2014

A module based approach for identifying driver genes and expanding pathways from integrated biological networks

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

A MODULE BASED APPROACH FOR IDENTIFYING DRIVER GENES AND EXPANDING PATHWAYS FROM INTEGRATED BIOLOGICAL NETWORKS

by

CHIA-LING HUANG

B.S., National Yang-Ming University, 2003
M.S., National Yang-Ming University, 2005
M.S., Boston University, 2009

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2014

Approved by

First Reader

Charles DeLisi, PhD
Arthur G.B. Metcalf Professor of Science and Engineering

Second Reader

Zhenjun Hu, PhD
Research Associate Professor of Bioinformatics

ACKNOWLEDGEMENTS

Foremost, I would like to thank my advisor, Dr. Charles DeLisi, for the continuous support of my Ph.D. study and research. His excellent guidance, caring and patience provide me an excellent atmosphere for doing research. The scientific freedom he gave me allowed me to explore the adventures in science. He leads me to understand the fun of scientific research. The most important thing I learned from him is one should enjoy the thing he/she is doing.

Special thanks to Dr. Daniel Ziemek and Dr. Enoch Huang for giving me a chance to do a one year internship at Pfizer Inc. I learned how to work and communicate with people from different departments and had a great industrial working experience. I would like to specially thank Dr. Ziemek for serving on my doctorial committee after I finished my internship. I am very appreciative for Dr. Ziemek's kindness, patience and encouragement. I am very grateful to my doctorial committee, Dr. Mark Kon, Dr. Zhenjun Hu and Dr. Gabriela Alexe for their great advice and continuous support.

I want thank the current and previous members in DeLisi Lab for their support. I particularly want to thank Yi-Chien Chang and Dr. Michael Seiler for their great help and inspiring discussions.

At the end, I would like to express my gratitude to my parents for their support and understanding. I also appreciate everyone I met in Boston and Taiwan. Their support and company made my life in graduate school wonderful.

A MODULE BASED APPROACH FOR IDENTIFYING DRIVER GENES AND EXPANDING PATHWAYS FROM INTEGRATED BIOLOGICAL NETWORKS

(Order No.)

CHIA-LING HUANG

Boston University Graduate School of Arts and Sciences

and College of Engineering, 2014

Major Professor: Charles DeLisi, Arthur G.B. Metcalf Professor of Science and Engineering

ABSTRACT

Each gene or protein has its own function which, when combined with others, allows the group to perform more complex behaviors, *e.g.* carry out a particular cellular task (functional module) or affect a particular disease phenotype (disease module). One of the major challenges in systems biology is to reveal the roles of genes or proteins in functional modules or disease modules.

In the first part of the dissertation, I present a data-driven method, Correlation Set Analysis (CSA), for comprehensively detecting active regulators in disease populations by integrating co-expression analysis and specific types of literature-derived causal relationships. Instead of investigating the co-expression level between regulators and their targets, I focus on coherence of regulatees of a regulator, *e.g.* downstream targets of a transcription factor. Using simulated datasets I show that my method can reach high true positive rate and true negative rate (>80%) even the regulatory relationships is weak (only 20% of regulatees are co-expressed). Using three separate real biological datasets

I was able to recover well-known and as- yet undescribed, active regulators for each disease population.

In the second part of the dissertation, I develop and apply a new computational algorithm for detecting modules of functionally related genes that are likely to drive malignant transformation. The algorithm takes as input the identity and locations of a small number of known oncogenes (a seed set) on a human genome functional linkage network (FLN). It then searches for a boundary surrounding a gene set encompassing the seed, such that the magnitude of the difference in linkage weights between interior-interior gene pairs, and interior-exterior gene pairs is maximized. Starting with small seed sets for breast and ovarian cancer, I successfully identify known and novel drivers in both cancer types.

In the third part of the dissertation, I propose a module based approach for expanding manually curated functional modules. I use the KEGG pathway database as an example and the results show that my approach can successfully suggest both validated pathway members (genes that are assigned to a particular pathway by other manually curated pathway databases) and novel candidate pathway genes.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

COSMIC	The Catalogue Of Somatic Mutations In Cancer
CSA	Correlation Set Analysis
DLBCL	Diffuse large B-cell lymphoma
FDR	False discovery rate
FLN	Functional linkage network
FLNP	Functional linkage network partitioning
GO	Gene ontology
GWAS	Genome-wide association studies
KEGG	Kyoto Encyclopedia of Genes and Genomes
OMIM	Online Mendelian Inheritance in Man
PPI	Protein-protein interaction
ROC	Receiver-operator characteristic
SNP	Single nucleotide polymorphisms
TCGA	The Cancer Genome Atlas

Chapter 1. Introduction

1.1 What constitutes a module in a biological network?

The interaction between genes, transcripts, proteins, and external stimuli is generally complex. One gene can be regulated by one or more proteins, and conversely, one protein can regulate several genes [1, 2]. The interaction between genes and proteins can be presented by an interaction network, such as a protein-protein interaction (PPI) network, a transcriptional regulation network, a metabolic network, or functional linkage network (FLN). Most real-world networks, including the biological networks we mentioned above, have some basic properties in their structure which cannot be found in random networks. First is the special behavior of degree distribution, which is the fraction of nodes having n links (n=0, 1, 2…). In a network that is generated by a stochastic process, the majority of nodes have similar number of links and only a small percentage of nodes have extremely large or small number of links. Therefore the degree distribution of a random network follows the binomial distribution or Poisson distribution in the limit of large graph size [3]. Real networks, however, have a power law distribution of degrees, which means they are scale free networks [4]. In a scale free network, most nodes only have a few links and relatively small number of nodes are highly connected (hubs). These hubs usually serve special roles in a network.

The second property is the modularity structure of a network. Modules are highly interconnected local regions in a network. In biological networks, modules could be groups of genes or proteins that form a protein complex, carry out a similar function, or contribute to a certain phenotype. In 2011, Barabasi *et al.* concluded that there were three basic modularity concepts in a biological network: a topological module, a functional module, and a disease module (Figure 1-1) [5]. Topological modules are locally dense regions on a network. The nodes in this module have a higher tendency to connect with nodes within the module than nodes outside the module. Functional modules represent groups of nodes that are involved in the similar or related functions. A disease module represents a group of nodes (genes or proteins) that contribute to the same disease. Typically, genes in a disease module share mutations, structure variations, or expression abnormalities linked to a particular disease phenotype. These modularity concepts lead to different methods for studying modules in a biological network. The first one is to identify a module for a particular disease phenotype. In this thesis, we focus on identifying driver modules in cancer. The second method is to predict new functional modules or suggest new members to an existing functional module.

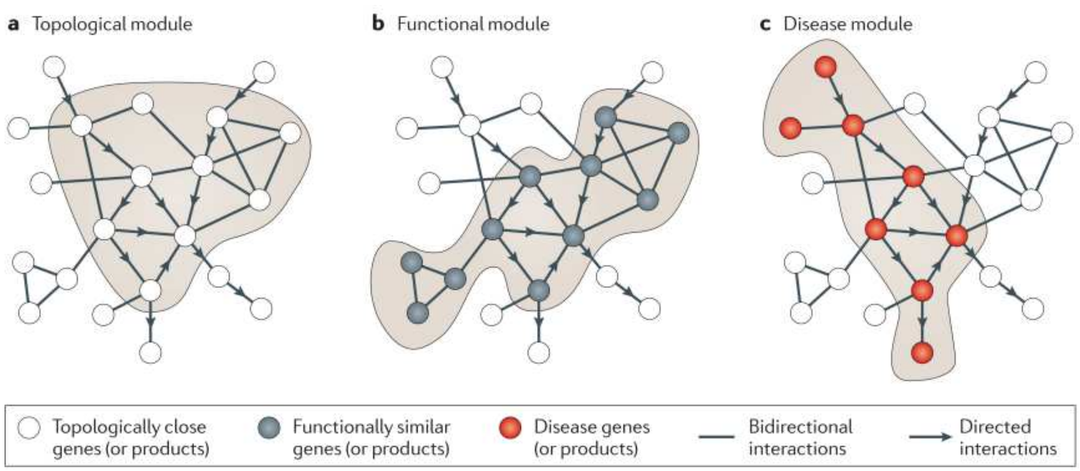

Figure 1-1 Modularity concepts in biological networks. (a) Topological modules are locally dense regions on a network. The nodes in this module have a higher tendency to connect with nodes within the module than nodes outside the module. (b) Functional modules are groups of nodes with related functions. The nodes in a functional module tend to interact with nodes in the same functional module than nodes outside the module. (c) A disease module is a group of nodes (genes or proteins) that contribute to the same disease. The mutations, structurural variations, or expression abnormalities of those genes/proteins are related to a particular disease. Figure adapted from Barabasi *et al.*, *Nature Review*, 2011.

1.2 Modules in cancer

Cancer is a genetic, epigenetic and phenotypic heterogeneous disease. Genetic and epigenetic heterogeneity means cancer may arise from a multitude of point mutations, copy number variations or epigenetic modifications. For example, Lafrate *et al.* found that malignant brain tumors contain amplifications of three different receptor tyrosine kinases (EGFR, MER, PDGFRA) in different tumor cells [6]. Breast cancer is a well-known example of phenotypic heterogeneous disease that has diverse morphology, expression profiles and immunohistochemical patterns [7, 8]. In the past few decades, cancer research has made remarkable advances. Hanahan and Weinberg first proposed six hallmarks of cancer in 2000, which includes: self-sufficiency in growth signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis (Figure 1-2a) [9]. In 2011, they added two emerging hallmarks – reprogramming of energy metabolism and evading immune destruction (Figure 1-2b) [10].

(a) (b)

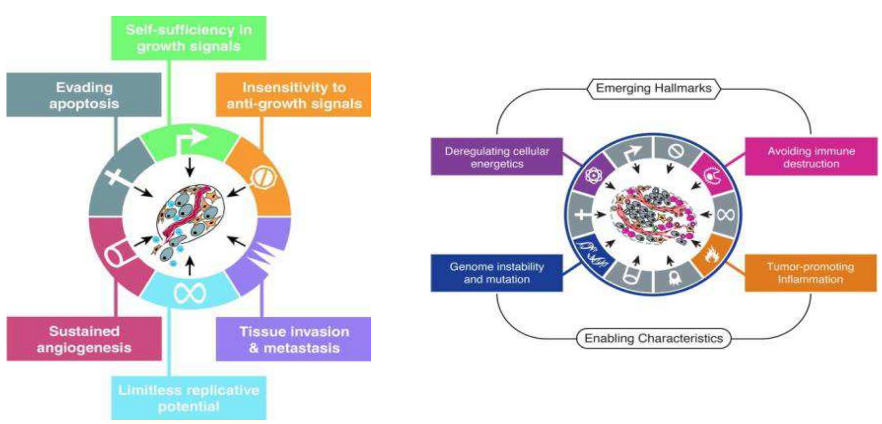

Figure 1-2. Hallmarks of cancer. Figure adapted from Hanahan and Weinberg, Cell, 2000 and 2011.

A central goal in cancer research is to distinguish mutations that drive the initiation and development of a tumor (drivers) from the many other mutations that are unrelated to transformation (passengers). This problem is difficult because the unidentified driver mutations occur infrequently, and must be identified from within a much larger pool of mutations that have no bearing on cancer. Moreover, with the advent of very high throughput sequencing, the background pool is increasing rapidly. Until recently, researchers have only revealed ~140 drivers [11]. Since knowledge of cancer causing alterations is essential to the identification of therapeutic targets, solving the driver identification problem is a critical as well as a difficult challenge. Progress toward its solution will not only add substantially to our understanding of cancer cell biology, but should also greatly increase therapeutic options.

1.3 Current driver identification approaches

Driver identification has been approached in a number of ways. The most widely used approach attempts to identify genes mutated more frequently in cancer cells than background mutation rate [12–15]. Those methods assume that a mutation that drives cancer should confer a growth advantage to tumor cells. They perform well on detecting popular driver genes, but may neglect drivers with low mutation frequency. Another widely adapted idea is to identify genes having high rate of functional mutations, *i.e.* non-silent mutations [16–18]. The assumption underlying this approach is that non-silent mutations are more likely than silent mutations to be potential drivers because non-silent mutations usually have higher impact on protein functions. Still, many other approaches are developed to provide a solution for driver identification, including searching for genes with a regional clustering of mutations [19], or genes having mutations in specific functional residues, *e.g.* phosphorylation sites [20]. Above approaches, however, only consider abnormal mutations in each gene and ignore the interaction between genes.

1.4 Module based approaches for predicting drivers in cancer

Recent years, there are other computational methods, such as MEMo and PARADIGM, they move the direction from identifying individual drivers to pathways that are dysregulated in cancer [21, 22]. However, one criticism of pathway approaches is that they ignore all the genes that are not currently in manually curated pathways.

To address this problem, in chapter 2 we first demonstrate how to identify active regulators in diseases by integrating a manually curated interaction network with expression information. The network we used in chapter 2 consists of relationships that describe the flow of causality directed from a specific regulator to a specific regulatee and covers ~10,000 genes. We restrict regulatees at transcript level only as because the expression data we used is measuring the activities of regulatees at the transcript level. In contrast, regulators can include transcript and protein levels. The more detail definitions and examples of regulators and regulatees will be illustrated in chapter 2. We present a novel algorithm that takes a regulator and its downstream regulatees as a regulatory module and incorporates expression data to identify candidate active regulators in diseases. This is based on the hypothesis that in most cases an active regulator under a given condition should activate or inhibit a subset of its regulatees. Using three separate real biological datasets we were able to recover well known and as yet undescribed, active regulators for each disease population.

Next, we focus on identifying "drivers" in cancer by applying a module based approach to a weighted functional linkage network. In chapter 3, we present a multifaceted computational approach –very different from those currently in use -- that draws upon multiple data sources to help solve this problem. The algorithm is based on the idea that oncogenes clustered on an evidence weighted functional linkage network are embedded in a field of genes and pathways that are likely to be associated with cancer cell physiology. More specifically, we present a method for determining the boundary of the field, thereby obtaining a module that is functionally rich in drivers, and enabling the identification of new oncogene candidates.

1.5 Module based approach for expanding manually curated functional modules

Functional modules are routinely identified based on so-called Gene Set Analysis (GSA) methods and a number of statistical procedures have been proposed, *e.g.* DAVID, GSEA, PWEA, MIRTFnet [23–28]. Most of the current functional modules, especially pathways, used in those GSA methods are from manually curated sources [29–34]. These manually curated functional modules offer highly confident and clearly annotated functions of modules and elements in each module. Because of this, however, these annotations are limited to both incomplete current biological knowledge and the requirement of substantial manpower. Due to this, the manually curated functional modules are usually small and the coverage is low. Another issue plaguing manually curated functional modules is the lack of standard criteria when collecting and translating the data from literature. Therefore, the same pathway or ontology may have different elements in different databases. For example, the well-known citric acid (TCA) cycle has 19 genes in Reactome and 31 genes in KEGG, yet has only 8 genes in BioCarta. Seven of these genes are treated as members of TCA cycle in all three databases but many more genes are assigned to TCA cycle by only one or two databases. In the last chapter of this thesis, we demonstrate how a module based approach can improve our knowledge of current manually defined functional modules. We use the KEGG pathway database as an example and successfully suggest new members with strong evidence supports to KEGG pathways.

Chapter 2. Correlation Set Analysis: Detecting active regulators in disease populations using prior causal knowledge

2.1 Introduction

Fundamental functions of living cells are controlled by regulatory relations between genes and proteins. Most cell types respond to changes in their environment (*e.g.* drug treatments or disease causing mutations) by altering their transcriptional patterns. More than a decade ago, it became possible to measure snapshots of all transcript levels in a given tissue sample using microarray technology. Since then, advances in technology have multiplied and the cost of experiments has decreased significantly. As a consequence, cell lines, animal models as well as clinical subjects in drug trials or in the general population [35] have been characterized on a molecular level. One crucial problem in such studies is the detection of active key regulators; *i.e.* genes or proteins that causally affect expression of downstream genes or proteins in the study population.

The detection of regulators or regulatory networks from the primary data alone has been studied extensively. Network reconstruction can be approached by identifying correlations in expressed genes [36], using any number of methods including those based on information theory [37], Bayesian models [38, 39] and regression models [40]. However, all purely expression-based methods assume that the expression of regulators and targets are directly (anti-) correlated or at least not independent of each other. Smet *et al.* showed that such models of high correlation between regulator and target don't match the actual regulator-target relation captured in RegulonDB [41]. One of the possible explanations is that a regulator, even a transcription factor, itself is not necessarily co-expressed with its targets, especially when the regulatory relation of the regulator and the targets is complex [42] or regulation acts beyond the transcriptional level, *e.g.* by phosphorylation. This suggests that in many cases the activity of regulator cannot be inferred from transcriptional data alone.

While methods to infer causal regulators directly from heterogeneous primary data types have been proposed and successfully validated in some cases [43], [44], these methods necessarily require very large sample sizes and a mix of different data types, *e.g.* genomic and transcriptomic data. Even when such large data sets are available, the choice of which hypothesis to pursue in follow-up experiments might not be easy to make as such methods usually don't relate their conclusions back to already known biological facts.

In this study we are interested in a method that suggests active regulators and corresponding perturbation experiments for a population based on expression data alone. With that goal in mind the above discussion suggests that investigating coherence between regulatee pairs rather than regulator-regulatee pairs might be more suitable for evaluating the activity of a given regulator. To associate regulatee coherence to a possibly only non-transcriptionally controlled regulator requires the use of prior knowledge.

Our knowledge of molecular biology has increased dramatically over the past several decades as evidenced by the 20 million articles currently indexed by the National Library of Medicine [45]. This immense body of knowledge contains many experiments that define the response of a biological system to a stimulus, *e.g.* by altering its transcriptional state. These experiments can be translated into causal regulatory relationships. Whenever the question of causal regulators is relevant (*e.g.* in finding potential intervention points in ovarian cancer), each experimentally validated finding constitutes a hypothesis that can be evaluated based on the data set at hand.

