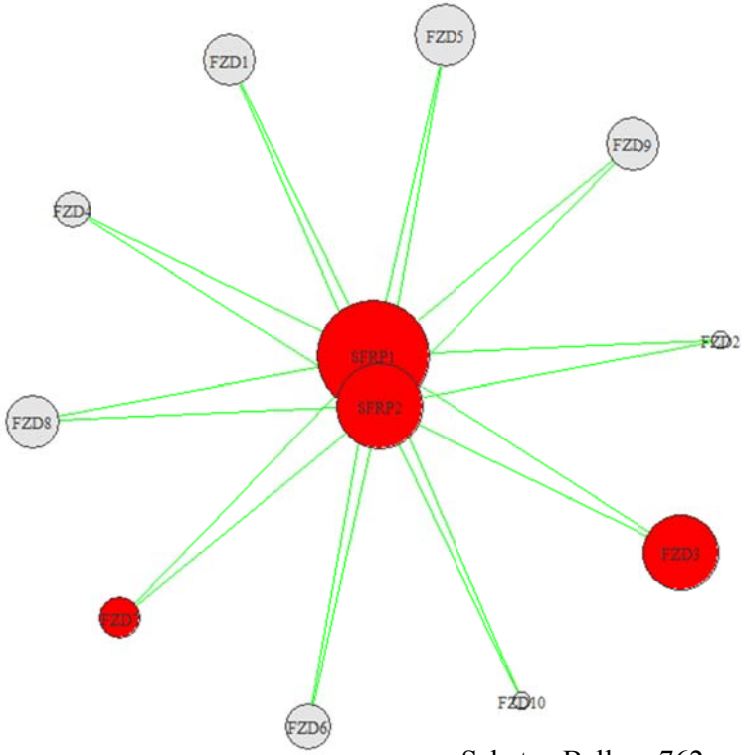
Sebates-Bellver 487



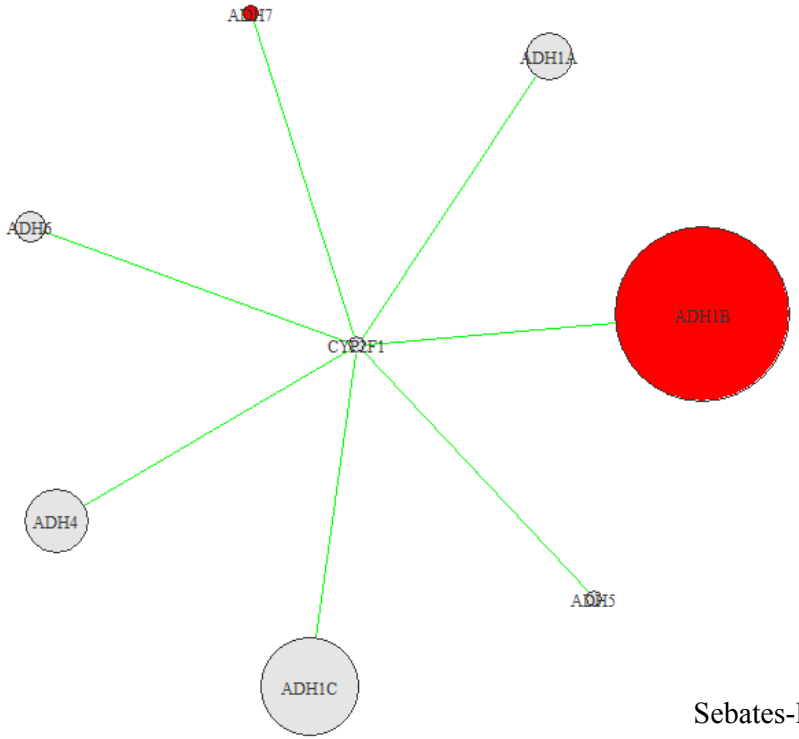
Sebates-Bellver 757



Sebates-Bellver 758

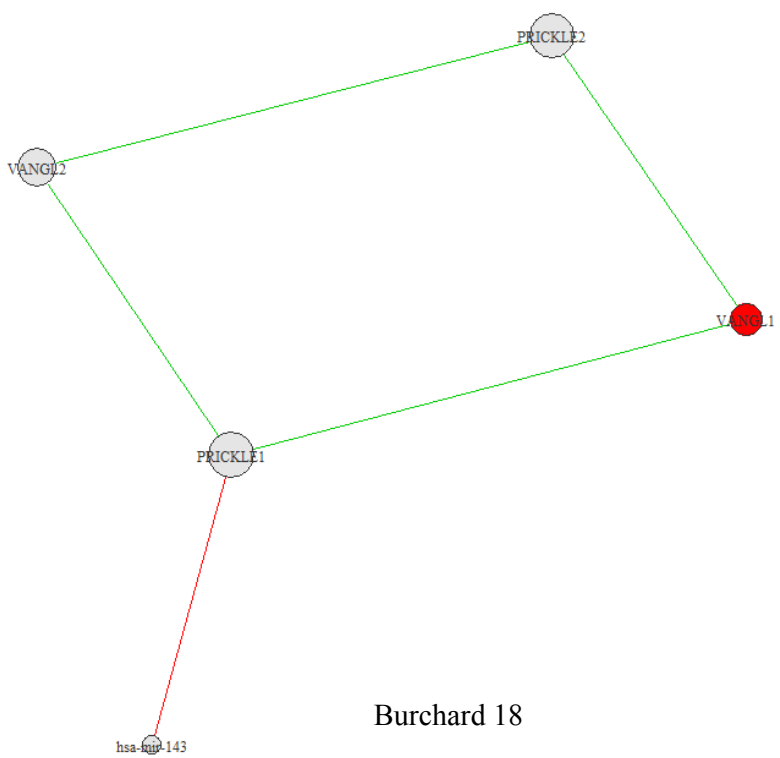


Sebates-Bellver 762

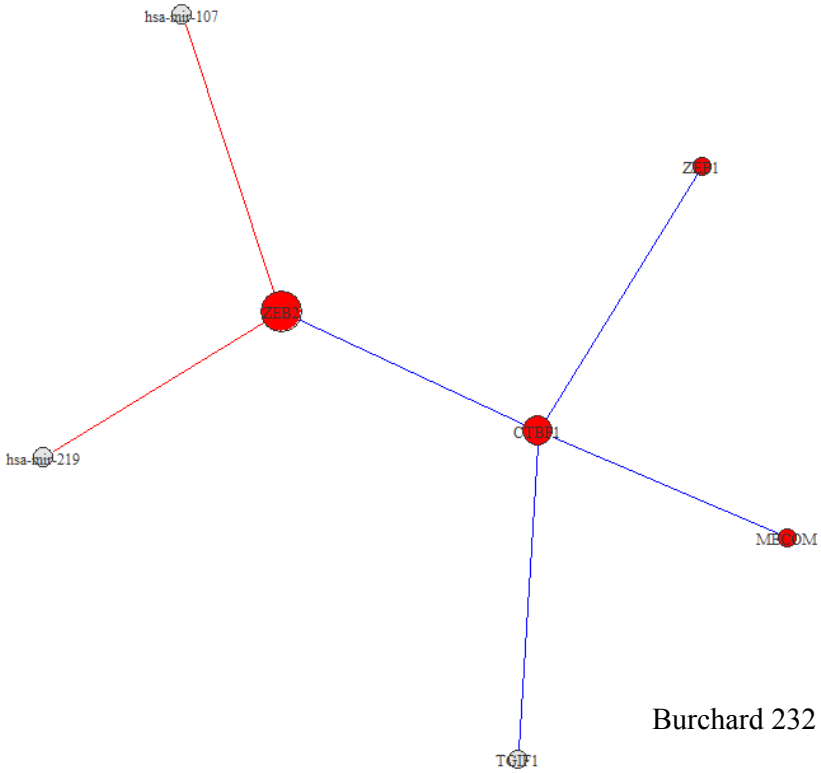


Sebates-Bellver 770

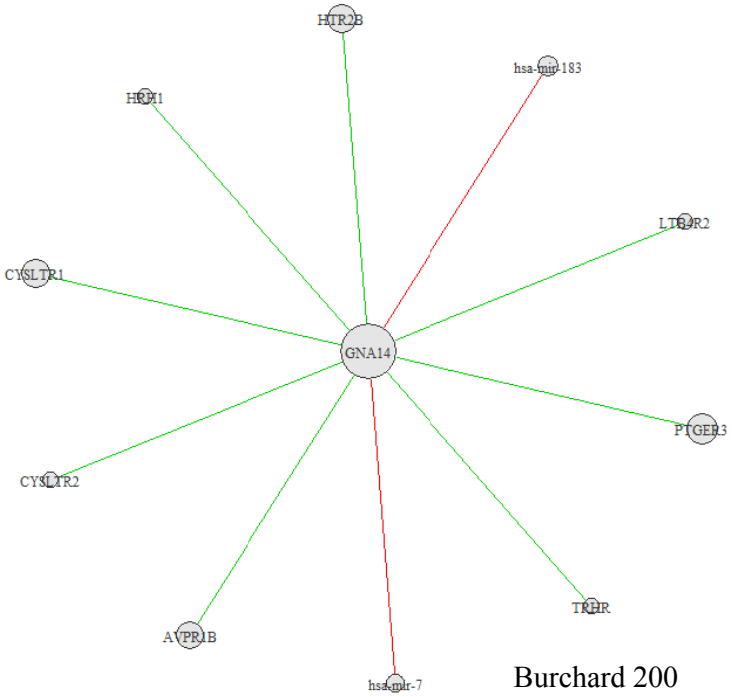
Significant Modules in Burchard 2010 Data