In contrast to many previous approaches, we rely on a very specific type of prior knowledge, namely relationships extracted from the literature that are (a) based on well-described laboratory experiments, (b) have a citation in the literature, and (c) establish a flow of causality; *i.e.* a directed flow of information form a well-defined perturbation experiment to observed molecular changes. Consequently, our method does not rely on association but on established causation and should be well-suited to providing hypotheses about the causal regulators that are active in a given dataset. This specifically allows us to make statements about regulators that are not limited to activation by changes in transcript abundance, but can include changes in protein abundance or post-translational modifications.

Prior biological knowledge has been used successfully in many contexts before. Relevant functional terms, dysregulated pathways in diseases as well as active miRNAs and transcription factors are routinely predicted based on so-called Gene Set Analysis (GSA) methods and a number of statistical procedures have been proposed, *e.g.* [23–27]. However, virtually all such methods focus on differential gene expression between two conditions as opposed to coordinated changes in a subject population. The situation is similar for methods that that utilize networks of prior information [46, 47]. A class of methods that has been proposed to detect subnetworks co-expressed across a population [48, 49] is related in that it utilizes expression data across all conditions in conjunction with a network. However, such methods aim at finding coherence between expression correlation and distance in the network in general. Our goal is to specifically assess whether a regulator is likely to be active based a given expression data set and consequently point the researcher to relevant perturbation experiments in the literature.

To this end, we introduce the *Correlation Set Analysis* (henceforth referred to as CSA) method in the following, provide evidence that it performs well on simulated datasets, and apply it to three different disease settings: ovarian cancer [50], metabolic disease [51], and diffuse large B-cell lymphoma (DLBCL) [52].

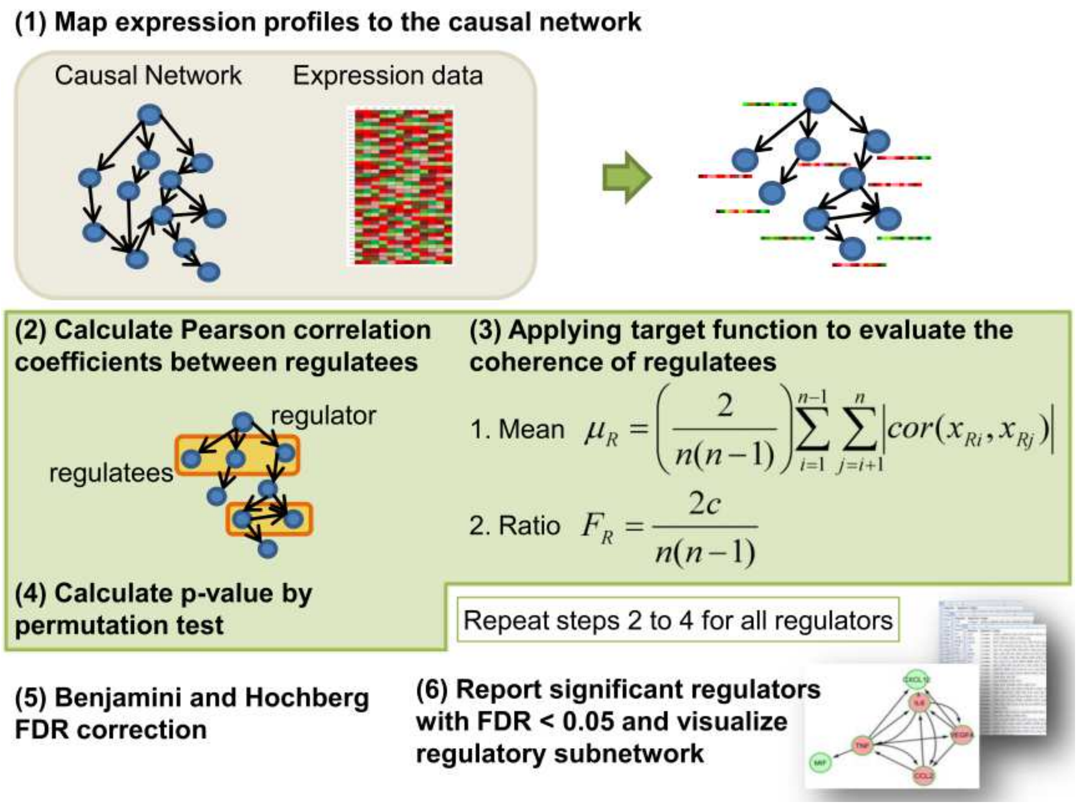

Figure 2-1. Schematic illustration of the Correlation Set Analysis (CSA) method.

2.2 Materials and methods

We identify regulators that are active under a given condition if (a fraction α of) its known targets show (a specified degree β of) correlated transcriptional change. This method does not infer novel regulatory relationships, but rather selects from a library of experimentally defined relationships which regulators test as active in the population of interest. Figure 2-1 provides an overview of CSA. We elaborate on each step below.

2.2.1 Constructing the causal network

The suggested method relies on a causal network to define regulators and regulatees and can only be as good as the encoded biological facts. The causal network consists of relationships that (a) are based on well-described laboratory experiments, (b) have a citation in the literature, and (c) most importantly, establish the flow of causality directed from a specific regulator to a specific regulatee. Consequently, our method does not rely only on association, but on established causation. For example, consider the following statements extracted from two articles represented in the Ingenuity [53] data:

1) "Binding of mouse Fyn protein and mouse Cnr1 (Pcdha4) protein occurs in mouse brain." (PMID 9655502)
2) "Blockade of CB1 (CNR1) increases expression of hepatic lipase (LIPC)." (PMID 20110567)

While statement 1 asserts a biologically correct binding event, it doesn't imply a directed flow of information and it is unclear what consequences the binding event has. In contrast, statement 2 describes a perturbation experiment that causally leads to observed changes. Only statement 2 allows for a meaningful definition of regulator and regulatee. Ultimately, the use of causal statements facilitates the interpretation of results and focuses the analysis on potential upstream drivers of the process under consideration.

Regulators and regulatees can be of different molecular types and include transcript levels, protein levels, protein activities and phosphorylation states. For the purpose of this method, we restrict our causal network to transcript regulatees as this is consistent with the population measures analyzed. In contrast, regulators include transcript and protein levels as well as protein modifications and activities. For CSA, we abstract these different forms into an undifferentiated node in the causal network based on their Entrez identifier [54].

It is important to note that the results of such transcriptional perturbation experiments do not necessarily capture direct physical relationships. In the example above, the increase in transcript levels of LIPC is certainly mediated by a cascade of other signaling molecules. Consequently, the CSA method is not limited to transcription factors as regulators, but encompasses many other classes of molecules amenable to perturbation experiments.

To ensure the reliability of the data, we only include manually curated statements. The substrate for the causal network is licensed from two commercial sources, Selventa Inc. [55] and Ingenuity Inc. [53] and, after filtering and post-processing, reduces to 6,942 regulators and 11,134 regulatees. Selventa Inc. has recently launched an initiative to provide access to a significant amount of causal information to academic researchers through the BEL-Portal (http://belportal.org). Among 6,942 regulators, 3,002 are proteins or mRNAs and 3,940 are chemical compounds or environmental factors (*e.g.* internal metabolites such as glucose or pyruvate, approved drugs such as Rosiglitazone or Doxorubicin, or environmental conditions such as hypoxia or oxidative stress). As described above from this we selected the subset of proteins and mRNA regulators.

After removing self-regulation and regulators that only have one regulatee, the causal network reduces to 1,783 regulators and 10,097 regulatees.

2.2.2 Scoring putative regulators

We assumed an active regulator under a given condition should activate or inhibit a subset of its regulatees. Across a set of conditions (*e.g.* in a patient population), this relationship should become apparent in a coordinated change in expression levels for regulatees downstream of an active regulator. We used different scoring functions to identify active regulators.

2.2.2.1 Mean scoring function

Pearson's correlation coefficient is one of the most widely used measures to evaluate similarities of gene expression profiles. For an expression dataset with m samples, the co-expression level of any two genes X and Y can be calculated by the correlation coefficient $cor(X,Y)$.

$$cor(X,Y) = \frac{\sum_{k=1}^{m}(X_k - \overline{X})(Y_k - \overline{Y})}{\sqrt{\sum_{k=1}^{m}(X_k - \overline{X})^2 \sum_{k=1}^{m}(Y_k - \overline{Y})^2}}$$

, where $\overline{X}$ and $\overline{Y}$ are sample means of gene X and gene Y respectively.

To assess the expression coherence of regulatee sets, we employed the simple test of measuring all pair-wise correlations within each set. Such a coherent regulatee set is consistent with the hypothesis that the corresponding regulator is active in the condition under consideration.

The causal network also specifies the type of regulation (*i.e.* up-regulation or down-regulation). Consider two regulatees, X and Y, that are under the control of a common regulator. If X and Y are regulated coherently, we expect their transcriptional profiles to be correlated. Conversely, if X is up-regulated and Y is down-regulated, we expect their profiles to be anti-correlated. We examined the correlation coefficients between up-regulated regulatees and down-regulated regulatees to test this hypothesis. However, we did not observe significant differences between correlation coefficients of regulatees regulated in the same direction and regulatees regulated in the opposite direction. Thus, we decided to use the absolute value of the correlation coefficient $|cor|$ in the scoring functions.

One intuitive way of detecting regulators with highly coherent regulatee pairs is to examine the average of all absolute correlation coefficients between all pairs of regulatees x_R for a regulator R.

$$\mu_R = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left| cor(x_{Ri}, x_{Rj}) \right|$$

Here, n is the number of regulatees of the regulator R. μ_R is referred to as the mean scoring function in the rest of this paper.

2.2.2.2 Ratio scoring function

If we expect that a substantial number of regulatees is affected by an active regulator, a test for a shift in mean pair-wise co-expression is sensible. However, we also investigated possible scenarios based on the biological data sets described in the results section. Figure 2-2a shows an example distribution of absolute correlation coefficients between regulatees which has higher average absolute correlation coefficients in a real network than in a randomized network. In this case, the majority of regulatees have similar expression patterns, which supports the hypothesis that this regulator is active. In some cases we observed a small bump at the high absolute correlation tail (Figure 2-2b), which indicates a small set of strongly co-expressed regulatees. This situation is more difficult to detect by examining the difference of average correlation coefficients. Hence, we propose an alternate way to detect active regulators: scoring regulators according to the ratio of highly coherent regulatee pairs over all regulatee pairs (referred to as the ratio scoring function).

$$F_R = \frac{2c}{n(n-1)}$$

,where c is the number of regulatee pairs, for a specified regulator R having n regulatees, with absolute correlation coefficient greater than a cutoff. Users can define biologically relevant pairs by setting the cutoff to levels appropriate to detect a desired effect size (say, correlation coefficient larger than 0.6). In this paper, we fix the cutoff, c, at the 95th percentile of the distribution of all pair-wise correlations for a given dataset. This alternate score will identify small sets of highly coherent regulatees. The decision rule based on μ_R and F_R is described in the next section.

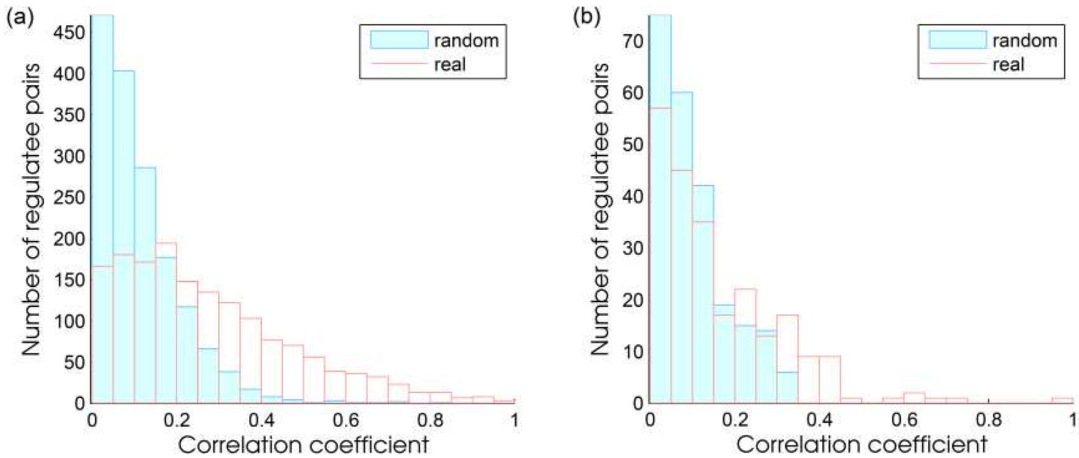

Figure 2-2. The example distributions of absolute correlation coefficients between regulatees of a regulator detected by different target functions. a) The average absolute correlation coefficient between regulatees in the real network (red) is significantly higher than it in the random network (blue). b) There is no significant difference between the absolute average correlation coefficients in the real network and the random network. However, there is a small bump at the right hand side, which means a small subset of highly correlated regulatees. The ratio scoring function was designed to detect such small subsets of regulatees.

2.2.3 Assessing statistical significance

These scoring functions provide rank-ordered lists of all regulators in the causal network based on the coherence of their downstream regulatees, and indicate which ones may be active. While true signals will tend to lead to high scores, high scores in any given result may be due to random noise. We therefore evaluated the statistical significance of the scores, μ_R and F_R of a regulator R, using a permutation test in two ways:

1) Gene permutation randomly assigns transcript profiles to regulatees and, thus, compares the score of the regulator R to the distribution of scores attained by regulators with the same number of randomly chosen regulatees.
2) Graph permutation generates a random causal network in which each regulator controls the same number of regulatees and each regulatee is controlled by the same number of regulators as in the original network.

The corresponding permutation scheme (Graph permutation) is more complex and also computationally more intensive. In each permutation, we evaluate SR on a random graph with the same degree distribution as our causal network. Randomizing a directed graph with a given degree sequence is an active field of research and we adopt a method from [56] relying on edge switching.

More precisely, at every iteration we pick two edges, say (a,b) and (c,d), uniformly at random from the set of edges E in the current graph, and replace them with the edges (a,d) and (c,b). This operation is known as an edge switch, and preserves the in- and out-degree distribution of the graph. If the resulting graph remains simple (no parallel edges) and weakly connected, it replaces the current graph. In order to save the computationally expensive connectivity checks, a batch of K edge switches can be performed before a connectivity check. If the check succeeds, K can be increased, while if it fails, K can be decreased. The particular adaptive algorithm we use to update K is described in [56]. We also adopt the commonly used rule of thumb [57] for the total number of edge switches to perform before declaring our graph to be sufficiently randomized, which is to perform an average of 3 edge switches per edge of the initial graph. Both permutation approaches assert the statistical significance of a score SR under the respective null hypothesis, and thus provide guidance to the biologists as to whether a particular regulator received a high score based on chance alone.

Both permutation approaches assess the statistical significance of a score under the respective null hypothesis, and thus provide guidance to the biologist as to whether a particular regulator received a high score based on chance alone. Note that the permutation of sample labels is not a meaningful option in the scenario of only one population when considering correlation though it is a preferred choice in many gene set analysis methods comparing two or more sample populations.

As the causal network contains more than one thousand potentially active regulators, the resulting p-values should be corrected for multiple testing. The false discovery rate (FDR) is an intuitive and well-accepted alternative measure of significance that is widely applied in similar applications. The Benjamini and Hochberg procedure was used to estimate the FDR based on the list of p-values [58]. Finally, CSA reports a results table of potentially active regulators (FDR < 0.05) which contains FDR, scores (ratio and mean), regulatees coherently up- or down-regulated by the regulator, non-coherently expressed regulatees, average correlation coefficient of regulator to regulatees, and the number of coherent regulatees. Users can rank regulators by the scores (FR or μ_R), the number of coherent regulatees (nc), or the average correlation coefficient of regulator to regulatees (μ_{RR}).

2.2.4 Generating simulated data

In order to make simulated data follow the distribution of the real data, we generate the simulated data based on the ovarian cancer data used in this study by applying the Cholesky decomposition method of Iman and Conover [59]. This approach is widely used to generate correlated random variables in finance. First, we randomly permute sample labels of a gene and repeat this step for all genes. Hence, the mean and standard deviation of each gene vector are invariant under permutation. The correlation between genes, however, is disrupted. Then, we randomly pick N regulators as test set of regulators. Each test regulator Rp can regulate a set of regulatees $T = \{T_1 \ldots T_q\}$, T is an n $\times$ m matrix, where n is the number of regulatees and m is the number of samples (m = 391 in this case). We want to generate a set of new regulatees $T^s = \{T_1^s \ldots T_q^s\}$ with same mean and variance as the original set of regulatees but correlate with each other at a desired correlation coefficient ρ. An n $\times$ n covariance matrix Σ s with desired correlation coefficient ρ is created. The Σ must be symmetric and positive-semidefinite. Therefore, it can be rewritten as $\Sigma = LL'$ by Cholesky decomposition, where L is the lower triangular matrix of Σ with positive diagonal entries. We can get "spike-in" regulatee matrix T^s with desired correlation coefficients ρ by postmultiplies T by L. Iterate above steps until all regulatees of the test regulators are modified to be correlated with correlation coefficient ρ.

2.2.5 Experimental data

Expression datasets of adipose tissue and DLBCL are downloaded from the Gene Expression Omnibus [60]. Adipose tissue samples from 701 individuals [GEO:GSE7965] with a range of age from 18 to 85 and average BMI nearly 30 were used in this study. Pretreatment tumor samples from 181 and 233 DLBCL patients [GEO:GSE10846] were used in this study. TCGA [35] provides mRNA measurements of serous ovarian cancer tissue using 3 array platforms: an Agilent array, and Affymetrix's U133A and exon arrays. Where genes are overlapping among the 3 platforms, we would like to combine the values into a consensus gene. Here, we follow an approach originally described by Verhaak *et al.*[61]. In short, the consensus gene is estimated using a standard factor model based approach:

$$\hat{x} = (\hat{\beta}^T (\hat{\beta}\hat{\beta}^T + \hat{\Psi})^{-1} y)^T$$

where $\hat{\beta}$ and $\hat{\Psi}$ are the platform specific coefficients and error covariance estimates, respectively, y is the 3-by-m dimensional gene expression values across the 3 platforms, and $\hat{x}$ is the m-dimensional, unified estimate for a single gene. For complete details, see Verhaak, *et al.*[61]. In those cases where only 2 genes are shared across the 3 platforms, we take the mean value.

For all of the 3 datasets, we used LSimpute to impute missing values in the expression profiles [62]. We discarded genes that are not included in our causal network and leave 9,052, 9,950 and 7,673 genes in adipose tissue dataset, DLBCL dataset and ovarian cancer dataset respectively.

2.2.6 Performance metrics

The receiver operating characteristic (ROC) curves are used to evaluate the performance of CSA. The true positive rate and false positive rate used for plotting ROC curves are calculated as following:

$$\text{True positive rate} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}}$$

$$\text{False positive rate} = \frac{\text{False positives}}{\text{False positives} + \text{True negatives}}$$

2.3 Results on simulation data

To assess the sensitivity and specificity of CSA, we generated simulated data sets with various characteristics. To retain a realistic scale for the data values, we derived our simulated data from the Ovarian Cancer dataset (see "Materials and Methods"). The dataset was derived from 391 ovarian cancer patients in TCGA [50]. To obtain a baseline dataset with no signal, we randomly permuted the sample labels for each gene vector separately. Consequently, each gene vector retains its original distribution, but correlations between gene vectors are disrupted. We labelled *n* genes as active regulators in the simulated data. Each induces expression profiles in *p*% of its regulatee pairs that have a Pearson correlation coefficient of *r*. Regulators and regulatees are defined according to the literature-based causal network described earlier. To evaluate CSA with respect to many different signal-to-noise characteristics, we varied the percentage of correlated regulatee pairs *p* in 10% increments from 0% to 100%. Similarly, we set the correlation coefficient r to {0.3, 0.4, 0.5, 0.6}. Details on the generation of dependent profiles can be found in the Materials and Methods section.

To test the robustness of the method to sample size, we generated additional datasets with a random subset of 20, 100, and 200 patients. Finally, we generated simulated sets based on *n*=10 as well as *n*=100 embedded active regulators. We found that the recovery of true positives was not affected by the number of embedded regulators. We therefore fixed the number of embedded regulators in the subsequent examples at *n*=100.

2.3.1 Evaluation of permutation methods and controlling false positives

While the scoring functions are able to rank embedded active regulators higher than non-active ones, they do not provide an objective cutoff value when investigating the biological significance of top results. In order to assess the suitability of our procedures to control the FDR, Figure 2-3a shows the false positive rate (for definitions, see

"Materials and Methods") of the two scoring functions with the two permutation methods on a representative simulated data set (r=0.5; p=50%). Our procedures are able to control the false positive rate effectively based on the estimated FDR while retaining a good true positive rate (Figure 2-4). In fact, when the data contains no or limited signal, CSA does not report any potentially active regulators at reasonable FDR cutoffs (FDR < 0.05). The same holds true for runs on randomized networks (data not shown).

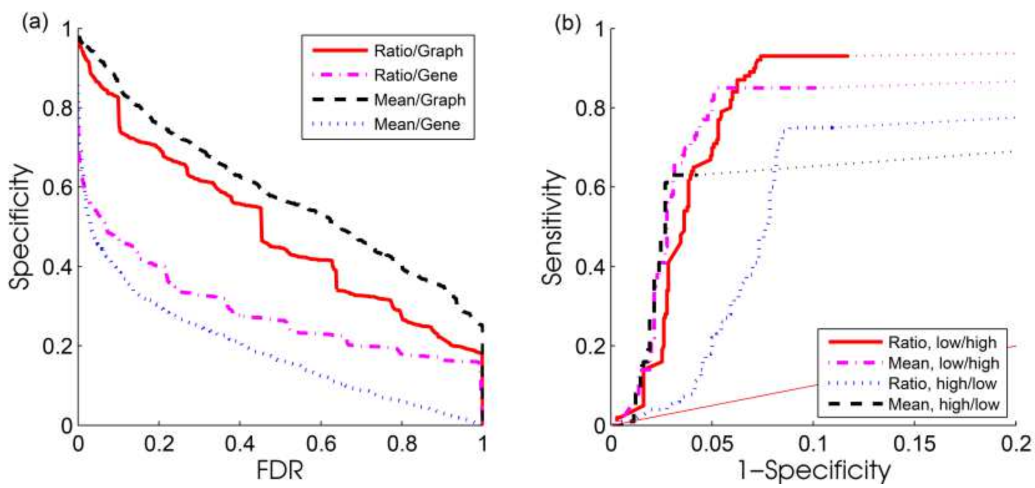

Figure2-3. (a) False positive rate of CSA with different parameter settings at different FDR levels. (b) ROC curves of mean function and ratio function of low/high and high/low simulated data. (a) The plot clearly shows that the estimated FDR can well control false positive rate of CSA. Both scoring functions with graph permutation reach low false positive rate when applying a reasonable FDR cutoff (FDR < 0.05). (b) The ROC curves suggests that the ratio scoring function reaches better true positive rate at the expense of a similar gain in false positive rate on datasets that contain few highly correlated regulatees.

Figure 2-3a also illustrates that Graph permutation is preferable to Gene permutation. ROC curves of Graph permutation and Gene permutation further prove that both graph permutation and gene permutation can reach good sensitivity and specificity, but graph permutation has higher specificity than gene permutation (Figure 2-5). The purpose of our method is to find active regulators in a certain condition, which means that specificity might be more important than sensitivity in our case. We will only focus on Graph permutation results in the following. In contrast, the mean and ratio scoring functions seem to perform comparably, and a more in-depth analysis is needed.

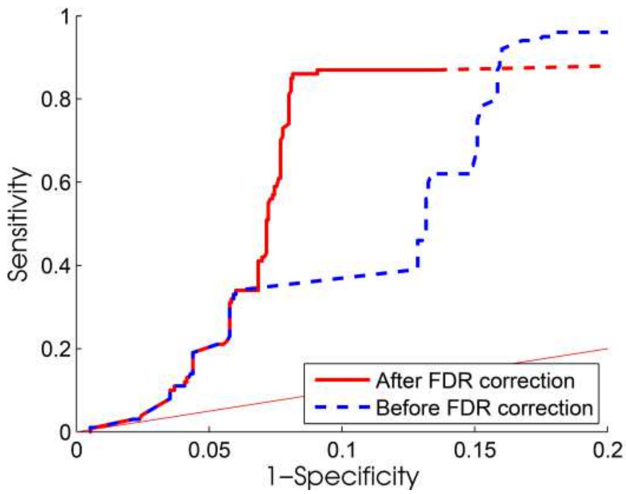

Figure 2-4. Effects of permutation on ranking regulators. ROC curves show that FDR calculated based on permutation can improve both sensitivity and specificity. Graph permutation is used here. Permutation and FDR correction can decrease false positives by filtering out regulators that have high scores but only regulate a small number of regulatees since a regulator can easily obtain a high score by chance if it only regulates a few regulatees. Similarly, permutation and FDR correction can increase true positives by recruiting regulators that have fair scores but can regulate a large number of regulatees.

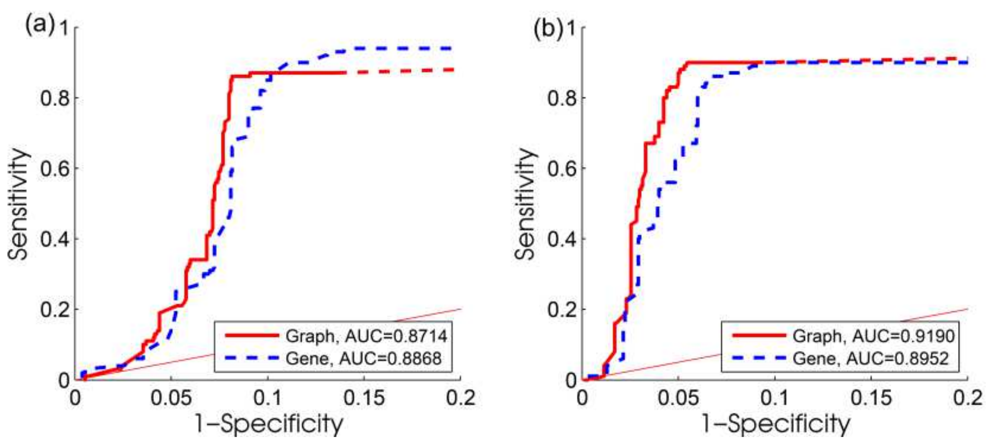

Figure 2-5. Comparison of graph permutation and gene permutation on two representative simulation data sets. (a) Simulation data set (r=0.5, p=50%). (b) Simulation data set (r=0.3, p=70%). Scoring function "ratio" is used in both cases. Both permutation methods can reach good sensitivity and specificity. However, graph permutation reaches slightly better specificity in most cases.

2.3.2 Evaluation of scoring functions

To understand the characteristics of the ratio and mean scoring functions, we focused on four datasets which differed in the strength of correlation r and the quantity of correlated regulatees p, namely, low/low (*r*=0.3; *p*=30%), low/high (*r*=0.3; *p*=80%), high/low (*r*=0.6; *p*=30%), high/high (*r*=0.6; *p*=80%).

Figure 2-3b depicts receiver-operator characteristic (ROC) curves (see “Materials and Methods” for definitions) for the low/high and high/low datasets. The only substantial difference between the two functions becomes apparent in the case of few highly correlated regulatee pairs, in which the ratio function is able to reach higher true positive rate at the expense of a similar loss in true negative rate. Note that the ratio function is explicitly designed to address this case. In most other situations, the two functions are comparable with the mean function performing slightly better.

2.3.3 Robustness to signal level and sample size

Figure 2-6 shows the ROC curves under a variety of signal levels to demonstrate the ability of CSA to detect active regulators. In this instance, we use the ratio scoring function, but curves based on the mean function give similar results (data not shown). The curves demonstrate that CSA is able to pick out true active regulators embedded in the simulated data. For large sample sizes, the true positive and true negative rates were consistently high (>80%) for a wide range of score cutoffs. As expected, performance deteriorated with decreasing signal, but remained useful, even for very low levels of signal (Figure 2-6a). In contrast, Figure 2-6b depicts the situation with only 20 patient samples. While for strong signals (*p*>70%), some regulators can be detected, weaker signals result in performance close to random. Together, this shows that our causal network based on literature information is informative enough to enable recovery of embedded signals, given enough patient samples.

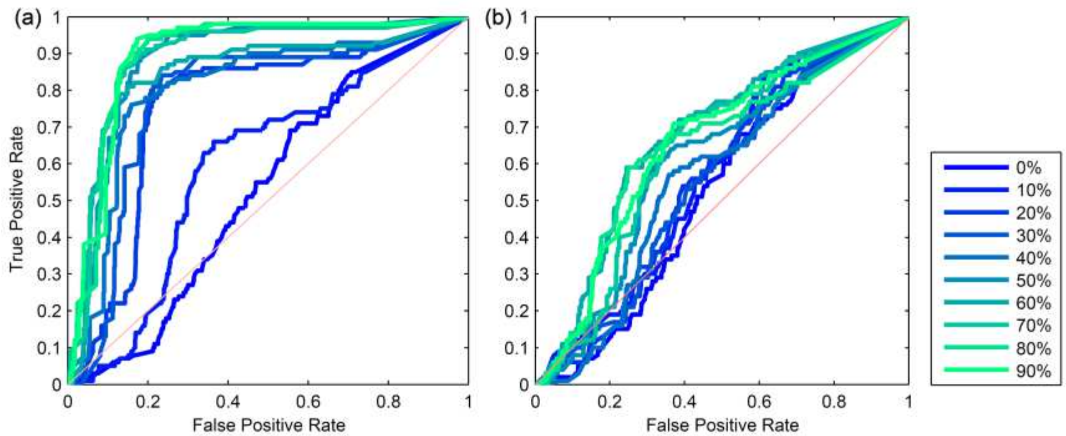

Figure 2-6. Robustness of CSA with respect to different levels of signal (0%-90%) and sample size. ROC curves based on (a) 391 samples and, (b) 20 samples. CSA reaches high true positive rate and low false positive rate for different signal levels.

2.3.4 Relevance of the causal network

To further illustrate the relevance of the underlying causal network to provide informative active regulators, we generated a randomized version of the causal network with the same degree distribution (using the edge-switching procedure described in the "Materials and Methods"). Running CSA based on this randomized network against the simulated data and ovarian cancer data from TCGA resulted in ROC curves that were indistinguishable from random, indicating that the causal network is biologically informative (data not shown).

2.3.5 Comparison to degree-based ranking

Finally, we compare CSA's results to an alternative approach that has been suggested as a general principle in many approaches to transcriptional network reconstruction, namely the prediction of key regulators or biomarkers based on their degree in the inferred network [37, 63, 64]. Here, we use the same representative simulation data set (*r*=0.5; *p*=50%) as we used in the previous section. Implementing a ranking strategy based on each candidate regulator's out-degree (number of targets they coherently regulate) gives an interesting baseline performance (Figure 2-7) that is clearly better than random. However, the ROC curves suggest also that the results based on our method are superior to a purely degree-based method.

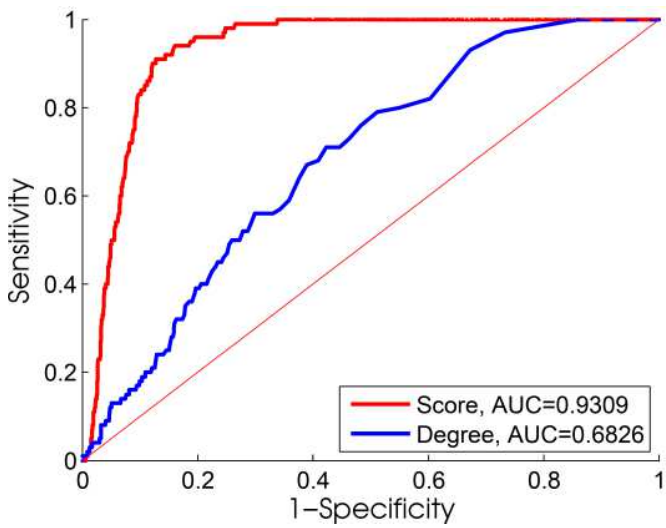

Figure 2-7. ROC curves of ratio score ranking and degree-based ranking. Red and blue ROC curves show ranking by ratio score and by out-degree (*e.g.* number of coherent expressed targets), respectively. The ROC curves suggest that degree-based ranking is better than random. However, ranking based on ratio score is superior to purely degree-based ranking.

2.4 Results on clinical data

To illustrate the utility of the CSA approach we describe here the results of comparing the directed perturbation experiments captured in the causal network to 3 different surveys of expression variation in 3 distinct disease and tissue settings – subcutaneous adipose tissue [51], ovarian cancer [50] and diffuse large B cell lymphoma [52].

2.4.1 Experimental results I – Adipose tissue

The first population dataset was subcutaneous adipose tissue from 673 individuals as described by Emilsson *et al.* [51], representing individuals from 3 generation families with a range of ages and degrees of obesity that was used to define loci affecting obesity in the Icelandic population. After matching transcripts measured in the adipose cohort to the causal network, the CSA method reported 246 of 1,762 (14%) regulators as potentially active at an FDR <0.05. This corresponds to 8,946 potential regulator:regulatee edges.

These data can be summarized by counting the number of CSA significant regulatees for each regulator (Table S1 Appendix A). Amongst the top ranked regulators in adipose were some well-known metabolic targets, including PPARG (nc=275), PPARA (nc =218), Insulin (nc=136) and PPARGC1A (nc=105). The top hit as judged by the size of significant regulatees was MYC (nc=391) which has been implicated in adipogenesis [65]. It is interesting to note that the well-known transcription factor, MYC, was not co-expressed with its regulatees in the adipose tissue dataset (average correlation coefficient=0.1161), but a subset of its regulatees were coherently expressed. This observation supported our hypothesis. An additional top hit was, NFE2L2 (also known as NRF2, nc =285), a master regulator of anti-oxidant response that has been implicated in many disease processes and in adipogenesis and obesity specifically [66].

Adipose tissue is composed of adipocytes and a stromal fraction including macrophages. Given this knowledge, we asked if CSA provided evidence for these sub-populations of cells. Perilipin (PLIN1) is a protein uniquely expressed in adipocytes (see Figure 2-8a and 2-8b) where it coats the surface of intracellular lipid droplets and protects them from degradation by lipases. CSA identifies 37 PLIN1 regulatees as cohesive in adipose tissue consistent with it being a significant regulator in human adipose tissue. A major conclusion of the adipose tissue study used here was that macrophages, as observed by macrophage-specific transcripts, are identified as causal drivers of obesity in humans [51] and mouse [67]. Consistent with this CSA finds a number of macrophage specific genes as active drivers including the chemokine receptor CCR1 (Figure 2-8c and 2-8d). One of the ligands of CCR1, RANTES is reported to be secreted by adipocytes and recruits macrophages to fat depots [68]. CCR1 appears as both a target of other regulators and as a regulator of downstream transcripts as judged by CSA (see Figure 2-8c). Furthermore the regulators of CCR1 were also found to be connected to each other consistent with a web of regulatory interactions affecting CCR1 and its downstream targets in macrophages in human adipose tissue.

Given the relative ease of experimentation, it is not surprising that many experiments reported in the literature were performed in cultured cell models. In this setting there is always a question of the relevance of the results to human populations. CSA potentially provides a data-driven way to assess this by testing whether any perturbation signature (or fraction of a signature) is significantly cohesive in a disease population. Interestingly, CSA identified many cell culture derived signatures as being relevant to the subcutaneous adipose tissue of humans (more than 53% individual references using cells in culture covered 47% regulator:regulatee edges). This included, for example, 88 of 254 regulatees for the classic adipose regulator PPARG and 68 of 116 insulin regulatees.