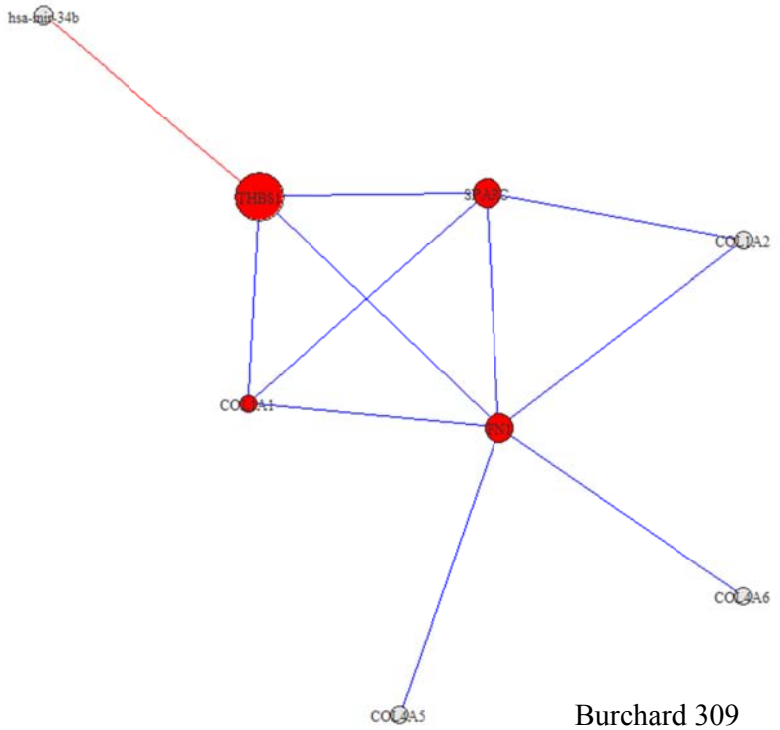
Burchard 18



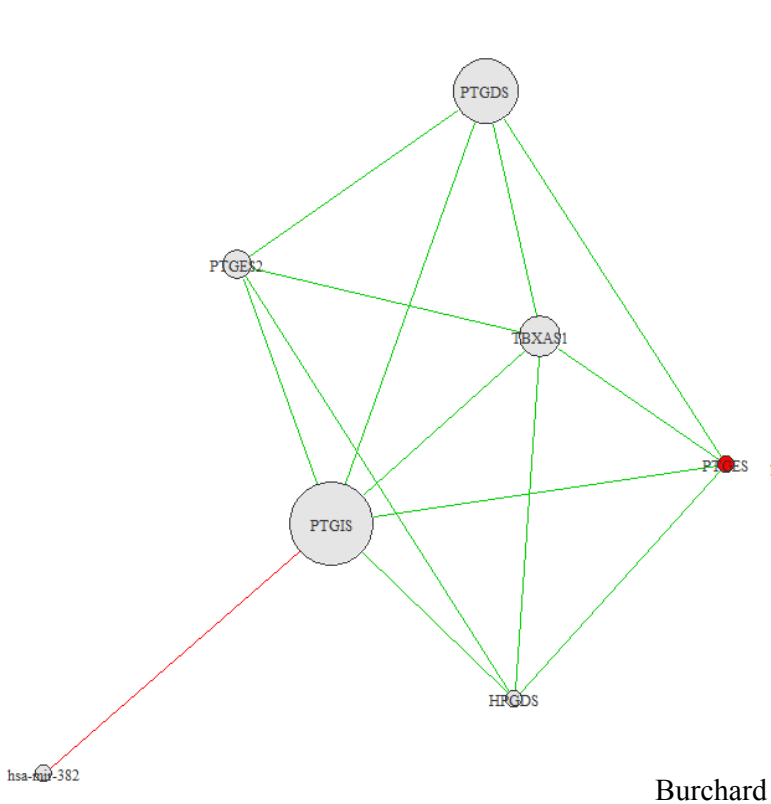
Burchard 232



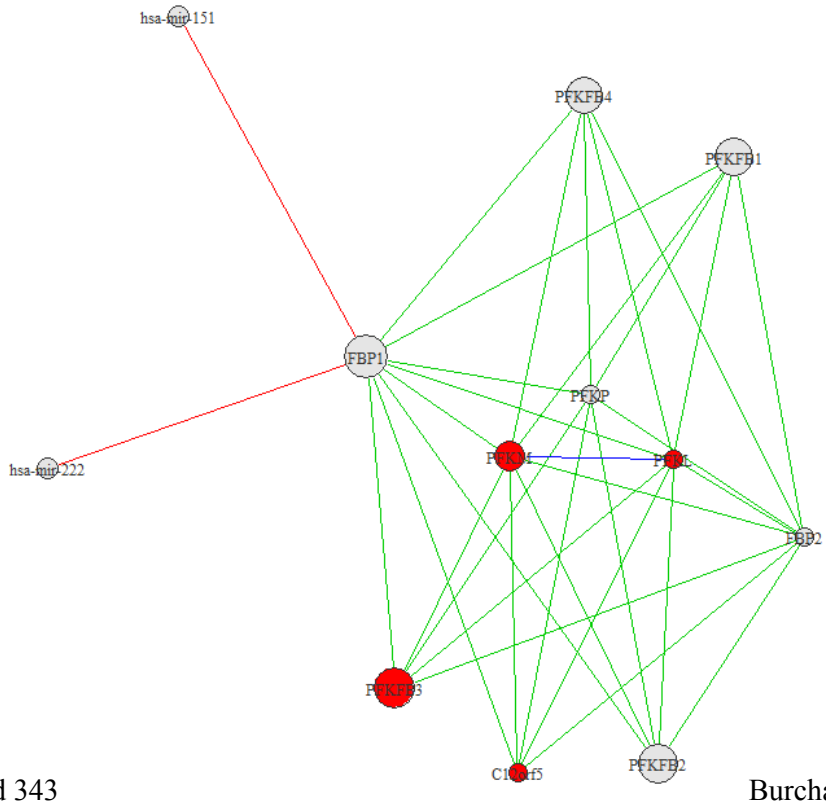
Burchard 200



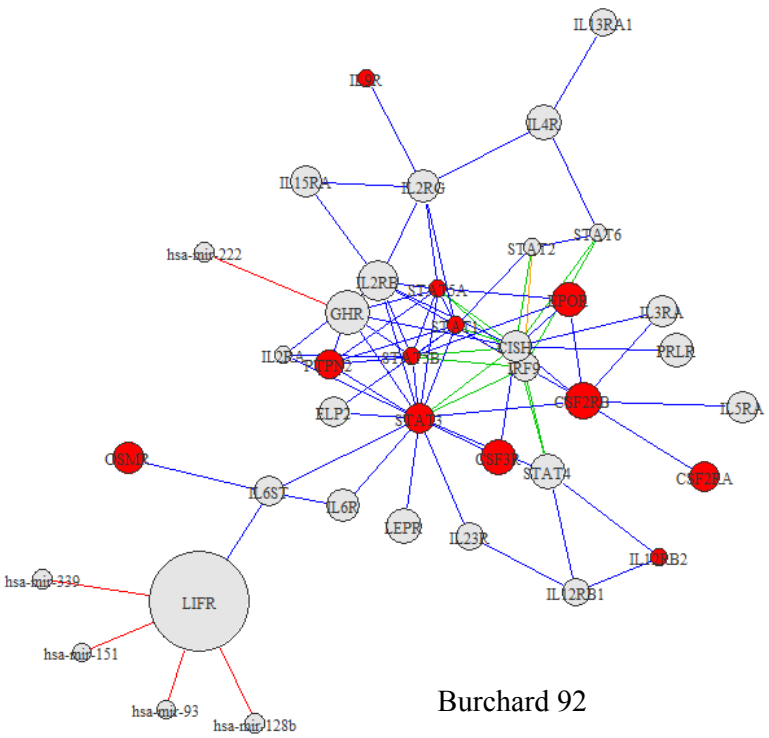
Burchard 309



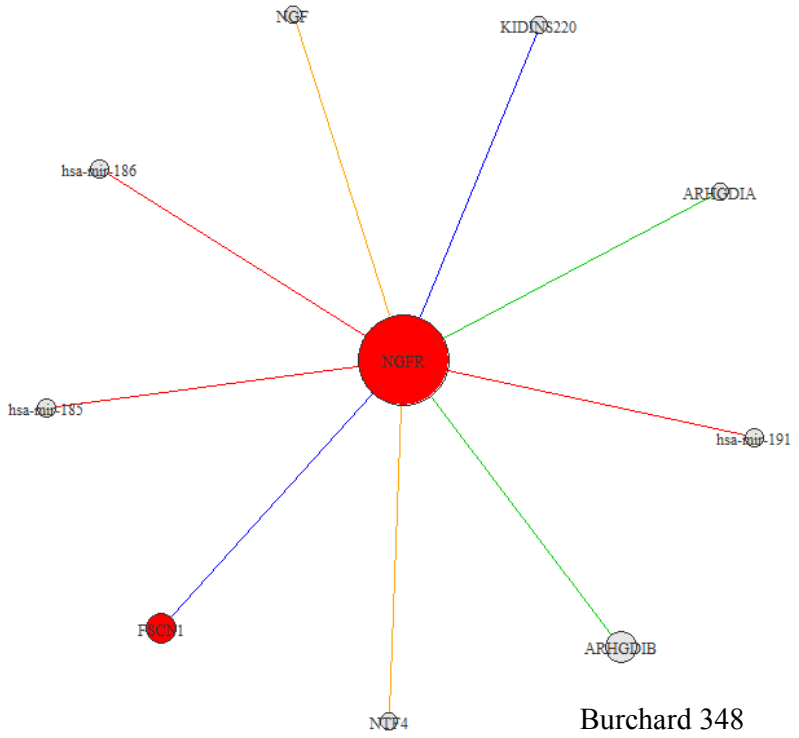
Burchard 343



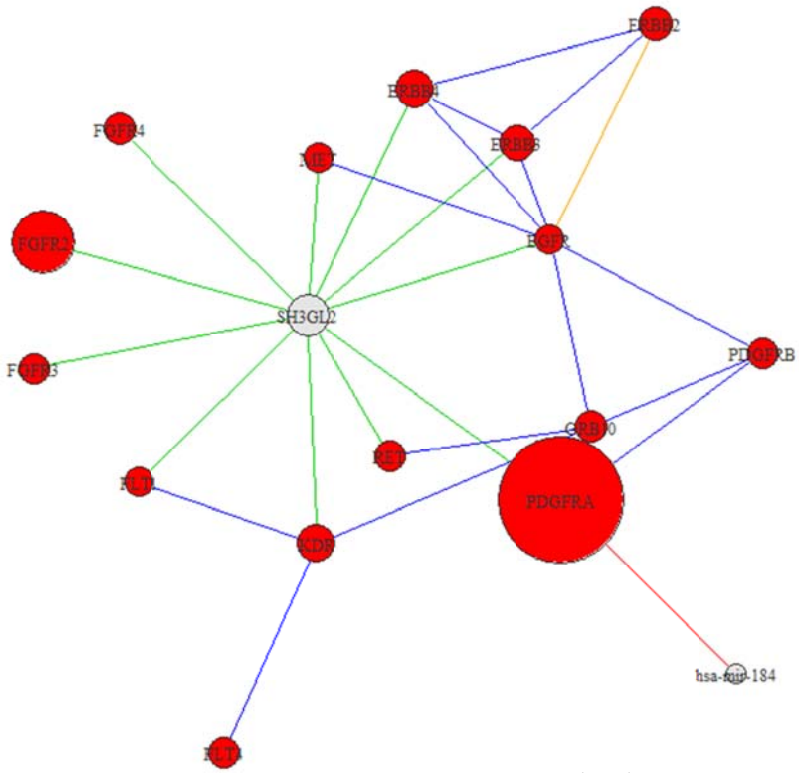
Burchard 318



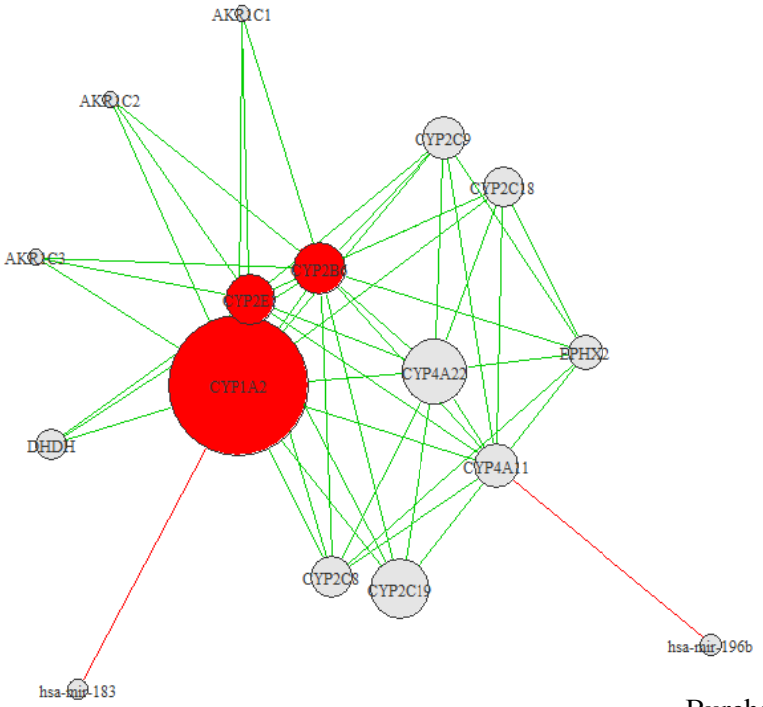
Burchard 92



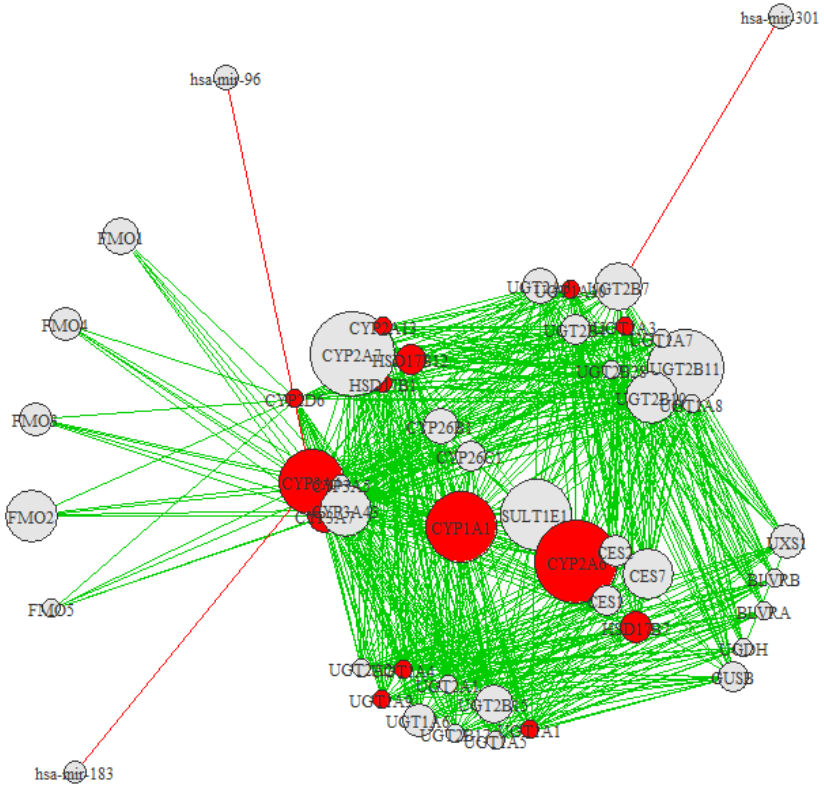
Burchard 348



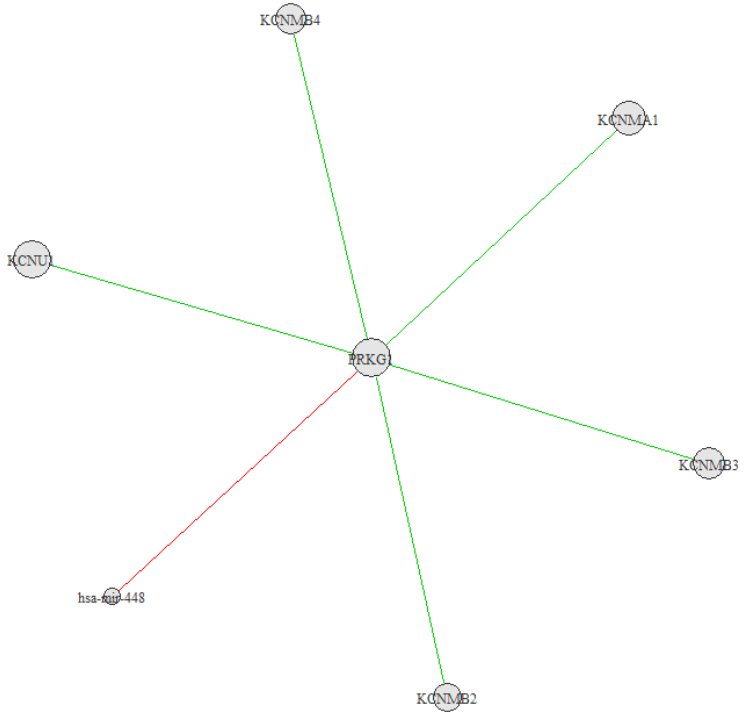
Burchard 567



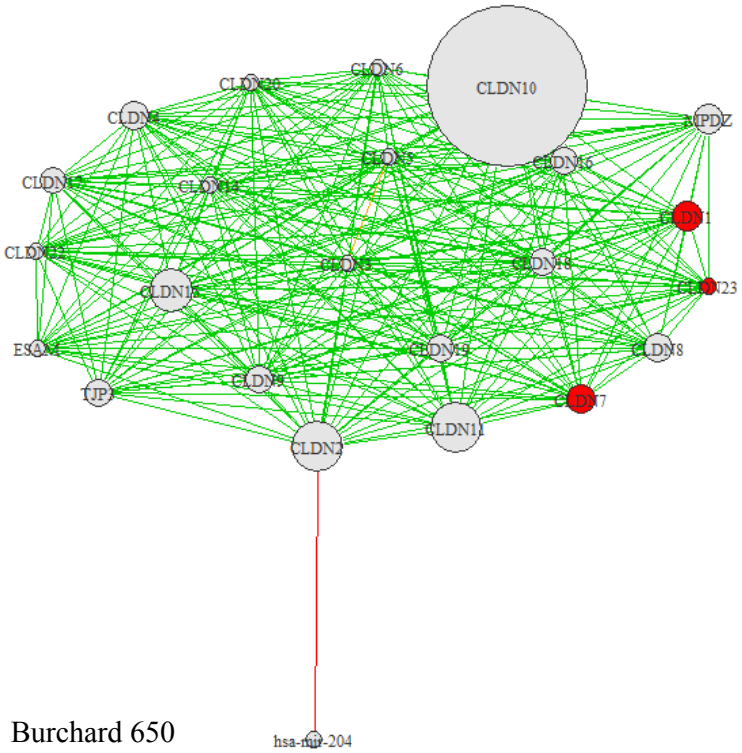
Burchard 583



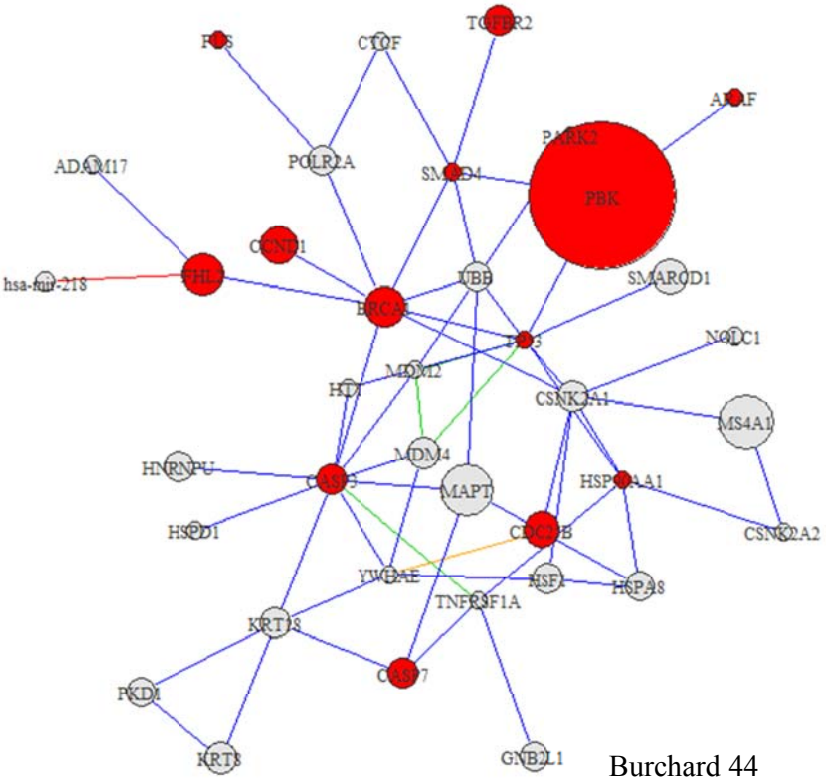
Burchard 186



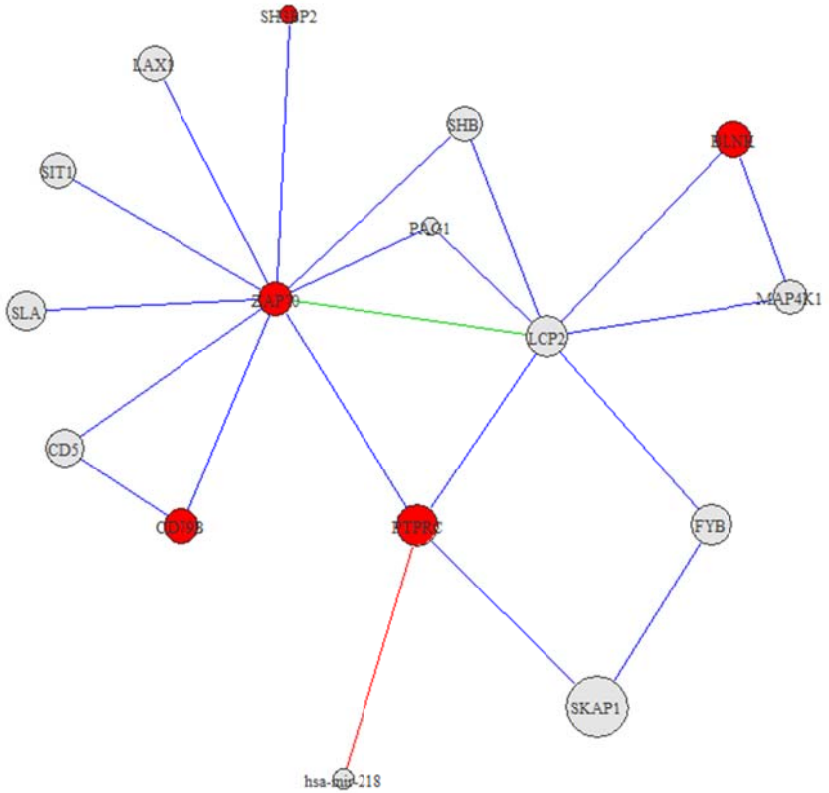
Burchard 647