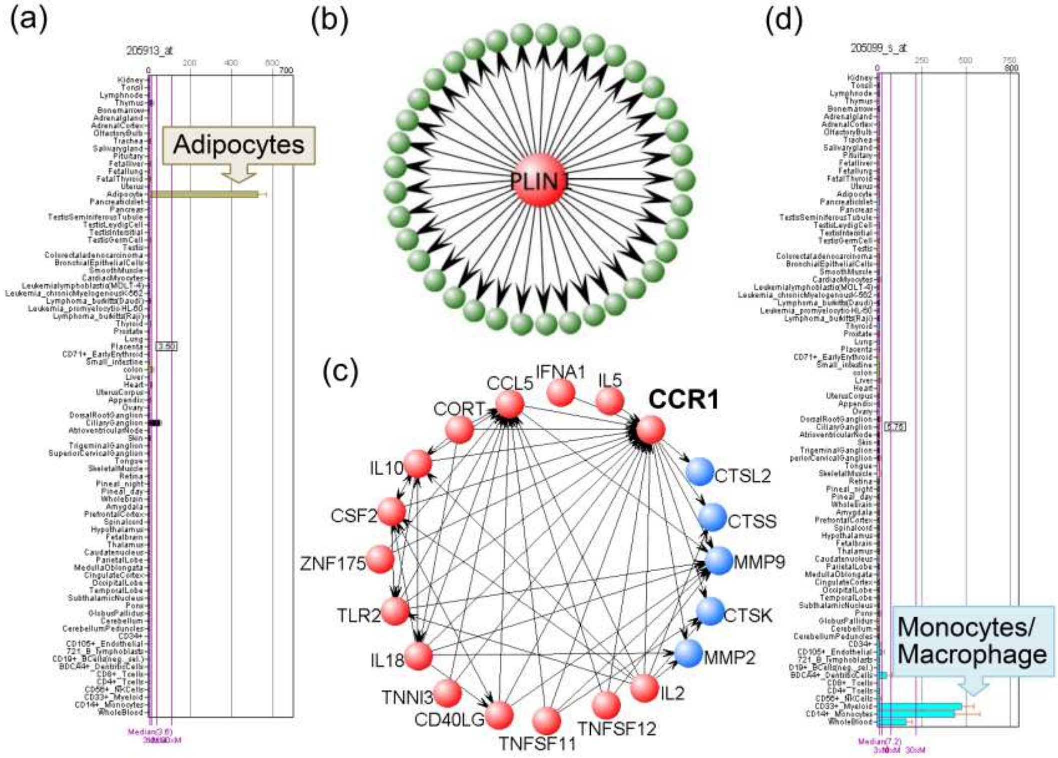

Figure 2-8. (a) Gene expression of PLIN1 in different tissues. (b) PLIN1 and some of its downstream regulatee. (c) CCR1 and its upstream regulators (red nodes) and downstream regulatees (blue nodes). (d) Gene expression of CCR1 in different tissues. (a) Gene expression of PLIN1 in different tissues. (b) PLIN1 and some of its downstream regulatees. (c) CCR1 and its upstream regulators (red nodes) and downstream regulatees (blue nodes). (d) Gene expression of CCR1 in different tissues. (a) and (d) are from BioGPS, which show that PLIN1 and CCR1 are uniquely expressed in adipocytes and Macrophages, respectively. (b) PLIN1 regulates 37 regulatees in adipose tissue. (c) CCR1 is regulated by numerous regulators in the causal network. CSA identified 14 potential active regulators of CCR1 in the adipose tissue (red). CCR1 is a regulator that can regulate several downstream regulatees (blue); at the same time, CCR1 is also regulated by many other regulators. These regulators regulate each other and also CCR1's regulatees.

2.4.2 Experimental results II – Serous ovarian cancer

The CSA method was also assessed against a collection of expression profiles from almost 400 human serous ovarian cancers available through TCGA project [50]. After matching transcripts measured in the ovarian cohort with the causal network, it was found that 357 of 1,398 (26%) regulators had an FDR <0.05 in the CSA analysis (see

Table S2 Appendix B). This identifies 12,860 potential regulator:regulatee edges as active in this cohort.

As before, the potential regulators can be ranked by the number of regulatees that pass the cutoff. Top ranked regulators include genes implicated in many cancers (see Table S2 Appendix B) such as TGFB1 (nc=520), IFNG (nc=485), FGF2 (nc=241), MYCN (nc=219) and VEGFA (nc=183). Again, we observed that another well-known transcription factor, MYCN, was weakly correlated with its regulatees (average correlation coefficient = 0.1015), but it had a subset of coherent regulatees and were identified by CSA as active in ovarian cancer dataset. CSA also identifies numerous potential drivers of various cyclins which are known to drive the cell cycle and be aberrantly regulated in cancer (see Figure 2-9a). Among 603 regulators that regulate cyclins in the causal network, CSA identified 77 as active in the serous ovarian cancer. As with the CCR1 example extensive cross-regulation of cyclin regulators was predicted, suggesting the presence of a complex network of causal interactions upstream of cyclins in ovarian cancers revealed by CSA. The derivation of the higher-order structure is non-obvious given the literature examples alone, but readily emerges from the CSA analysis as relevant in this context. Using the ranking by the number of edges testing significant in CSA, TNF was found to have 657 regulatees (reactive to TNF) and 153 regulators (causal regulators of TNF) in this cohort. High levels of secreted TNF protein (ranked first by CSA) were reported to cause high levels of the proteins IL6, MIF, CCL2, CXCL12, VEGFA, also secreted in ovarian cancer cell lines [69]. Intriguingly, these genes in the CSA analysis of the ovarian cancer cohort were also found to have many regulatees (304, 0, 13, 0, 183, respectively) and/or regulators (158, 5, 64, 18, 26, respectively), and directly and reciprocally regulate each other (see Figure 2-9b). As reported by Kulbe *et al.* [69], knockdown of TNF in cells with high TNF results in failure to form tumors in mice. The interpretation of such established cell line models of cancer can be difficult. The CSA analysis however indicates that the basic regulatory findings of this work were replicated by CSA and found to be the most dominant feature in approximately 400 ovarian clinical cancer samples. Together, these findings may identify a critical driver of ovarian tumor growth in vivo.

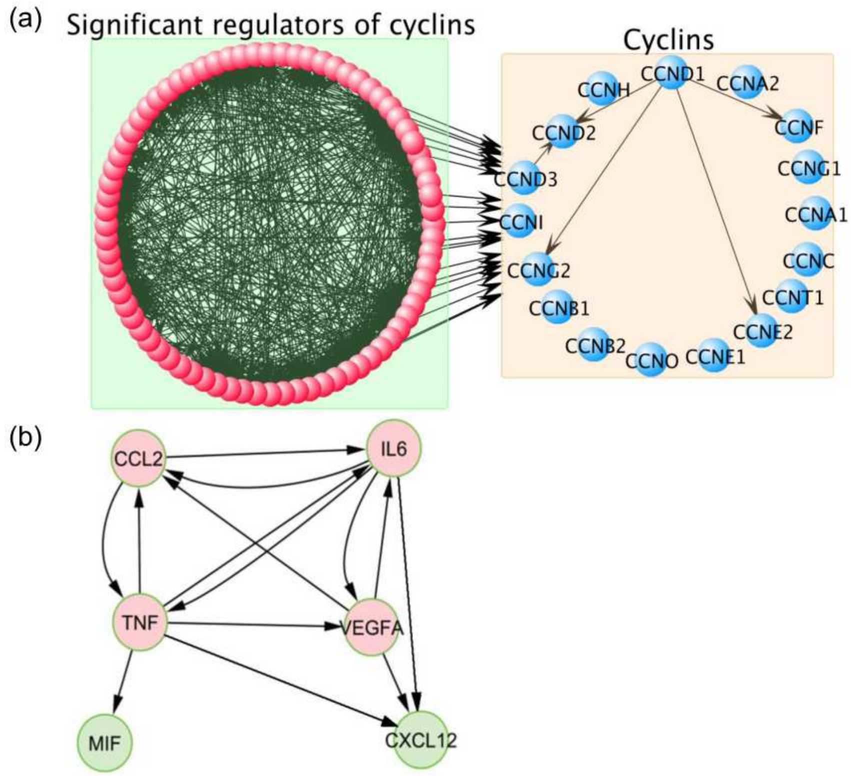

Figure 2-9. (a) Regulators regulate cyclins in serous ovarian cancer. (b) Hypothesis regulatory model of secreted proteins in serous ovarian cancer. (a) Cyclins have 603 regulators in the causal network. CSA identified 358 potential active regulators in serous ovarian cancer; 77 of 358 regulators were found to regulate cyclins. Regulators (red nodes) regulated cyclins (blue nodes) and also regulated each other, which implies that these regulators work cooperatively to regulate cyclins. (b) Secreted proteins TNF, IL6, VEGFA and CCL2 were identified as regulators (red nodes) in serous ovarian cancer by CSA. They regulated each other and two other secreted proteins, MIF and CXCL12 (green nodes). TNF, IL6, VEGFA and CCL2 are also used as therapeutic targets of several different kinds of cancers [70–73].

2.4.3 Experimental results III – DLBCL

The final example is a study of diffuse large B-cell lymphoma (DLBCL) in which expression profiles of 2 patient populations who subsequently received different treatments were examined for signatures that predict the clinical course of the disease [52]. For the purposes of this analysis the subsequent treatments are not relevant. The first cohort (CHOP) included 181 samples and the second cohort (R-CHOP) included 233 samples. As described in [52], 3 signatures were derived in a multivariate analysis that predict survival in the 2 cohorts. The Germinal Center B-cell signatures contained 37 genes, the Stromal-1 signature contained 264 genes and the Stromal-2 signature contained 61 genes. CSA analysis was applied to each of the cohorts and potentially active regulators identified that pass the FDR cut-off (218 and 220 of 1780 significant hits for CHOP and R-CHOP, respectively (see Table S3 Appendix C). Using these significant hits we then asked if any of the regulators regulated genes involved in the 3 predictive signatures (Germinal Center B-cell, Stromal-1 or Stromal-2). Interestingly, although the Stromal-1, and -2 signatures were found by a multivariate analysis, suggesting they are independent, CSA analysis identifies genes that can regulate both signatures jointly. Among the 131 regulators that regulate at least one gene in either the Stromal-1 or -2 signatures, 53 (40%) regulate genes in both cohorts. Furthermore, we calculated the significance of the enrichment of each regulator’s regulatees for overlap with the 3 predictive signatures by Fisher’s exact test. Significant enrichments for the two Stromal signatures were found (see Table 2-1). Figure 2-10 shows the regulators enriched for Stromal-1 and -2 signatures in the CHOP and R-CHOP cohorts and their target genes in all three signatures. 11 regulators were found enriched for Stromal-2 signature in both cohorts. Surprisingly, all of these 11 regulators are also enriched for

Stromal-1 signature, indicating that it is possible the 2 signatures arise because of the same regulator(s).

The candidate regulators can be ranked by the number of predictive signature genes they regulate (limiting to those enriched for the signatures). This results in the identification of some very familiar drivers of many cancers including MYC, MYCN and CCND1 (see Table 2-1).

Table 2-1. Top 15 regulators found in CHOP and R-CHOP cohorts.

	No. of Regulatees[a]		Germinal Center B cell[b]		Stromal-1[c]		Stromal-2[d]	
Regulator	**CHOP**	**R-CHOP**	**CHOP**	**R-CHOP**	**CHOP**	**R-CHOP**	**CHOP**	**R-CHOP**
TGFB1	723	632	0.2379	0.3081	1.36E-12	2.30E-14	0.0003	7.36E-05
IFNG	624	553	0.5868	0.6600	0.0041	0.0006	0.1928	0.1008
MYC	439	410	0.1870	0.2094	0.0013	0.0005	0.1427	0.0673
IL6	370	317	0.5962	0.6678	0.0281	0.0497	0.2135	0.0854
MYCN	272	248	0.3566	0.3908	8.22E-06	7.20E-06	0.4979	0.5463
ERBB2	253	219	0.3731	0.3666	7.50E-07	2.79E-07	0.0149	0.0088
IFNB1	251	233	0.3864	0.4139	0.1869	0.1256	0.4674	0.4826
IL10	249	221	0.3727	0.3672	0.0950	0.0543	0.1307	0.1080
CDKN1A	234	224	0.3704	0.3680	0.0953	0.0554	0.4817	0.2436
F2	233	212	0.4139	0.4484	0.0009	0.000372	0.0409	0.0317
IL2	205	188	0.4605	0.4912	0.3236	0.4066	0.6388	0.6769
CCND1	204	184	0.4622	0.4987	7.86E-05	2.38E-05	0.0013	0.000777
VEGFA	189	170	0.4894	0.5259	3.27E-05	3.60E-05	1.86E-06	7.91E-06
MAPK1	183	166	0.5006	0.5340	0.0063	0.003266	0.3192	0.3550
TNFSF11	166	150	0.1038	0.0898	0.5435	0.6276	0.1905	0.1714

[a]number of cohesive regulatees.
[b,c,d]the enrichment p-values of Germinal Center B cell, Stromal-1 and Stromal-2 signatures in CHOP and R-CHOP groups.

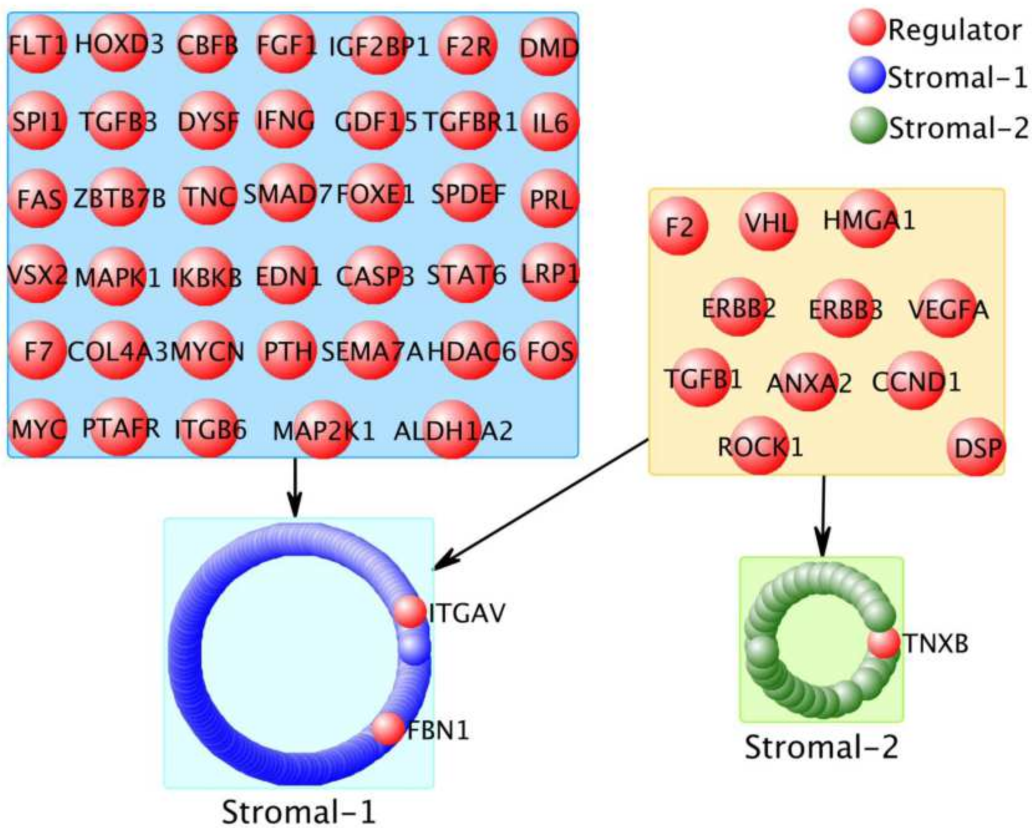

Figure 2-10. Regulators enriched for Stromal-1 and Stromal-2 signatures. Red nodes are regulators. Yellow nodes are genes in Germinal center B-cell signature. Blue nodes are genes in Stromal-1 signatures. Green nodes are genes in Stromal-2 signatures. We did not find any regulator enriched for Germinal Center B-cell signature works in both cohorts. Instead, CSA identified 41 and 11 regulators enriched for Stromal-1 and -2 signatures in both cohorts. Furthermore, the regulatory model showed that the majority of genes in Stromal-2 signature are regulated by regulators that also regulate Stromal-1 signature.

2.4.4 Results on clinical data with the public causal network

Selventa Inc. has recently launched an initiative to provide access to a significant amount of causal information to academic researchers. We performed CSA on the ovarian cancer dataset with the public causal network released by Selventa Inc. The result suggested that 121 of 170 regulators reported by CSA (FDR < 0.05) with the public causal network were found in our previous result (Table S4 Appendix D).

Compared with 358 causal regulators identified by using the complete causal network, CSA can recover about 1/3 of the regulators in the ovarian cancer dataset. The results suggested that CSA works well with the public causal network although it does not report as many causal regulators as with the complete causal network.

2.5 Conclusions

The advent of inexpensive high-throughput transcriptomics measurement techniques has enabled the characterization of cell lines, animal models, and, more recently, cohorts of clinical patients on a molecular level. A crucial research question in such studies (*e.g.* in ovarian cancer patients) is the identification of causal regulators of the observed transcript changes. In this study, we sought to develop a method, Correlation Set Analysis (CSA), to identify directed perturbation experiments relevant to a disease population of interest in an unbiased data-driven manner. The method relies on the hypothesis that regulatees of a regulator active in a population will be significantly correlated to each other across the population. As described the CSA method was developed and evaluated on simulated data where it found to be quite sensitive in identifying known regulator effects.

To test the method and illustrate its uses, we present the results of CSA analysis in three disease settings, where in each case known key regulators and as yet un-described regulators were found. In the regulatory networks of CCR1 in adipose tissue and cyclins in serous ovarian cancer, we observed that upstream regulators of CCR1 and cyclins not only regulated CCR1 and cyclins but also regulated each other. For example, both TNSF12 and CCL5 regulate CCR1 and CCR1's downstream regulatee, MMP1, but CCL5 is also regulated by TNSF12 at the same time. Also in the DLBCL dataset, we found that of 11 regulators enriched for Stromal-2 signature in both CHOP and RCHOP cohorts were also enriched for Stromal-1 signature. This finding implies that Stromal-1 and -2 signatures are actually regulated by the same regulators. These examples indicate that CSA is able to bring together isolated regulatory relationships into higher order regulatory networks, and thereby adding new knowledge and insights.

Based on this we propose that CSA has four key advantages. Firstly, it provides a simple data driven way to connect population data with the many directed perturbation experiments reported in the literature. Secondly, CSA identifies plausible changes in regulators at the protein level, since we do not require correlation between regulator and regulatees. A third advantage of this approach is that the results can be simply ranked providing the user with an intuitive way to prioritize any follow up investigation. Finally, the regulatory relationships inferred by CSA were found in all cases to form a contiguous network, with many cases of regulators that were also regulatees of upstream regulators. In other words CSA provides a glimpse of the true underlying higher order structure of biological systems which can be systematically tested for causal effects on disease progression.

In addition to the advantages it is worthwhile noting that there are some limitations to CSA that stem from its reliance on the literature-derived causal network. Clearly, the relationships identified are limited to the causal network used. With the continued expansion of our understanding of regulatory relations, the accuracy of CSA should improve with time. Another possible limitation is in cases where overlapping sets of regulatees are controlled by distinct regulators. In this case, CSA will not be able to distinguish between the regulators resulting in an increased false positive rate. One possible solution for this problem is to cluster regulators by the similarity of their respective regulatee sets. In our experience, however, the overlap between regulatee sets is still sufficiently small to allow for meaningful inference. This is supported by the excellent performance of CSA on simulated data.

As we mentioned before, the causal network in its original form contains not only genes and proteins, but also chemical compounds and environmental factors. Since our approach does not restrict regulators to genes on the microarray, such entities can appear as candidate regulators, though we did not report them in this work. Hypotheses based on compounds might be useful for assessing environmental factors driving disease progression or for supporting drug repositioning efforts.

In the future we plan to apply the CSA method to large scale compendia of gene expression sets to understand patterns of active regulation across a span of phenotypes. Given the promising performance of CSA when recovering embedded regulators as well as providing insights for biological datasets, we believe that it will greatly enhance our understanding of general and specific mechanisms driving disease and other relevant phenotypes.

Chapter 3. Identification of driver modules in cancer by using a weighted functional linkage network

3.1 Introduction

The number of causal molecular alterations that have been identified in cancer cells is far smaller than the total number of identified alterations; *i.e.* large numbers of alterations occur sporadically, and display in a relatively small fraction of cancer cells. This large background of mutations and the paucity of prior evidence for causality make systematic experimental investigation impractical and render reliable statistical identification of drivers difficult.

Driver identification has been approached in a number of ways. The most widely used approaches attempt to identify single nucleotide polymorphisms (SNPs) or copy number variants (CNVs) having abnormal occurrence frequencies in cancer cells, which are statistically significant [13, 14]. The assumption underlying this approach is that a mutation that drives cancer will confer a growth advantage on the cancer cell, and the mutation will therefore become common. Other approaches exploit the idea that non-synonymous mutations are more likely than synonymous mutations to be potential drivers [18, 74].

More recent methods focus on pathways--such as proliferation, apoptosis, angiogenesis, and metastasis—which can mediate transformation [5]. Since a pathway can in principle be altered by a mutation in any of its genes, the likelihood of finding the same aberrant gene across a set of samples is far smaller than that of finding the same aberrant pathway. Some computational methods therefore focus on the identification of processes that distinguish normal from cancer phenotypes [21, 22]. Many traits, however--including cancers--are not readily distinguished by sets of genes represented by traditional pathways, but reflect crosstalk between and among subsets of genes in different pathways, or previously ungrouped sets of genes. This suggests that methods aimed at identifying phenotypic distinctions should include correlations between genes. For example, Cerami *et al.* identified drivers in glioblastoma by finding modules in a literature curated network [75]. Another study by Torkamani *et al.* constructed coexpression networks in cancers and identified modules enriched in somatic mutations in glioblastoma, breast and colon cancer [76].

Our approach also focuses on functional modules of genes rather than individual genes. In particular we develop and use a novel approach, Functional Linkage Network Partitioning (FLNP), to find breast and ovarian cancer specific functional gene modules that are highly enriched in known oncogenes. The basic idea is simple. Starting with a minimum boundary surrounding a set of known cancer genes that are connected on an evidence weighted functional linkage network, we move genes in and out of the boundary until we reach a module that maximizes a target function, which we take to be the difference between the average intragenic and intergenic weights (Methods). Since the modules consist of genes that have strong functional relations, we classify all other genes in the module as candidate drivers, provided enough is known about their function to plausibly implicate them in processes associated with transformation.

3.2 Materials and methods

Previous studies have shown that proteins involved in the same disease tend to be functionally related, or serve a common process [77, 78]. With that in mind we have developed an algorithm FLNP that detects sets of genes that are in the functional neighborhood of well-known drivers of cancer, where functional neighborhood is defined on our functional linkage network (FLN), as we now explain.

3.2.1 Functional linkage network

The FLN consists of genes (nodes) connected by evidence weighted links [79, 80]. A particular network topology is realized by choosing a threshold for accepting a link; if the link between two nearest neighbor nodes exceeds the threshold, the link is realized, otherwise it is suppressed. FLNs have been used successfully in many contexts, including gene and protein annotation [81–83], genotype-phenotype association for diseases and disorders, [79, 84–86] and the identification of functional modules [87]. In the FLN we used in this study, the link weights were established, by integrating 16 genomic features using a Naïve Bayesian classifier [79]. Unlike protein-protein interaction networks or regulatory networks, the FLN topology reflects functional similarity. Previous studies have shown that FLNs can strongly improve both sensitivity and specificity in identification of candidate disease genes [79, 88].

3.2.2 Seed selection

Our algorithm for finding a module starts by choosing a seed set: a small number of known oncogenes associated with the cancer of interest. Many databases have collected mutations in a variety of diseases, including cancers. The Online Mendelian Inheritance in Man (OMIM) database provides comprehensive and highly reliable contexts of human genes and genetic phenotypes [89]. OMIM covers 14,013 genes and 7,296 genetic phenotypes, including 91 major types of cancers. The Catalogue Of Somatic Mutations In Cancer (COSMIC) is another human curated database which collects somatic mutations in human cancers from literatures [90]. Together OMIM and

COSMIC provide 651 validated oncogenes for about 90 types of cancers (as June 2012); those oncogenes are treated as drivers in this study.

The Genetic Association Database (GAD) extracts candidate genes associated with complex diseases from published papers based on genome-wide association studies (GWAS) [91]. Although GWAS can provide evidence that a gene is disease associated, the association is less reliable than it is for OMIM and COSMIC genes, owing to insensitivity to rare single nucleotide polymorphisms (SNPs), false positives caused by multiple testing, and substantial variation between studies [92, 93].

We select known mutated genes from OMIM, and require that a continuous pathway on the FLN must exist between oncogenes if they are to be considered members of the seed.

3.2.3 Identification of functional disease module

A well-accepted and intuitive definition of a module in a network is a group of densely connected vertices, which have many edges between vertices within the group but with only a smaller number of edges between vertices of different groups [94]. In a biological network, we extend the definition of a module from a set of densely connected vertices to a set of strongly functional related genes. Namely, genes are grouped into a module based on the strength of functional connections rather than number of edges between them. The function may refer to a biological process, molecular function or a phenotype [5]. Therefore, a functional linkage network which represents the functional correlations between genes is a good way for identifying functional modules.

We take as an *ansatz* that genes interior to a module have a higher functional correlation with one another, than they have with exterior genes. Specifically, for a module $M = \{k\}$, consisting of k genes, the set of neighboring exterior genes $N_G(k)$, we define a modularity score, S_d as

$$S_d = \frac{1}{2p}\sum_{i,j}\omega_{i,j}[i \in M][j \in M] - \frac{1}{q}\sum_{i,j}\omega_{i,j}[i \in M][j \in N]$$

where ω is the weighted adjacency matrix, p is total number of edges within the module M, and q is the total number of edges between module M and the neighborhood community $N_G(k)$. The element $\omega_{i,j}$ in the weighted adjacency matrix ω represents the weight of the edge between genes i and j.

The objective is to maximize the target function, S_d, subject to the constraint that it must be greater than θ. θ is the threshold for defining a module, which we take to be zero for maximum sensitivity.

Modules can of course be defined in a number of different ways. For example, KEGG pathways are also modules [95]. If our definition is to be useful for discovery, we'd expect our modules to include those obtained using other definitions. We therefore calculated scores for KEGG pathways and GO terms.

KEGG. We found that 242 of 245 (98.8%) KEGG pathways have a modularity score greater than 0, with an average S_d of 1.27.

Gene Ontology (GO). We classified GO into 12 coverage levels as defined in our previous work [96]. Coverage increases as level decreases; *i.e.* terms become increasingly specific as level increases.

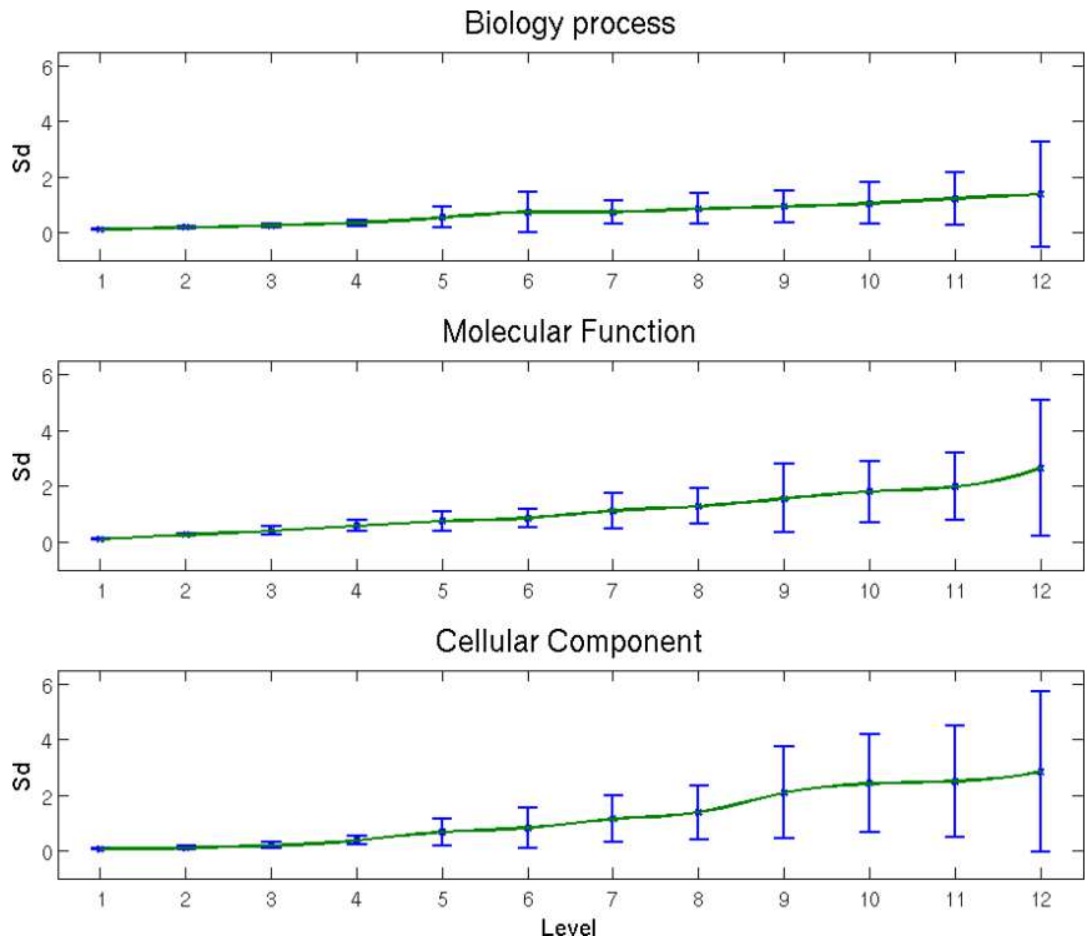

Figure 3-1. Distribution of modularity score S_d of GO terms.

Figure 3-1 shows the distribution of S_d for GO terms (having at least 2 genes) at different levels. The results indicate that 8,833 of 9,865 (89.5%) GO terms in the biological process branch, 2,530 of 2,649 (95.5%) GO terms in molecular function branch and 1,070 of 1,147 (93.3%) GO terms in cellular component branch have greater than 0. In addition, S_d increases with specificity, especially GO terms in cellular component branch. This implies that the compartmentalization of modules may affect the modularity score. Genes or proteins that are active in the same part of the cell tend to have higher functional correlations. Although we did not consider the compartmentalization of modules in this study, it is worth to take this into account in the future work.

3.2.4 Implementation of the algorithm

FLNP starts by choosing a small number of connected seed genes on the FLN without applying any cutoff on edge weight. In each cycle, genes in $N_G(k)$ are moved sequentially to the interior, and S_d is computed. At the end of the cycle, we retain in the interior the gene that has increased the score the most. The algorithm performs a greedy search and stops when no gene in $N_G(k)$ can be found to increase the modularity score (Figure 3-2). Because the algorithm does not guarantee a global optimum, it is possible that the removal of some genes from the final module will increase the score. To determine how large this effect might be, we sequentially removed each gene and recomputed the modularity score. For breast cancer, there is no gene whose removal increases the S_d by more than 0.1. The removal of 22 of 81 genes in the breast cancer module will slightly increase S_d by 0.01 to 0.02. For ovarian cancer, none of the genes in the module decreases S_d by more than 0.1. The removal of 20 of 47 genes in the ovarian cancer module will slightly increase S_d by 0.01 to 0.05. Considering the inherent approximations in the procedure used to obtain the weights, these differences are considered to be within the margin of error of the method. An FLNP interface is available at http://visant.bu.edu/FLNP.

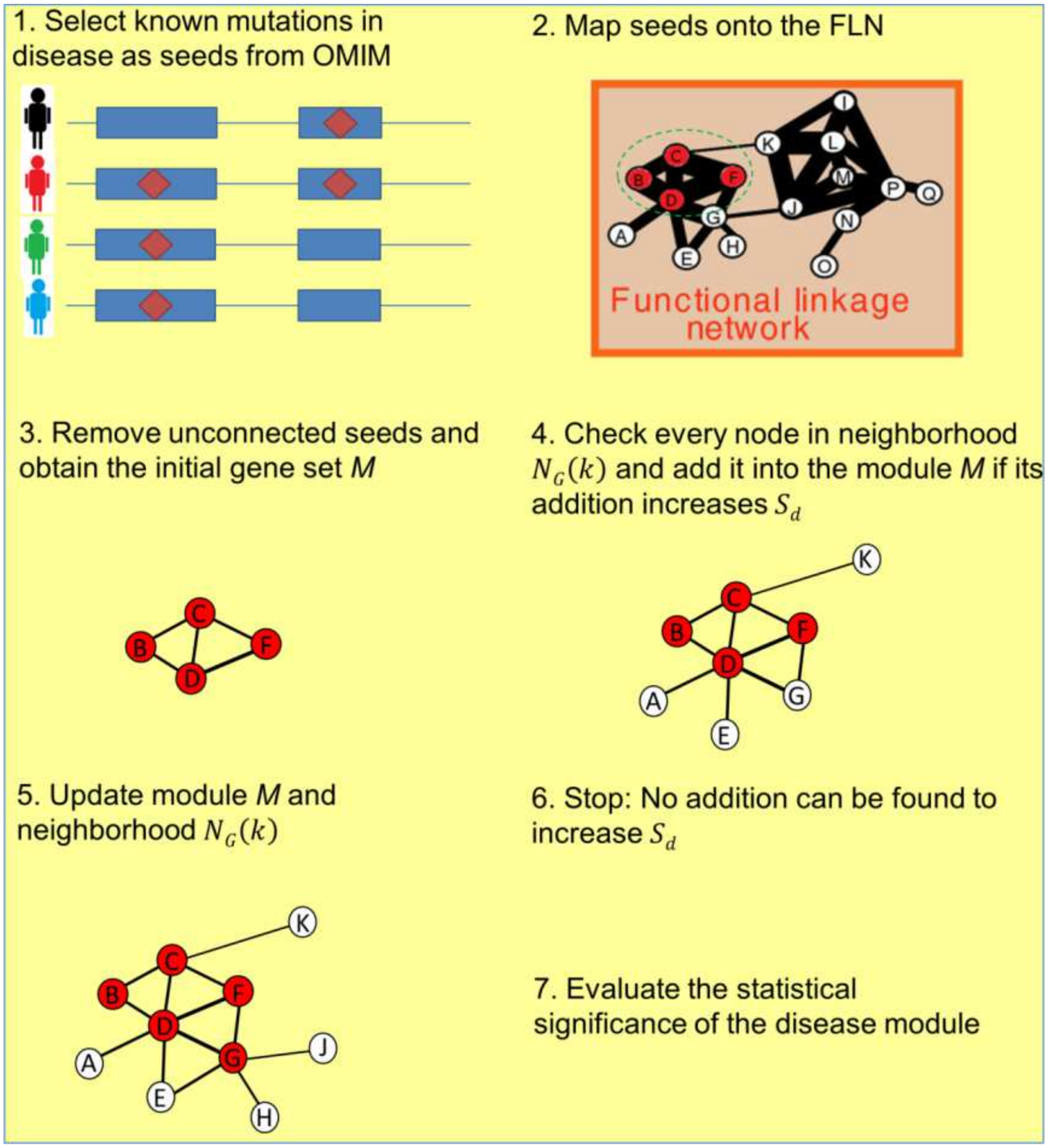

Figure 3-2. Overview of the algorithm.

3.2.5 Statistical significance

We are interested in both modularity and driver context, and use a permutation test to evaluate the significance of each. To assess the significance of the modularity score we used a background distribution of modules based on randomly chosen seeds, with the same number of seed genes as the unique oncogene module. The procedure is repeated 2000 times.

To assess the biological significance of the identified oncogene module we estimate the frequency with which the number of oncogenes found in our oncogene module, would be found in a set of random modules. In each of 1000 iterations, we randomly selected the same number of seed genes used in our oncogene module, excluding oncogenes, and then use the algorithm described above to obtain a module.

3.3 Results and discussion

3.3.1 Case study I – Breast cancer

3.3.1.1 Seed selection

OMIM lists mutations in 22 genes that are possible contributors to breast cancer. A gene may have one or more mutations in breast cancer, but for the purpose of this paper, we are concerned only with the binary classification of the gene. We excluded uncertain mutations (entry with braces, “{}”, in OMIM, which represents mutations that are highly possible disease related mutations but yet to be confirmed) leaving 11 genes that are confirmed to drive breast cancer. After removing singletons (genes that do not connect to any other genes in the seed set), 9 genes remained, and we used all 9 as our seed set (Table 3-1). With θ set to 0.1, FLNP uncovered a breast cancer module with 81 genes (Table S1 Appendix E). A sampling calculation (Methods) indicates that the value attained by target function is well above that of random gene modules of comparable size (p-value = 0.017).