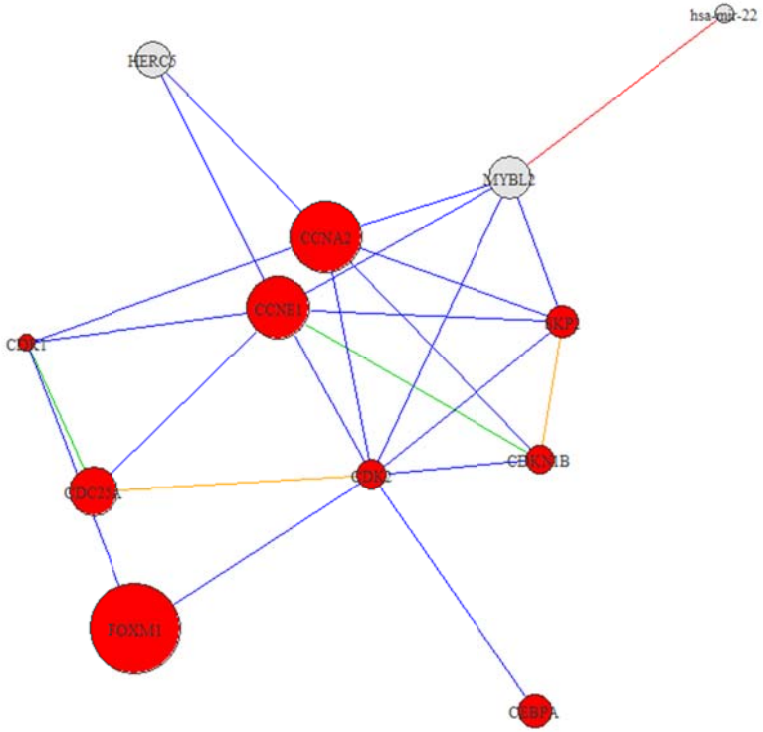
Burchard 650



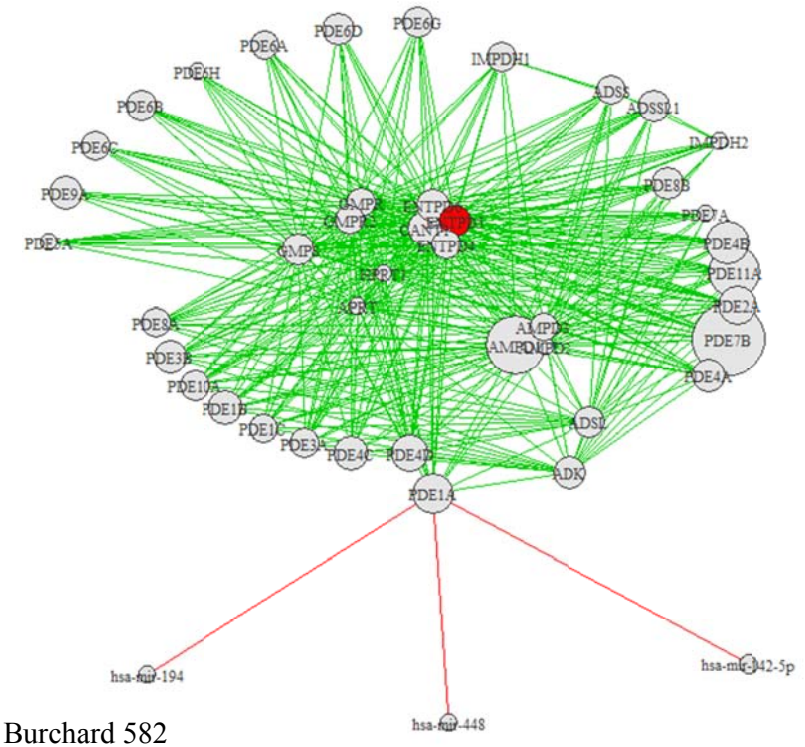
Burchard 44



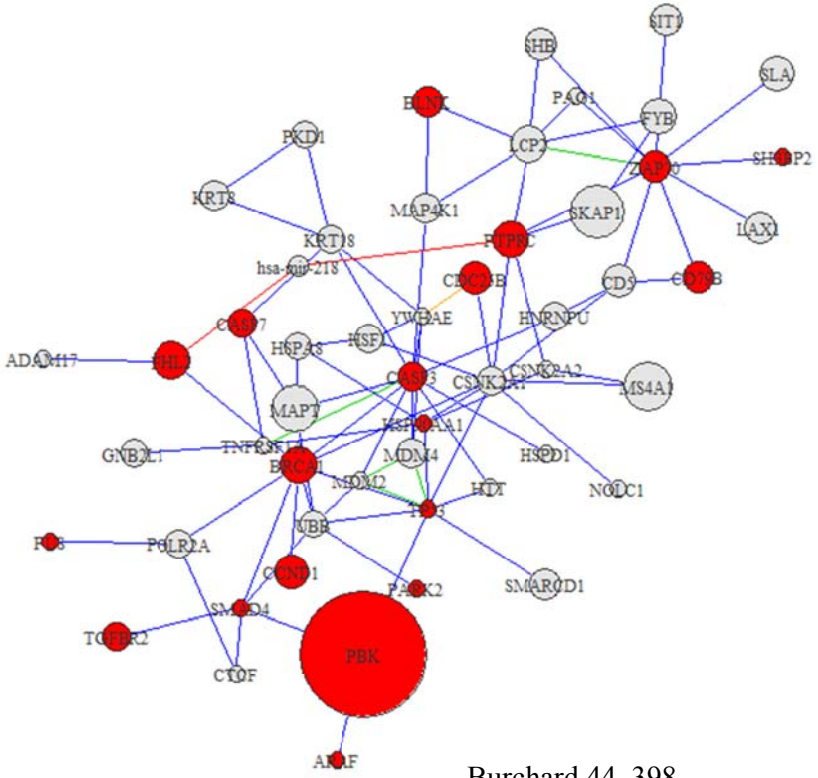
Burchard 398



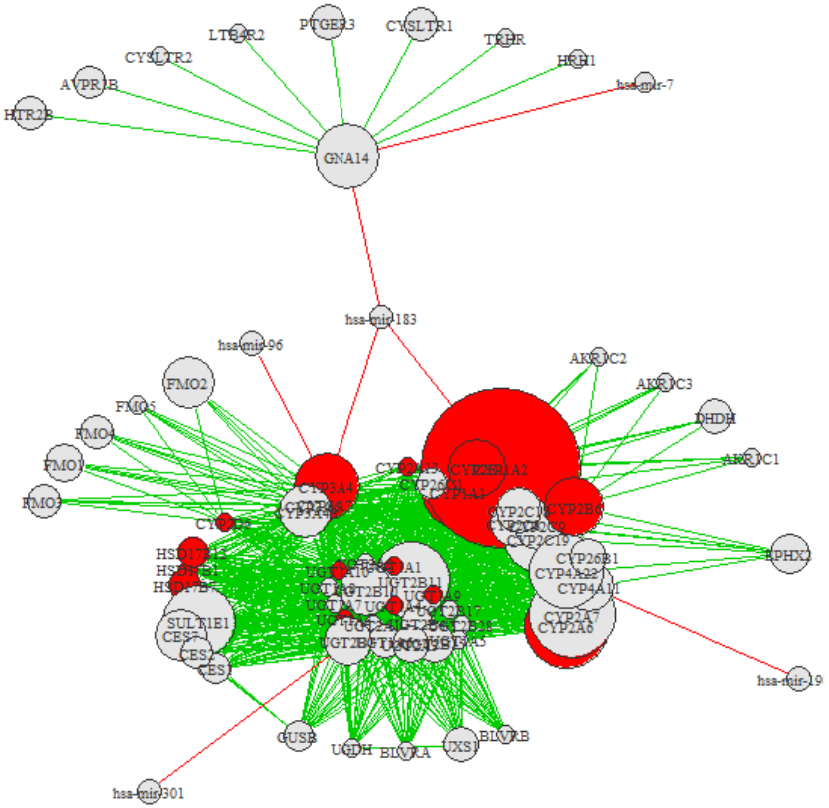
Burchard 389



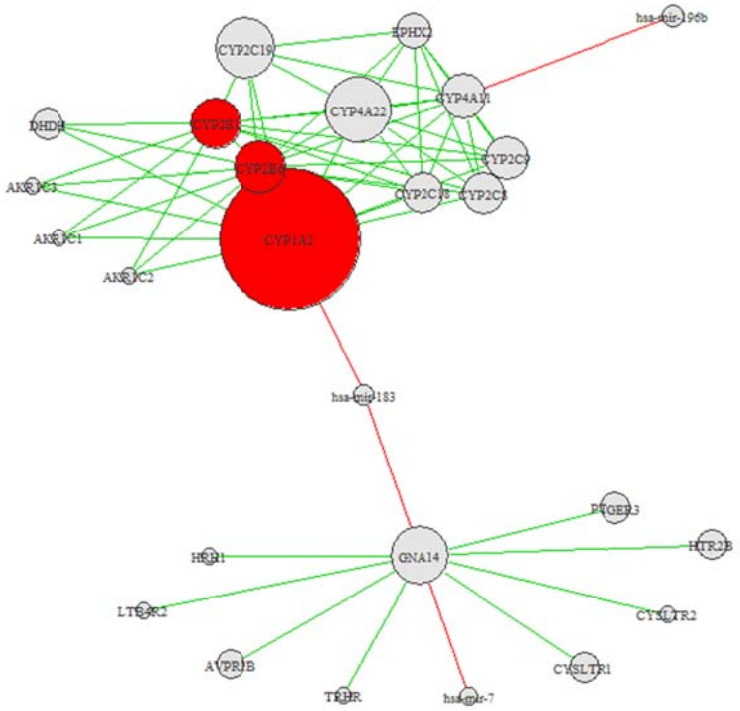
Burchard 582



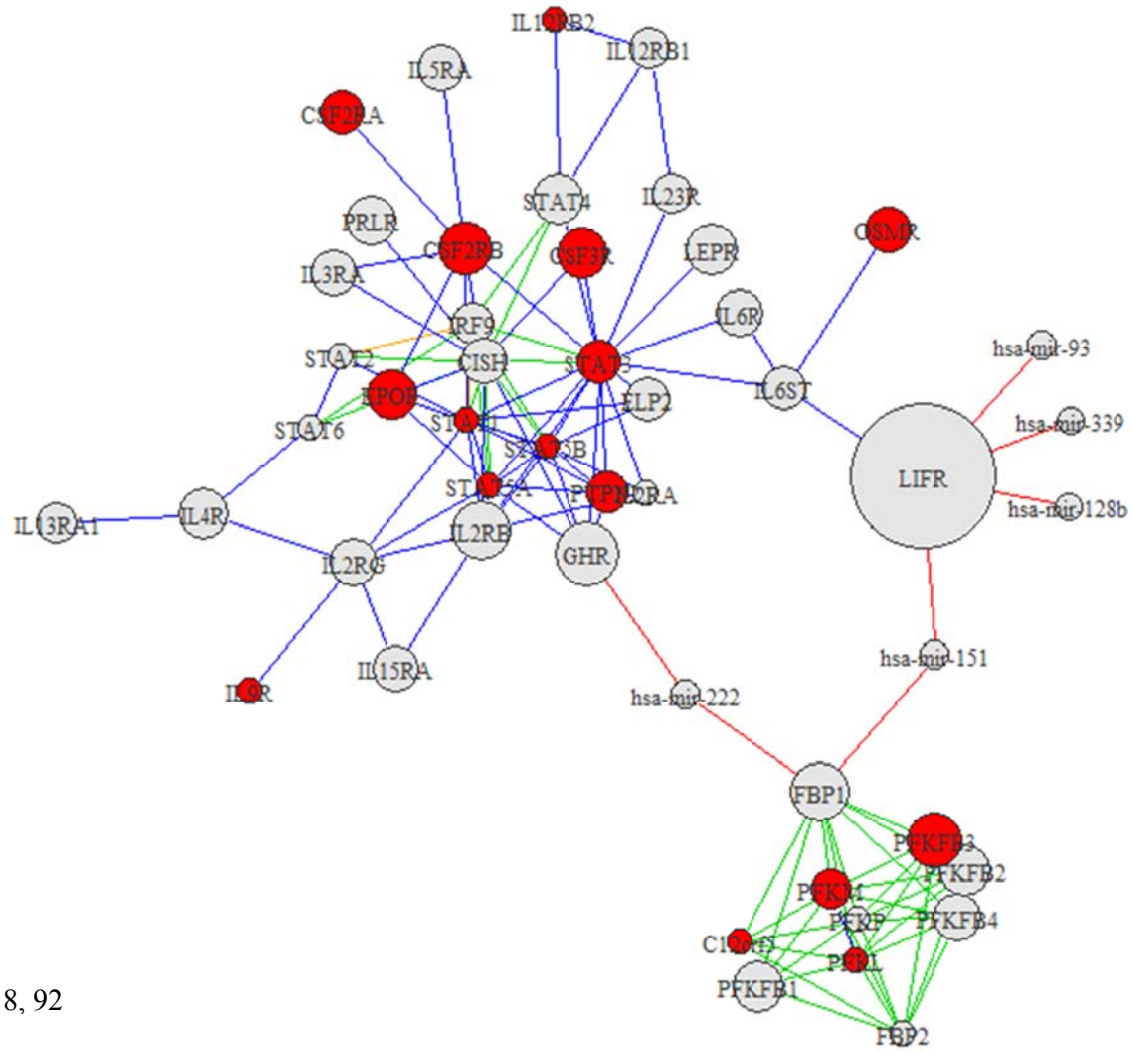
Burchard 44, 398



Burchard 583, 200, 186

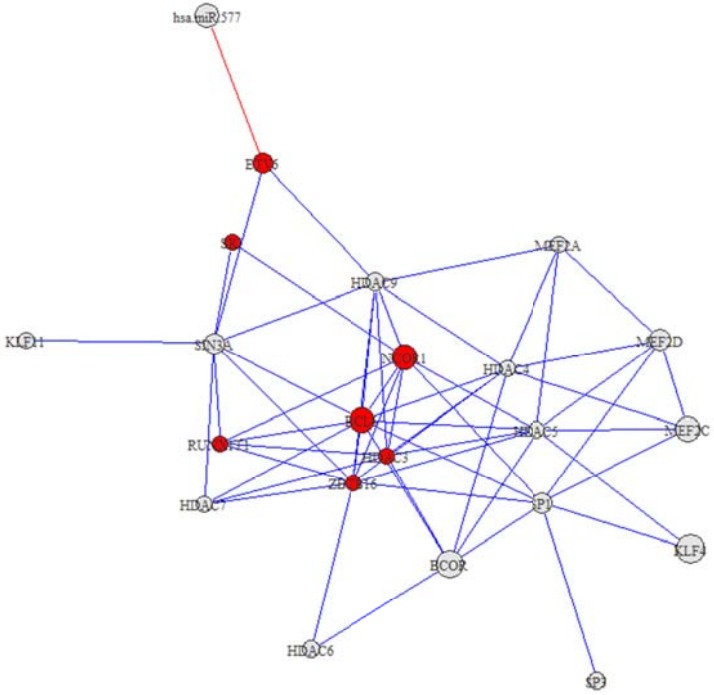


Burchard 583, 200

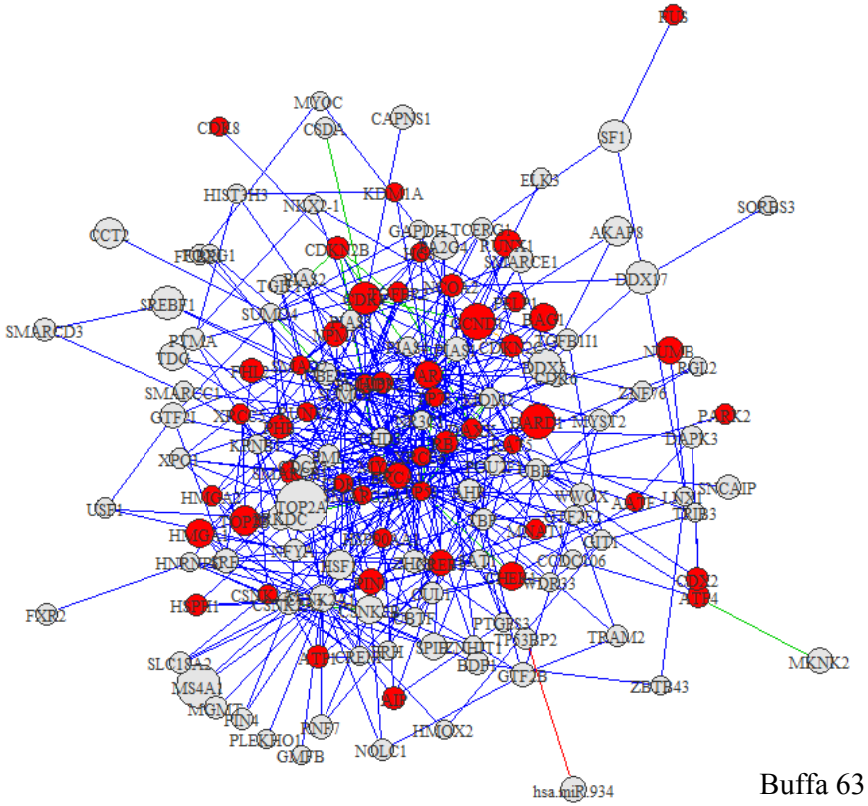


Burchard 318, 92

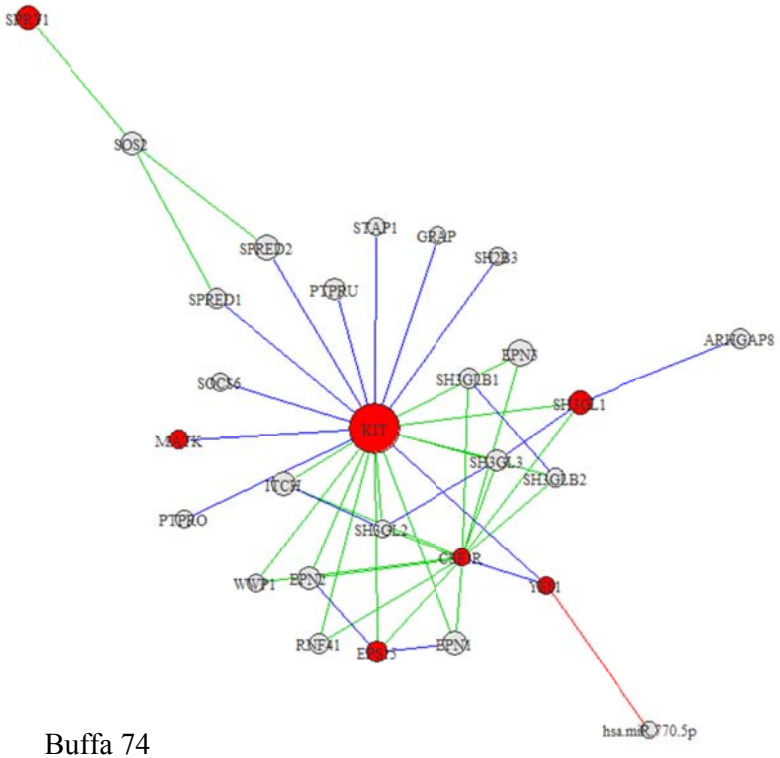
Significant Modules in Buffa 2011 Data



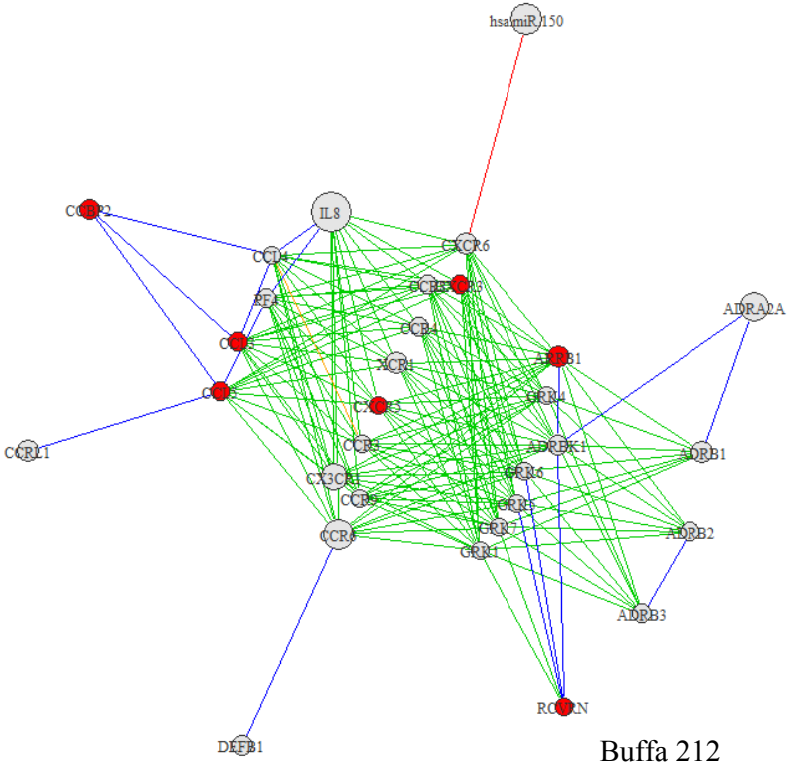
Buffa 22



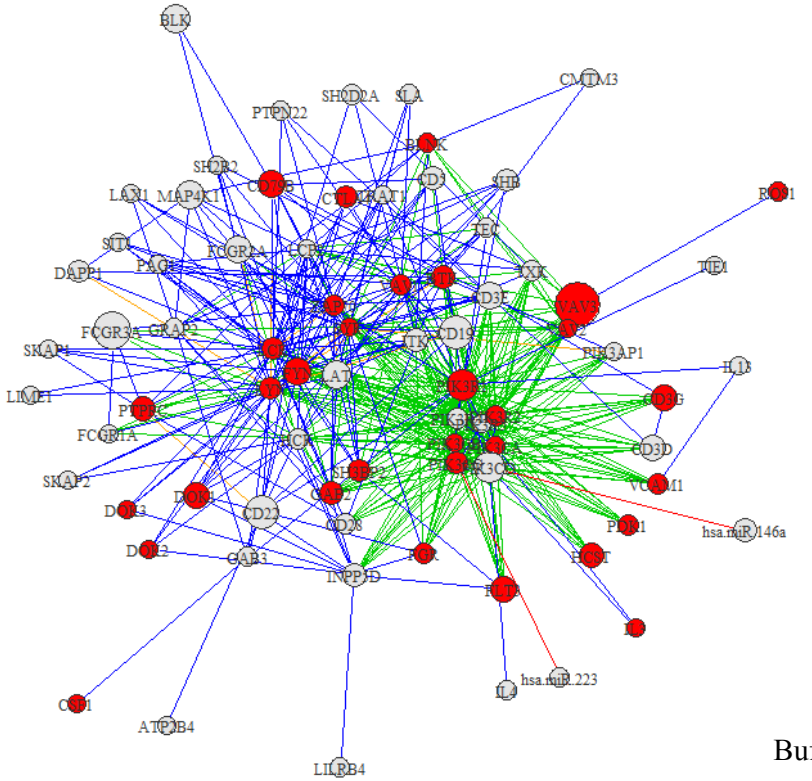
Buffa 63



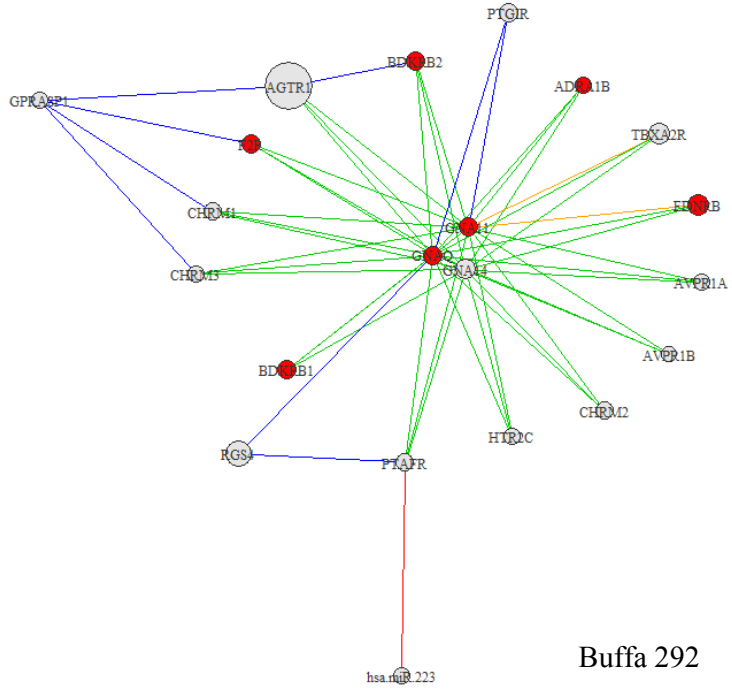