Table 3-1. Seeds set of breast cancer

Gene symbol	Entrez Gene ID	Aberration	References
PPM1D	8493	Amplification	Li *et al.* (2002)
TP53	7157	Mutation (pro151-to-thr and pro151-to-ser)	Carrere *et al.* (1993), Chen *et al.* (1991)
BRIP1	83990	Somatic mutation (pro47-to-ala and met299-to-ile)	Cantor *et al.* (2001)
AKT1	207	Somatic mutation (glu17-to-lys)	Carpten *et al.* (2007)
KRAS	3845	Somatic mutation (gly13-to-asp)	Kozma *et al.* (1987)
PIK3CA	5290	Somatic mutation (his1047-to-arg, his1047-to-leu and glu545-to-lys)	Campbell *et al.* (2004)
RB1CC1	9821	Deletion of exons 3-24 and exons 9-23	Chano *et al.* (2002)
SLC22A18	5002	gain of imprinting	Gallagher *et al.* (2006)
TSG101	7251	loss of heterozygosity	Li *et al.* (1997)

3.3.1.2 KEGG pathways that intersect the module

Pathways that contribute to altered physiology in cancer cells occur at an unusually high rate in our module. In particular we used DAVID [97, 98] to evaluate the overlap between KEGG pathways and the breast cancer module and found nine such KEGG pathways (excluding pathways specifically labeled with the term cancer; *e.g.* the “prostate cancer” pathway) shared a certain fraction of their genes with those in the breast cancer module with false discovery rate (FDR) smaller than 0.05. These included the VEGF signaling pathway (angiogenesis); the Adherens junction, Focal adhesion, and regulation of actin cytoskeleton pathways (metastasis); the erbB signaling, GnRH signaling, MAP Kinase signaling and Progesterone-mediated oocyte maturation pathways (growth); and the apoptosis pathway. Table S2 in Appendix F lists genes that are in the breast cancer module for each cancer-related pathway and the p-value and FDR of each pathway.

3.3.1.3 Oncogene enrichment

Among the 81 genes in the breast cancer module, 72 were not members of the seed set. These can be placed into four categories. (a) Thirty are known oncogenes found in COSMIC and OMIM (Boxes in Figure 3-3). The estimated probability for chance occurrence is smaller than 0.001, which means none of the 1000 random modules has higher concentration of oncogenes than the breast cancer module. This was obtained by using the statistical test we described in the method section. Of the (72 - 30 =) 42 genes which have not been previously identified with breast cancer, we estimate that (42×39/81 =) 20 will be drivers. This point estimate is likely to be conservative since there is independent evidence (discussed below) supporting driver status. (b) Four are in the Genome Association Database (GAD) and implicated by genome wide association studies (GWAS). GWAS and our findings are independent and therefore mutually supportive. (c) A separate analysis performed by our collaborator Gabriela Alexis (unpublished data) indicates that 6 of the remaining 38 have a significantly high non-silent mutation or amplification/deletion rate across 781 breast cancer samples in The Cancer Genome Atlas (TCGA) with an FDR < 0.05. Here, too, our results and those obtained by analyzing TCGA are independent, and therefore mutually confirmatory.
(d) For the remaining 32 genes there is no evidence for a causal role in initiation or progression of cancer, although 12 are in one or another of the 9 cancer-related pathways that overlap the module.

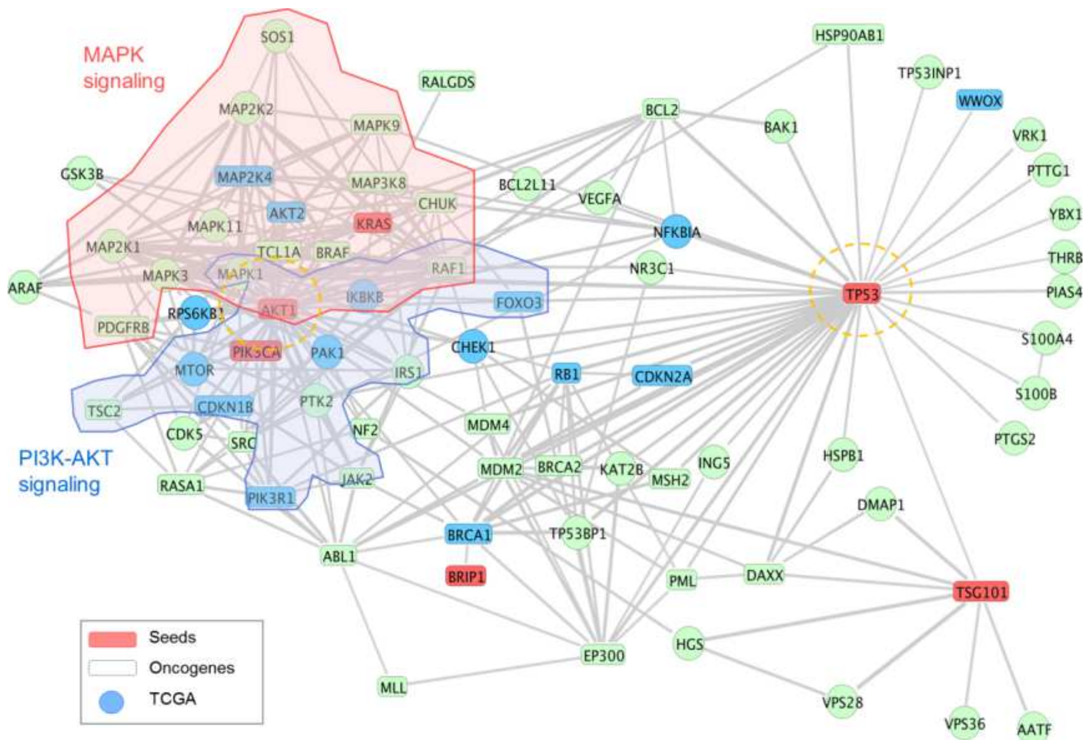

Figure 3-3. Breast cancer module. Boxes are known oncogenes from OMIM and COSMIC. The red nodes are breast cancer drivers obtained from OMIM, which were used as seeds; blue nodes are genes with high mutation frequencies in the TCGA breast cancer population; green nodes are the remaining genes in the breast cancer module. Edges between genes are from the functional linkage network, which represents the functional relations between genes; edge thickness reflects the strength of functional relation.

3.3.1.4 Comparison with other methods

We compared our algorithm with NetBox [75] and with an FLN based method previously developed in our Lab [79]. NetBox identifies disease modules from an integrated literature curated network by starting from a group of altered genes. Using our seed set as the input into NetBox resulted in a breast cancer network with 92 genes, 85 of which were not members of the seed set. Of the 85, 25 (29.4%) were known oncogenes as indicated by OMIM and COSMIC. The breast cancer module identified by FNLP was slightly more enriched in oncogenes (30 of 72, 41.7%). More importantly, to the extent that reproducibility is an important condition for driver candidacy, we note that 23 genes were reported by both methods (Table 3-2).

We used DAVID to find pathways that are enriched in candidate breast cancer drivers identified by NetBox, and found that the nine cancer-related pathways identified by FLNP were also enriched in NetBox (FDR< 0.05).

The third method (historically, the first) [79] is a simple and direct FLN (DFLN) based prioritization; viz., genes are ranked by the sum of the weights of their links to neighboring oncogenes. DFLN is not module based, and provides no guidance on where to threshold for candidacy. In order to compare results we picked the top 72 genes determined by DFLN, the number in our breast cancer module. The results show that FLNP performance is similar to that of DFLN, but with a smaller seed set (Table 3-2).

Of the 12 genes identified by all three methods, 6 are oncogenes - DAXX, EP300, MDM2, RAF1, BCL2 and SRC. For the other six genes that are not listed as oncogenes by either OMIM or COSMIC, there is independent evidence supporting the possibility that they are breast cancer drivers. For VEGFA the evidence comes from GWAS [99]. A separate analysis performed by our collaborator Gabriela Alexis (unpublished data) indicates that CHEK1 and NFKBIA have a significantly high non-silent mutation or amplification/deletion rate across all samples in the TCGA breast cancer cohort (FDR < 0.05). Finally, GSK3B and PTK2 have been implicated in breast cancer related pathways, *e.g.* the ErbB signaling pathway, focal adhesion pathway and VEGF signaling pathway [100–102].

Table 3-2. Number of oncogenes, GAD listed genes and genes with high alteration rate in TCGA breast cancer samples identified by FLNP (A), NetBox (B) and DFLN (C).

	FLNP (A)	NetBox (B)	DFLN (C)	(A,B)	(A,C)	(A,B,C)
Module size (excluding seeds)	72	85	72	23	18	12
Known oncogenes	30 (41.7%)	25 (29.4%)	35 (48.6%)	12	11	6
GAD	10 (13.9%)	11 (12.9%)	12 (16.7%)	5	4	3
TCGA	19 (26.4%)	16 (18.8%)	16 (22%)	6	4	2

Our results significantly overlap the recent findings Stephens *et al.* who constructed a landscape of breast cancer genes from 100 tumors, and identified 40 genes possessing potential driver mutations [103]. 12 of 40 are in our breast cancer module, including (a) genes previously implicated in breast cancer development, *e.g.* TP53, AKT1, BRCA1, RB1 and PIK3CA, (b) novel genes with driver mutations, *e.g.* AKT2 and CDKN1B, and (c) likely drivers found in other cancer types, *e.g.* KRAS and MAP2K4. Notably, all 12 genes are oncogenes associated with least one kind of cancer. The results are summarized in Table 3-3 and Figure 3-3.

Table 3-3. Breast cancer module genes with high mutation rate or amplification/deletion in TCGA breast cancer population and the study by Stephens *et al.*

Gene	Function	Abnormal type in TCGA breast cancer population	FDR	Stephens *et al.*
CDKN1B	Tumor suppressor	Deletion	0.007795	Truncation/ Missense Novel breast cancer driver
CDKN2A	Tumor suppressor	Deletion	5.43E-27	
CHEK1	Protein kinase	Deletion	5.17E-15	
AKT1*	Oncogene	Missense	1.14E-08	Known breast cancer driver

FOXO3	Tumor suppressor	Deletion	6.70E-07	
MTOR	Protein kinase	Deletion	1.32E-24	
IKBKB	Protein kinase	Amplification	1.66E-27	
NFKBIA	Tumor suppressor	Amplification	0.01066	
PAK1	Oncogene	Amplification	5.89E-58	
WWOX	Tumor suppressor	Deletion	3.62E-06	
PIK3CA*	Oncogene	Amplification	2.32E-05	Known breast cancer driver
PIK3R1	Enzyme	INDEL/Missense Deletion	1.13E-10 (Mutation) 2.31E-06 (CNV)	
RB1	Tumor suppressor	Deletion	1.22E-22	Known breast cancer driver
RPS6KB1	Protein kinase	Amplification	1.27E-56	
MAP2K4	Tumor suppressor	INDEL/Missense/Nonsense Deletion	1.13E-10 (Mutation) 6.17E-16 (CNV)	Likely driver
BRCA1	Tumor suppressor	Deletion	0.000214	Known breast cancer driver
AKT2	Oncogene			Missense Novel breast cancer driver
TP53*	Tumor suppressor	Missense	<8.64e-12	Known breast cancer driver
KRAS*	Oncogene			Likely driver

*seed genes

In order to more clearly see relations between genes, we remove weak links (functional linkage weight < 0.2) from the breast cancer module. This results in two major sub-groups (Figure 3-3), one centered on TP53 and the other centered on AKT1. The AKT centered sub-group includes genes in the MAPK signaling pathway and the PI3K-AKT signaling pathway. It is connected to TP53 via a series of mutated genes in breast cancer (blue boxes). Previous studies also proved that the MAPK signaling pathway and the PI3K-AKT signaling pathway is critical to the development of breast cancer and is a possible therapeutic target [104–106]. These subgroups are consistent with common breast cancer subtypes. In particular, the two luminal subtypes are dominated by PI3K oncogene mutations, whereas mutations in P53 and in the MAPK pathway are widely distributed across samples of basal-like tissue.

We also found genes that have not yet been reported as drivers in large scale studies, but for which there is suggestive evidence of association with breast cancer progression. For example, PAK1 has a high mutation rate in the TCGA breast cancer population, though it is not reported as a driver of any cancer in either OMIM or COSMIC. PAK1 is linked to AKT1, RAF1 and some MAP kinases. A recent study carried out by Shrestha *et al.* indicates that PAK1 is a breast cancer oncogene and may cause a predisposition to breast cancer by activating MAPK and MET signaling pathway [104].

Our results included not only known or possible drivers but also novel candidates. BRAF encodes a serine/threonine protein kinase that helps regulate the MAPK signaling pathway. MAPK signaling is related to several important cellular functions, including differentiation and proliferation, and plays an important role in various types of cancer, including breast cancer. Mutations in BRAF are known to drive melanoma, non-small-cell lung cancer, adenocarcinoma and colorectal cancer [107–110], possibly by disrupting phosphorylation of MEK and the activation of ERK in the MAPK signaling pathway[111]. Although it has not been previously classified as a driver of breast cancer in either OMIM or COSMIC, our results, and the known biology of BRAF, as well as its implication in other cancers, taken collectively strongly suggest that it drives breast cancer.

In conclusion, we found that the breast cancer module contains many well-known drivers, and therefore expect that many of the remaining genes are good candidates for drivers. Consequently for this 81 gene module, we recover 30 known oncogenes, and have reasonable *prima fascia* evidence for 11 genes that have not previously been classified as cancer drivers.

3.3.2 Case study II – Ovarian cancer

3.3.2.1 Seed selection

We collected 7 genes that are known to be mutated in ovarian cancer from OMIM. One of them was a singlet, not connected to any of the others, and was therefore removed from the seed set (Table 3-4). With θ set to 0, FLNP converged on a 58 gene module (Figure 3-4 and Table S3 Appendix G). As shown in Methods, the value obtained by the target function is well above that associated with random modules (p-value = 0.0035).

3.3.2.2 Oncogene enrichment

Among the 58 genes in the ovarian cancer module, 52 were not members of the OMIM seed set and 13 are known oncogenes (a non-redundant set from OMIM and COSMIC, excluding the seed genes). The estimated probability for chance occurrence is smaller than 0.001, which means none of the 1000 random modules has higher concentration of oncogenes than the ovarian cancer module. The relatively high, but none the less significant, p-values reflected the correlation structure of the FLN. Of the (52 – 13 =) 39 genes that are not firmly established as oncogenes, we estimate that (39×13/52 =) 10 which have not been previously identified with ovarian cancer, will be drivers. Among the oncogene candidates in the module, BRAF and PIK3CA are reported as rare but important mutations in the previous study carried out by the TCGA network [50]. PIK3CA, PIK3R1, EGFR and ERBB2 are implicated as oncogenes in ovarian cancer [112].

Table 3-4. Seeds set of ovarian cancer

Gene symbol	Entrez Gene ID	Aberration	References
CTNNB1	1499	Somatic mutation (ser37-to-cys missense)	Sagae *et al.* (1999), Wright *et al.* (1999),
ERBB2	2064	somatic 2570A-G transition in the ERBB2 gene that caused an asn857-to-ser (N857S) substitution	Cancer Genome Project and Collaborative Group (2004)
AKT1	207	Somatic mutation (glu17-to-lys)	Carpten *et al.* (2007)
PIK3CA	5290	Somatic mutation (his1047-to-arg, gln546-to-lys and glu545-to-lys)	Campbell *et al.* (2004)
RRAS2	22800	Somatic mutation (leu72-to-gln)	Chan *et al.* (1994)
CDH1	999	Somatic mutation (ser838-to-gly) Somatic loss of heterozygosity	Risinger *et al.* (1994)

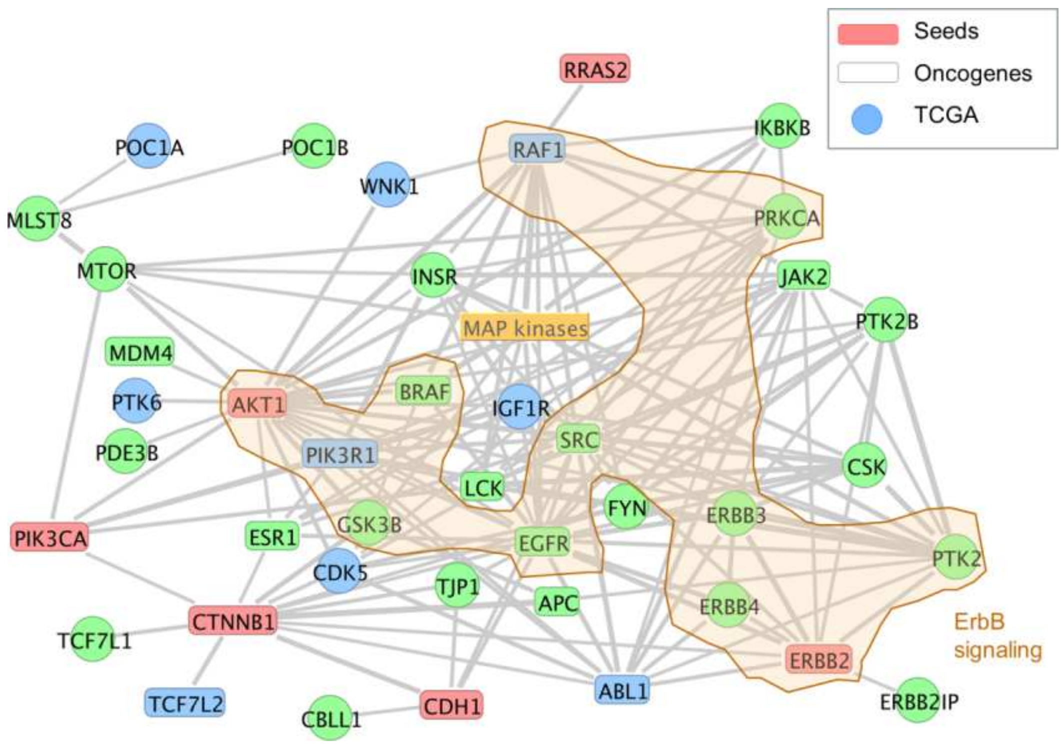

Figure 3-4. Ovarian cancer module. Boxes are known oncogenes from OMIM and COSMIC. The red nodes are seeds; the blue nodes are genes with high mutation frequencies in the TCGA ovarian cancer population; the green nodes are rest of genes in the ovarian cancer module. Edges between genes are from the functional linkage network, which represents the functional relations between genes, and the thickness of the edge reflects the strength of functional relation between two genes.