Buffa 74



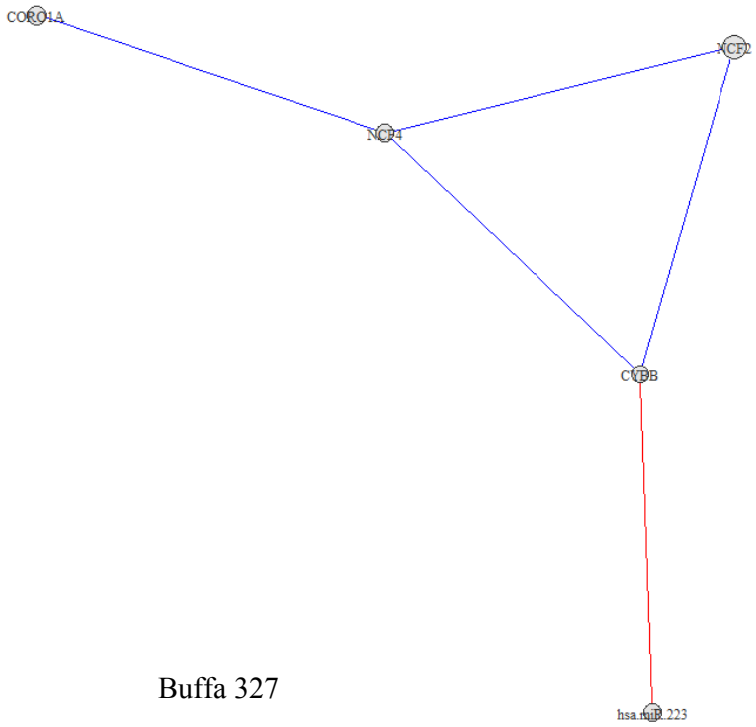
Buffa 212



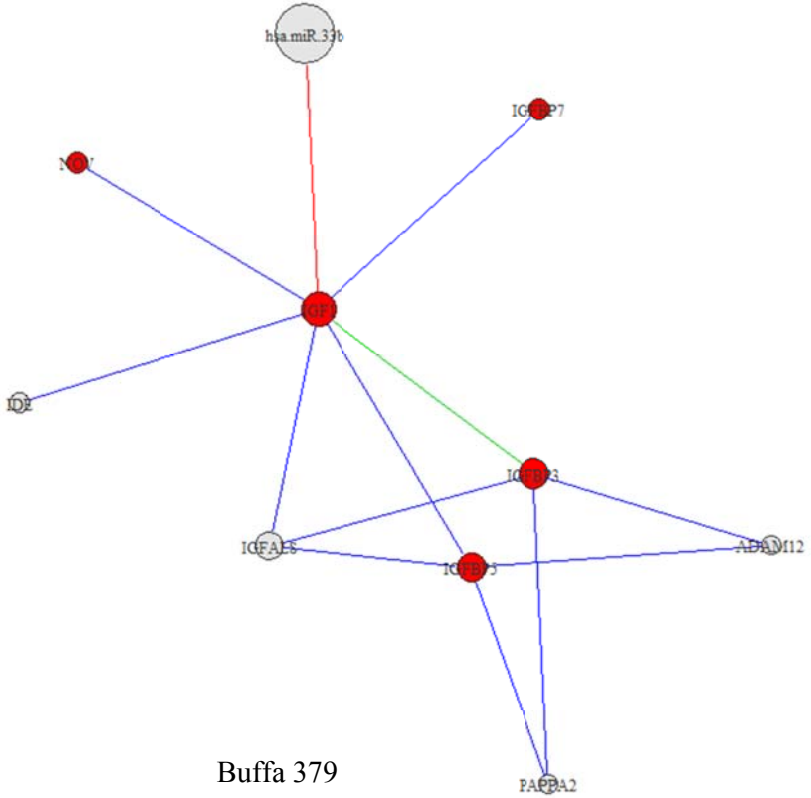
Buffa 269



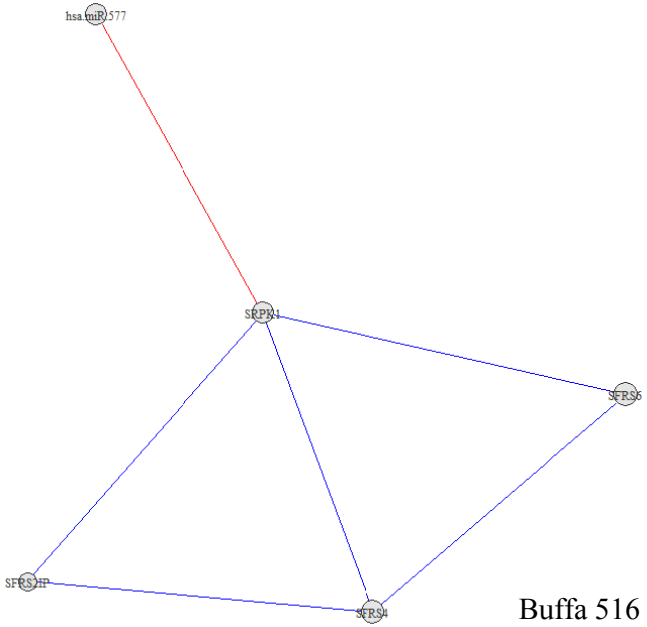
Buffa 292



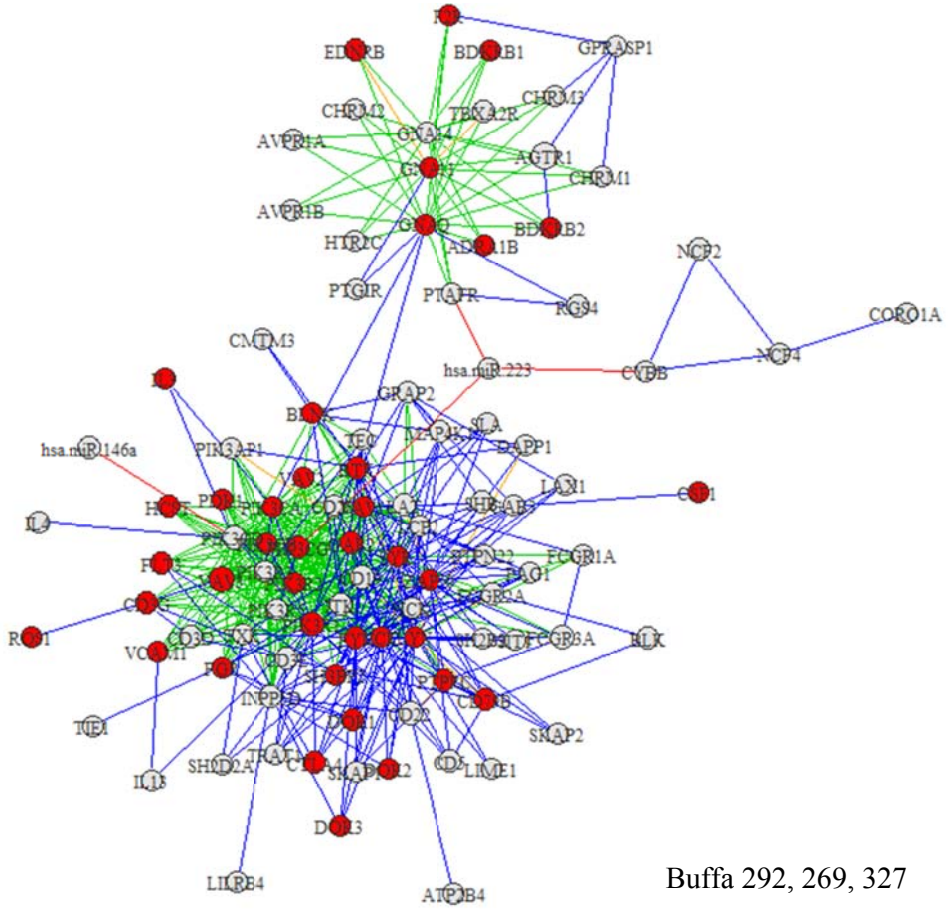
Buffa 327



Buffa 379



Buffa 516



Buffa 292, 269, 327

miRNA Evaluation Table

Evaluation Statistics	TOTAL	Cancer Enriched (CE)	MiRNA Enriched (ME)	CE + ME	Precision	Recall	Matthews Correlation Cofficient	TOTAL	Cancer Enriched (CE)	MiRNA Enriched (ME)	CE + ME	Precision	Recall	Matthews Correlation Cofficient
Results Object	**Unfiltered**							**Filtered**						
EA OM DS1	223	59	25	13	0.5200	0.2203	0.2716	223	59	3	2	0.6667	0.0339	0.1088
EA OM DS2	276	65	20	7	0.3500	0.1077	0.1109	276	65	5	2	0.4000	0.0308	0.0568
Net OM DS1	225	62	20	12	0.6000	0.1935	0.2705	222	61	10	6	0.6000	0.0984	0.1727
Net OM DS2	307	74	11	7	0.6364	0.0946	0.1906	331	68	3	1	0.3333	0.0147	0.0319
FW OM DS1	168	44	18	8	0.4444	0.1818	0.2044	222	61	10	6	0.6000	0.0984	0.1727
FW OM DS2	335	70	12	2	0.1667	0.0286	-0.0060	332	71	5	3	0.6000	0.0423	0.1202
Results Object														
EA B3 DS1	223	61	30	13	0.4333	0.2131	0.2264	223	61	15	10	0.6667	0.1639	0.2657
EA B3 DS2	276	67	23	9	0.3913	0.1343	0.1475	276	67	12	5	0.4167	0.0746	0.1018
Net B3 DS1	228	63	14	4	0.2857	0.0635	0.0311	226	64	12	7	0.5833	0.1094	0.1757
Net B3 DS2	327	67	12	6	0.5000	0.0896	0.1571	329	68	5	5	1.0000	0.0735	0.2505
FW B3 DS1	221	58	12	4	0.3333	0.0690	0.0597	224	60	12	4	0.3333	0.0667	0.0556
FW B3 DS2	332	66	20	10	0.5000	0.1515	0.2240	313	75	5	5	1.0000	0.0667	0.2325
Results Object														
EA B5 DS1	223	59	37	12	0.3243	0.2034	0.1879	223	61	19	11	0.5789	0.1803	0.2497
EA B5 DS2	276	67	30	10	0.3333	0.1493	0.1407	276	67	16	6	0.3750	0.0896	0.1007
Net B5 DS1	234	66	34	14	0.4118	0.2121	0.2134	224	63	17	10	0.5882	0.1587	0.2281
Net B5 DS2	338	68	25	13	0.5200	0.1912	0.2713	327	65	9	7	0.7778	0.1077	0.2577
FW B5 DS1	245	57	23	8	0.3478	0.1404	0.1419	234	62	19	10	0.5263	0.1613	0.2144
FW B5 DS2	349	69	38	12	0.3158	0.1739	0.1821	334	66	12	8	0.6667	0.1212	0.2451
Results Object														