A separate analysis performed by Gabriela Alexis (unpublished data) indicates that 5 of the remaining 39 have a significantly high non-silent mutation or amplification/deletion rate across all samples in the TCGA (FDR < 0.05). Our results and those obtained by analyzing TCGA are independent, and therefore mutually confirmatory.

3.3.2.3 KEGG pathways that intersect the module

For the remaining 34 genes, there is no supporting evidence for a causal role in initiation or progression of cancer, although 19 are in one or another of the 9 cancer related pathways described in the previous section. Not surprisingly, the nine cancer-related KEGG pathways are enriched in ovarian cancer module genes (Table S4 Appendix G) with FDRs ranging from 0.0004 (apoptosis pathway) to 10^{-23} (ErbB signaling pathway). As suggested in a number of reports, ErbB (EGFR), ERBB2 (HER2) and several members of EGFR family are overexpressed or altered in many cancers, including ovarian cancer [113]. Our result also suggests that members of the ErbB signaling pathway and the EGFR family are tightly connected to other oncogenes in the ovarian cancer module and may play critical roles in the development of ovarian cancer.

3.3.2.4 Comparison with other methods

We used the same seed set as the input for NetBox. The resulting module includes 97 genes, 92 of which were not members of the seed set. 28 of them (30.7%) were known oncogenes as indicated by OMIM or COSMIC, whereas the ovarian cancer module identified by our method has 13 oncogenes (25%). At a rank cutoff of 52 (the size of our ovarian cancer module), DFLN had greater specificity than the other two methods (Table 3-5).

We used DAVID to find pathways that are enriched in candidate ovarian cancer drivers identified by NetBox. Eight of the nine cancer-related pathways predicted by FLNP were also predicted by NetBox (FDR < 0.05). The exception is the Progesterone-mediated oocyte pathway.

Nine genes were found by all three methods, including six oncogenes – EGFR, ESR1, PIK3R1, RAF1, LCK and SRC. Among them only PIK3R1 is a well established driver. There is independent evidence supporting the candidacy of the other eight genes. GSK3B is involved in the Wnt and EGFR signaling pathways, cell migration, hyper-methylation of CpG island, ER overload response, negative regulation of apoptosis and negative regulation of MAPK activity, which is central to cellular transformation. ERBB3 over expression in a significant proportion of ovarian cancers is correlated with proliferation, differentiation and poor prognosis [114–116]. PTK2 (also known as FAK) is a focal adhesion kinase and is involved in early step of cell growth. The activation of PTK2 is prevalent in ovarian cancers and is related to cell growth, migration, and progression [117, 118]. Halder *et al.* showed that a PTK2 (FAK) inhibitor can suppress ovarian cancer growth *in vitro* and *in vivo* [119].

Table 3-5. Number of oncogenes, ovarian cancer associated genes and genes with high alteration rate in TCGA ovarian cancer samples identified by FLNP (A), NetBox (B) and DFLN (C).

	FLNP (A)	NetBox (B)	DFLN (C)	(A,B)	(A,C)	(A,B,C)
Module size (excluding seeds)	52	91	52	14	17	9
Known oncogenes	13 (25%)	28 (30.7%)	30 (57.7%)	8	9	6
GAD	1 (1.9%)	6 (6.6%)	8 (15.4%)	1	1	1
TCGA	9 (17.3%)	14 (15.4%)	14 (26.9%)	2	5	2

Figure 3-4 and table 3-6 show that genes with high mutation rate in the TCGA ovarian cancer cohort (blue nodes) have strong connections with seed genes (red nodes) and other oncogenes (boxes). The result shows that the ErbB signaling pathway is critical in the ovarian cancer module. Several studies have reported that the ErbB signaling pathway is critical in development of many cancer types, including ovarian cancer [120–122]. We found a number of genes that have not been connected to ovarian cancer by large scale studies, but for which there is strong evidence implicating them in ovarian cancer. For example, the tumor suppressor gene IGF1R has been suggested as a potential therapeutic target for ovarian cancer by several studies [123–125]. The protein tyrosine kinase PTK6, which functions as an intracellular transducer in epithelial tissues, is another candidate driver for ovarian cancer. Ludyga *et al.* proved that PTK6 directly interacts with ERBB2 and knockdown of PTK6 will decrease activation of ERBB2 in breast cancer cell lines [126]. It is reasonable to presume that PTK6 may have similar interaction with ERBB2 in ovarian cancer. Irie *et al.* further showed that PTK6 can form a complex with IGF1R and modulate anchorage-independent survival of breast cancer and ovarian cancer [127]. Our result indicates that PTK6 is tightly connected to AKT1, which is a driver of ovarian cancer. Thus, we speculate that PTK6 contributes to the development of ovarian cancer via interacting with AKT1, IGF1R and/or ERBB2.

Considering both breast cancer and ovarian cancer, the FLNP algorithm and Netbox showed comparable performance. In addition, although they predicted a common core set of candidates, many of the candidates they predict are complementary. It would therefore appear that there is some gain in using both methods.

Table 3-6. Ovarian cancer module genes with high mutation rate or amplification/deletion in TCGA ovarian cancer population.

Gene	Function	Abnormal type in TCGA ovarian cancer population	FDR	Comments
PTK6	Protein kinase	Amplification	0.000195	
CDK5		Amplification	0.000152	
WNK1	Protein kinase	Amplification	1.25E-06	
IGF1R	Tumor suppressor	Amplification	3.26E-06	Potential therapeutic target [123–125]
AKT1*	Oncogene	Amplification	0.042123	Carpten *et al.* (2007) [128]
WDR51A		Deletion	0.001159	
RAF1	Tumor suppressor	Deletion	0.000802	
ABL1	Tumor suppressor	Deletion	5.70E-10	Le *et al.* (2011)[129]
CDH1*	Tumor suppressor	Deletion	0.000304	Risinger *et al.* (1994) [130]
PIK3R1		Deletion	1.55E-66	Bast *et al.* (2009) [112]
TCF7L2	Protein kinase	Deletion	1.14E-05	

*Seed genes

3.5 Conclusions

We provide a quantitative formulation of the idea that oncogenes clustered on an evidence weighted functional linkage network are embedded in a field of genes and pathways that are likely to be associated with cancer cell physiology. More specifically we present a method for determining the boundary of the field, thereby obtaining a module that is functionally rich in drivers, and enabling the identification of new oncogene candidates. We estimate that 20 of the 42 and 10 of the 39 module genes that have not been previously identified with breast and ovarian cancer respectively, will be drivers. The results also reveal the detailed interrelations within and among the three tiered information hierarchy of genes, networks and modules underlying these two forms of cancer.

Chapter 4. Expansion of KEGG pathways

4.1 Introduction

Pathways describe the interactions between a group of genes, enzymes and/or molecules that carry out a particular cellular task, *i.e.* apoptosis, fructose metabolism or focal adhesion, or contribute to a specific disease phenotype. Pathways are routinely used in so-called gene set analysis to detect which pathways are perturbed in a disease [23, 24, 131, 132]. Many pathways are manually curated, which allows for highly confident and clearly annotated functions of pathways and components in each pathway [29, 32, 33, 133]. Because of this, however, these pathway annotations are limited to both incomplete current biological knowledge and the requirement of substantial manpower. Due to this, manually curated pathways are usually small and the coverage is low. Figure 4-1 shows that most current manually curated human pathways can only cover up to around 7,000 genes, which is about one-third of the human genome, and the average size of pathways are from 7 to 81.

Another issue plaguing manually curated functional pathways is the lack of standard criteria when collecting and translating the data from literature. Therefore, the same pathway may have different elements in different databases. For example, the well-known citric acid (TCA) cycle has 31 genes in KEGG and 19 genes in Reactome, yet has only 8 genes in BioCarta. Seven of these genes are treated as members of TCA cycle in all three databases but many more genes are assigned to it by only one or two databases.

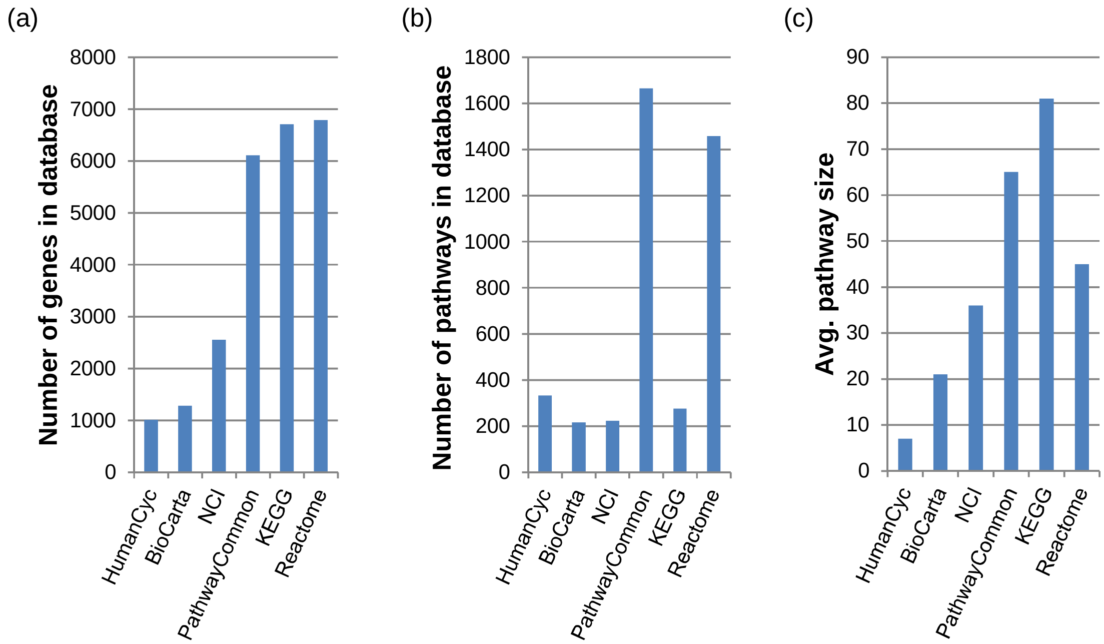

Figure 4-1 Statistics of current manually curated human pathway databases.

The advent of high-throughput experiments enables the measurement of a great number of physical interactions and functional associations between proteins, genes, and small molecules. Many computational approaches have been developed to identify functional modules or disease modules from interaction maps such as protein-protein interaction (PPI) networks, gene regulation networks, or functional linkage networks (FLNs) [49, 86, 87, 134–138]. A number of studies demonstrated that network-based approaches can successfully detect genes known to be associated to a disease or participate in a given cellular function as well as suggest novel candidate genes for the same [5, 75, 77]. The module detection algorithms can be classified into two major categories: divisive methods and agglomerative methods. Divisive methods try to partition graphs into regions of differing connectivity and optimize the modularity, while agglomerative methods try to grow modules from densely interconnected regions. Most of divisive methods need to define a modularity function and try to find a best partition that can optimize the modularity, which is the fraction of edges found to be within modules minus the expected fraction if edges are randomly placed [139, 140]. A good network partition must comprise many within-module edges and as few as possible between-module edges.

However, divisive methods have several drawbacks. Frist, divisive methods usually perform a greedy optimization and require complete knowledge of the entire network. If a network is huge, divisive method could be very time consuming. Second, divisive methods aim to achieve global optimization and attempt to assign all nodes in a network into some cluster. This could result in many false positives if we do not have comprehensive knowledge of the entire network. Third, it only considers network topology and most of divisive methods cannot deal with weighted networks. Also, divisive method partitions a network into several subnetworks, so it cannot allow overlapped modules. Biological modules are usually highly overlapped, which means a gene or protein may participate in several similar or different functions (modules). Therefore, those methods may not be suitable for detecting functional modules. To address this issue a number of agglomerative algorithms have been proposed to detect overlapped modules on a biological network, *e.g.* MCODE, HUNTER, SPICi and ClusterOne [141–146]. Unlike divisive methods, agglomerative methods aim to find local dense regions on a network. These methods predicted a number of functional modules; however, due to the limitation of current knowledge, it is difficult to annotate the resulting functional modules and therefore constrains their applications.

In contrast to most of the previous studies that aimed at searching for all possible topological modules or functional modules on a network, we focus on suggesting new members for existing pathways. In this study we propose a context-based locally optimal algorithm that starts from a small set of pathway genes as seeds and followed by searching for functionally related genes on the functional linkage network. We use the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database as an example and select genes from KEGG pathway database release 64.0 as seeds. We demonstrated that our algorithm can successfully identify genes participating in a particular pathway with evidence from KEGG pathway database release 68.0 and other pathway databases, as well as corresponding Gene Ontology (GO) annotations.

4.2 Materials and Methods

4.2.1 KEGG pathway database

KEGG pathway database is one of the most popular pathway annotation databases. KEGG updates its database about five times every year. In its Release 64.0 on October 1, 2012, KEGG has 206 human pathways and 4,794 genes, where in the Release 68.00 published one year later (October 1, 2013), there are 276 human pathways in the KEGG pathway database covering 6,715 genes. In order to better evaluate the performance of our algorithm, we expand pathways in KEGG Release 64.00 and compare our results with the latest Release 68.00 as well as other pathway databases (HumanCyc, NCI pathway database, BioCarta, Reactome and Pathway Commons) and GO.

4.2.2 Functional Linkage Network

Module identification methods have been applied on different types of networks, *e.g.* PPI networks, co-expression networks, gene regulatory network, and FLNs [49, 86, 87, 134–

138]. The FLN integrates a variety of genomic data and represents the functional connections between genes (nodes) by evidence weighted links [79, 147]. In the FLN we used in this study, the link weights were established by integrating 16 genomic features using a Naïve Bayesian classifier [147]. Previous studies have shown that such integrated FLNs can provide more reliable information than a single source network [148, 149]. Notably, the sources used to construct the FLN did include any of the pathway databases used in this study. Furthermore, the FLN used in this study was constructed in 2009, which means the information used to improve KEGG Release 64.00 to Release 68.00 is not included in the FLN.

4.2.3 Implementation of algorithm

4.2.3.1 Expansion of pathway modules

In this study we modified the module identification algorithm, FLNP, which we described in Chapter 3. In short, for a module $M = \{k\}$, consisting of k genes, the set of neighboring exterior genes $N_G(k)$, we define a modularity score, Sd as

$$Sd = \frac{1}{2p}\sum_{i,j}\omega_{i,j}[i \in M][j \in M] - \frac{1}{q}\sum_{i,j}\omega_{i,j}[i \in M][j \in N]$$

where ω is the weighted adjacency matrix, p is total number of edges within the module M, and q is the total number of edges between module M and the neighborhood community $N_G(k)$. The element $\omega_{i,j}$ in the weighted adjacency matrix ω represents the weight of the edge between genes i and j.

We expect the modularity scoring function we defined should fit to well-defined functional modules, such as pathways. Therefore we calculated the modularity score of 245 KEGG pathways and found that 242 of 245 (98.8%) KEGG pathways have a modularity score greater than 0, with an average Sd of 1.27 (Figure 4-2).

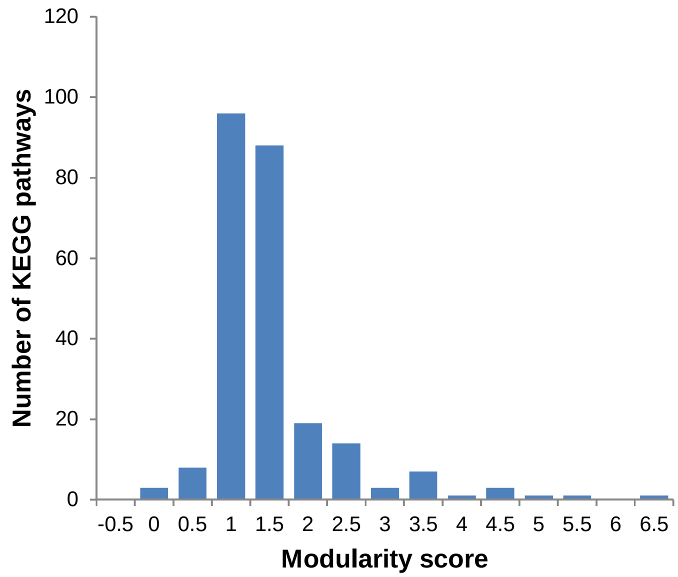

Figure 4-2 Distribution of modularity score Sd of 245 KEGG pathways, retrieved April 2013.

FLNP starts by selecting a small number of connected seed genes on the FLN without applying any cutoff on edge weight and recruits genes for maximizing the target function Sd , subject to the constraint that it must be greater than θ. θ is the threshold for defining a module, where we set θ as 0.1 for getting better specificity. Our previous work has shown that with a given context, *e.g.* oncogenes for a particular cancer types, FLNP can successfully identify group of genes that are enriched in that context (Details are described in Chapter 3). In this study, our goal is to expand an existing KEGG pathway, therefore we choose a small number of connected genes from a particular KEGG pathway as seeds and use FLNP to identify candidate genes for additional pathway membership.

The sensitivity and specificity of the identified modules are correlated to the context of input seeds, which is again correlated to the number of input seeds. The larger number of input seeds will result in a stronger context. We used the KEGG cell cycle pathway as an example to demonstrate the effect of the input seeds context on the resulting module. First, we divided the cell cycle pathway into two parts, a training set that contained 89 genes and a test set that contained 30 genes. To evaluate the effect of the number of seeds, we applied our algorithm on the training set with different number of seeds (seed number = 1, 2, 3 … 89) 100 times each, and then calculated the average sensitivity and specificity for each seed set. Figure 4-3 shows the relationship between the number of seeds and the specificity and sensitivity of the identified modules. As the number of seeds is reduced, there is a corresponding increase in sensitivity but a decrease in specificity. To obtain a balance between sensitivity and specificity, we take into account the pathway size when selecting the seeds and use a weighted factor λ to adjust the number of seeds of pathways with different size.

$$SN_p = \left\lfloor \frac{N_p}{\lambda} \right\rfloor$$

, where SN_p is the number of seeds for the pathway P and N_p is the size of the pathway P. λ is a weighted factor and varied with N_p, where

$$\lambda = 4 \times \left\lceil \frac{N_p}{100} \right\rceil$$

We iteratively choose SN_p genes from a pathway P as seeds and attempt to detect a candidate module for 500 times.