EA All DS1	223	61	100	41	0.4100	0.6721	1.2655	223	61	97	41	0.4227	0.6721	1.2291
EA All DS2	276	67	59	20	0.3390	0.2985	0.3295	276	67	45	15	0.3333	0.2239	0.2249
Net All DS1	495	8	83	4	0.0482	0.5000	0.9607	272	70	130	47	0.3615	0.6714	1.4077
Net All DS2	699	50	70	9	0.1286	0.1800	0.2049	342	65	47	21	0.4468	0.3231	0.3967
FW All DS1	37	8	3	1	0.3333	0.1250	0.1298	83	27	28	12	0.4286	0.4444	0.5489
FW All DS2	351	67	80	27	0.3375	0.4030	0.5089	324	70	42	19	0.4524	0.2714	0.3257

VITA

Deanna Petrochilos earned a bachelor's degree in Anthropology and Organismal Biology from Yale Univeristy where she studied the comparative anatomy of ancient fossils and evolutionary biology. In the interim between college and graduate school, she worked for ten years in the biotechnology and software industries where she developed her interest in bioinformatics. She has a Master's degree in Genetic Epidemiology from the Harvard School of Public Health and a Doctorate of Philosophy in Biomedical and Health Informatics from University of Washington. She has been involved in bioinformatics and public health research at Incyte Pharmaceuticals, the Joint Genome Institute, Genentech, and the Fred Hutchinson Cancer Research Center. Her doctoral studies and current interests are in the application of novel methods to study genomic interactions involved in complex disease, specifically in analysis of large-scale cancer data.