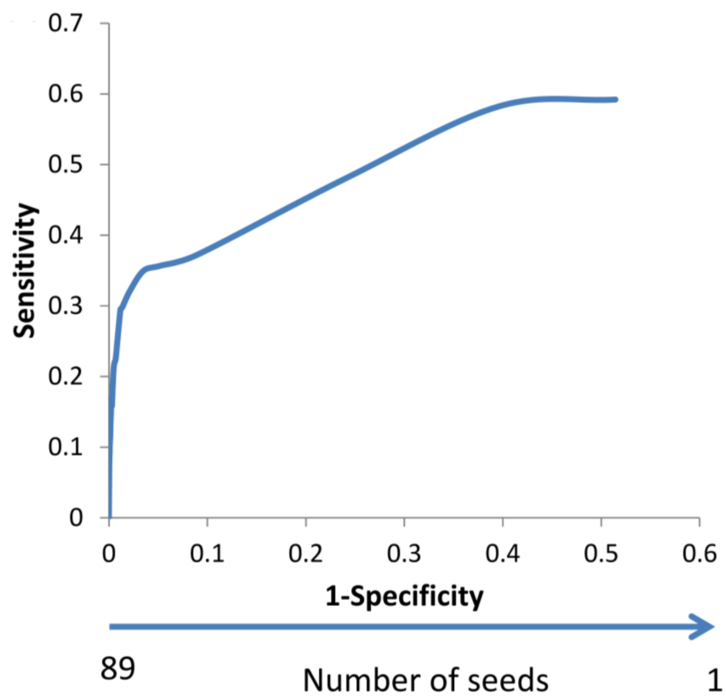

Figure 4-3 Sensitivity and specificity with different context of seed genes. We used the KEGG cell cycle pathway as an example to demonstrate the effect of input seeds context on the resulting module. The specificity increases with the seed size while the sensitivity behaves in the opposite way.

4.2.3.2 Merge pathway modules

Each of the 500 modules we obtained by applying our algorithm contains genes that are associated to part of the target pathway. Here we use a majority rule to merge the modules. More specifically, we count the number of times a gene has been elected by our algorithm as a candidate for a particular pathway. Genes that are selected by our algorithm more than *n* times will be included in the final module.

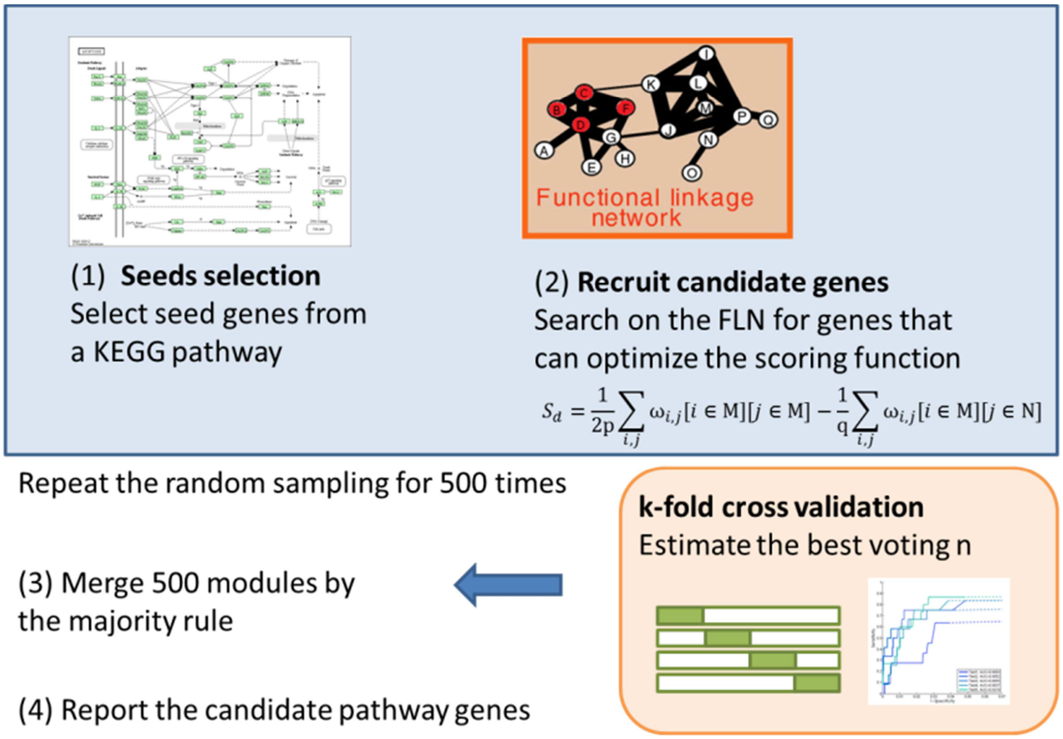

Figure 4-4 Overview of the algorithm.

Defining the threshold *n* is an important task. Here we use *k*-fold cross validation to determine the best threshold *n*. For a particular pathway, we partition the pathway into *k* equal size subgroups. Of the *k* subgroups, a single subgroup is retained as validation data and the remaining *k-1* subgroups are used as training data. The number of subgroups is set to five for pathways with size smaller than 100 and set to ten for pathways with size larger than 100. We selected the threshold *n* at the point of inflection, which is widely used for determining the number of clusters [150]. The median *n* of all cross-validation results is then used. Figure 4-4 illustrates the steps of our algorithm.

4.3 Results and discussions

4.3.1 Comparison with other methods

We applied our algorithm on 105 KEGG pathways in KEGG release 64.00 excluding (a) disease pathway because they are not functional modules and (b) pathway with genes fewer than 20 because those pathways cannot provide enough context for expansion. Our algorithm is able to expand 100 KEGG pathways and the expanded pathway size is from 32 to 826. 48 of 100 expanded KEGG pathways contain genes that are added into the same pathway in KEGG release 68.00. In total, 6,272 genes were covered in our results.

We performed a density-based network clustering (*k*-means) on the FLN and found 33 clusters on the FLN. We used a hypergeometric test to find the best matching KEGG pathway for each cluster. The sizes of the identified clusters were from 3 to 320 and the average size was 51. On average, 64% of the cluster genes overlapped with the matched KEGG pathway. This high overlap rate implies that modular structures in the FLN we used indeed represent certain biological function. Although the simple density-based method can identify functional clusters that are highly overlapping with KEGG pathways, it can only match to 26 KEGG pathways and expand 23 KEGG pathways.

We also compared our results with a comparative method, PathExpand, which extends pathways by including densely interconnected interaction partners on a PPI network [151]. PathExpand is able to expand 159 BioCarta pathways, 90 KEGG pathways and 52 Reactome pathways. In the following example, we demonstrated the results of our algorithm on a KEGG metabolic pathway, glycolysis, and a KEGG signaling pathway, apoptosis, and compared our predictions with PathExpand.

4.3.2 Expansion of KEGG glycolysis pathway

Glycolysis is a metabolic pathway that describes the process of converting glucose into pyruvate. In each of 500 iterations, we randomly selected 16 seed genes from the 64 genes of the glycolysis pathway in the KEGG pathway database release 64.0. Then, we partitioned the glycolysis pathway into five equal size subgroups and performed the cross-validation procedure. Figure 4.5 shows the receiver operating characteristic (ROC) curves of the cross-validation experiments and indicates that our algorithm can successfully recover genes in the test data with high sensitivity and specificity.

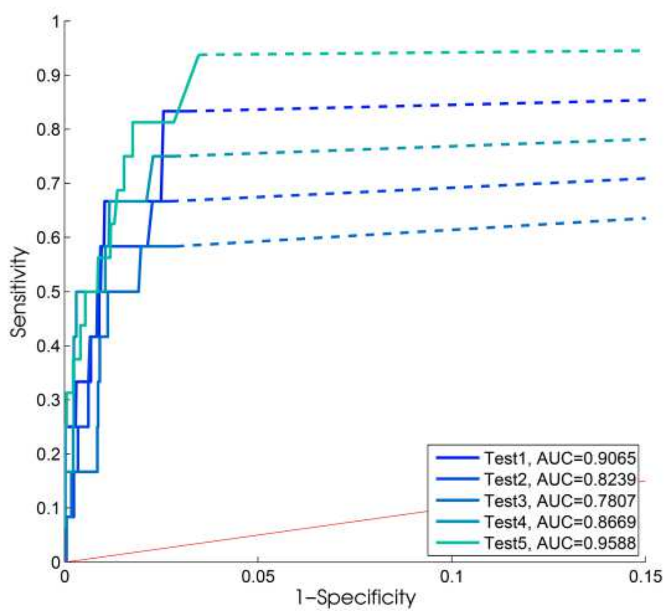

Figure 4-5 ROC curves of cross-validation of glycolysis pathway. AUC is the area under curve.

From cross-validation we estimate that the best threshold *n* for the glycolysis pathway is 30. In total, our algorithm reported 20 genes as candidate members for inclusion in the glycolysis pathway. Among these 20 genes, PDHX and GAPDHS are included in the six new genes for the glycolysis pathway in the latest KEGG pathway database release 68.0.

Five genes, EDARADD, OGDH, MDH2, MDH1 and OGDHL, are not yet included in any pathway database but are members of the gene ontology (GO) term “glycolysis” (GO:0006096). The expanded KEGG pathway is shown in Figure 4.6.

Our algorithm not only included genes that have strong evidence from other manually curated databases supporting their role in glycolysis but also suggested novel genes and possible interactions. For example, transketolase (TKT) is an enzyme that connects the pentose phosphate pathway to glycolysis [152]. Our results suggested that TKT may contribute to glycolysis via the interactions with TPI1, PDHX, OGDH and FBP1. In a very recent study, Bogorad *et al.* constructed a non-oxidative glycolysis (NOG) pathway both *in vitro* and *in vivo* in *Escherichia Coli* [153]. The NOG pathway provides a more efficient way for carbon conservation in sugar metabolism and is important for carbon management. NOG consists of eight core enzymes including F/Xpk, Tkt, Tal, Rpi, Tpi, Fba, Fbp and Rpe, which matches to our hypothesis of the role of TKT in glycolysis.

PathExpand suggested only two new genes, PDK3 and PDK4, to the KEGG glycolysis pathway with default settings. However, these two genes were not found in our results, nor were they found in any glycolysis related pathways in other pathway databases, *e.g.* HumanCyc, NCI pathway interaction database, BioCarta, Reactome, or the “glycolysis” GO term.

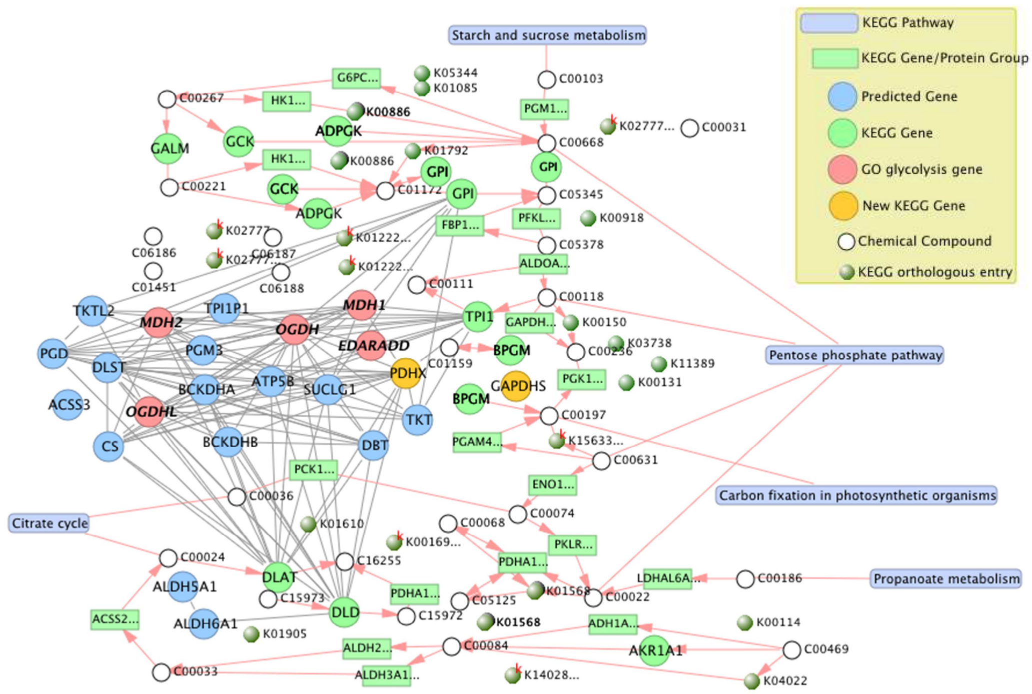

Figure 4-6 Expanded KEGG glycolysis pathway. Green nodes are KEGG glycolysis pathway genes. Blue nodes are candidate glycolysis genes identified by our algorithm. Orange nodes are glycolysis genes identified by our algorithm which were added into the glycolysis pathway in KEGG release 68.0. Red nodes with bold text are GO glycolysis genes identified by our algorithm. Pink links are interactions in the KEGG glycolysis pathway and gray links are interactions retrieved from the FLN. In order to clearly show the interactions between genes, we remove weak links on the FLN (normalized linkage weight < 0.01).

4.3.3 Expansion of KEGG apoptosis pathway

Apoptosis is the process of cell death that occurs naturally in living cells. The malfunction of apoptosis may cause atrophy or uncontrolled cell proliferation, which is very common in cancer development. We randomly selected 22 seed genes from the 89 genes present in the apoptosis pathway in the release 64.0 of KEGG. Again, the apoptosis pathway was partitioned into five equal size subgroups and performed the cross-validation experiments. Figure 4-7 shows the ROC curves of the cross-validation experiments and indicates that our algorithm can successfully recover genes in the test data with very high specificity and decent sensitivity.

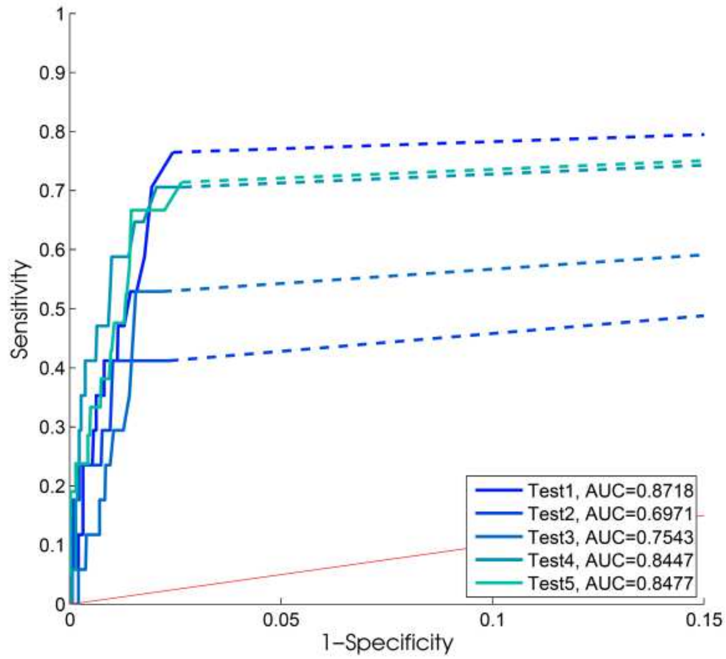

Figure 4-7 ROC curves of cross-validation of apoptosis pathway. AUC is the area under curve.

Our cross-validation experiments estimated that the best threshold *n* for the apoptosis pathway is 3. Our algorithm reported 13 new genes as candidate members for the apoptosis pathway. Although we did not identify the three new members of the apoptosis pathway of KEGG release 68.0, other pathway databases provide evidence to support our findings. We collected pathways from several widely used manually curated databases, including HumanCyc [154], NCI pathway interaction database [155], BioCarta [156], Pathway commons [157], and Reactome [158]. We found that six of the 13 genes we identified were suggested to participate in apoptosis by at least one of those pathway databases (Table 4-1). One of the remaining seven genes is involved in the “apoptotic process” GO term (GO:0006915). The expanded KEGG pathway is shown in Figure 4-8.

Table 4-1 Predicted genes for the KEGG apoptosis pathway with external evidence.

Gene symbol	Entrez Gene ID	References
TRAF6	7189	BioCarta, NCI
RAF1	5894	BioCarta
BAK1	578	Reactome, BioCarta
PARP1	142	BioCarta, NCI
NGFR	4804	NCI
PDPK1	5170	BioCarta, NCI
MCL1	4170	Apoptotic process (GO:0006915)

We found a number of genes that had not yet been assigned to the apoptosis pathway by any of the manually curated pathways or GO, but for which there is strong evidence implicating them in the apoptosis pathway. PI3K/AKT/mTOR pathway is an important signaling pathway in cancer, which is activated by insulin signaling pathway and contributes to cell proliferation, regulation of protein synthesis and apoptosis [159, 160]. Previous studies suggested that PI3K activates AKT, then AKT regulates apoptosis by phosphorylating BAD, FKHR, and NF-κB, or effects a growth transition by phosphorylating mTOR [159]. Our result suggested that the mammalian target of rapamycin (mTOR) may regulate apoptosis as well. In some cancers, like myeloma, rapamycin can induce apoptosis in cancer cells by activating apoptosis [161, 162]. The mechanisms leading to cancer cell death in response to rapamycin treatment, however, are not yet clear. Nevertheless, it is known that rapamycin can inhibit mTOR and induce apoptosis [163], implying that mTOR may be an apoptosis inhibitor. Some studies have proposed that mTOR inhibits apoptosis via its downstream target S6K, which then phosphorylates and inactivates the pro-apoptotic molecule BAD [162, 164, 165]. Figure 4-8 shows that mTOR interacts with PDPK1 and AKT3, which in turn influence BAD, which agrees with the above assumption.

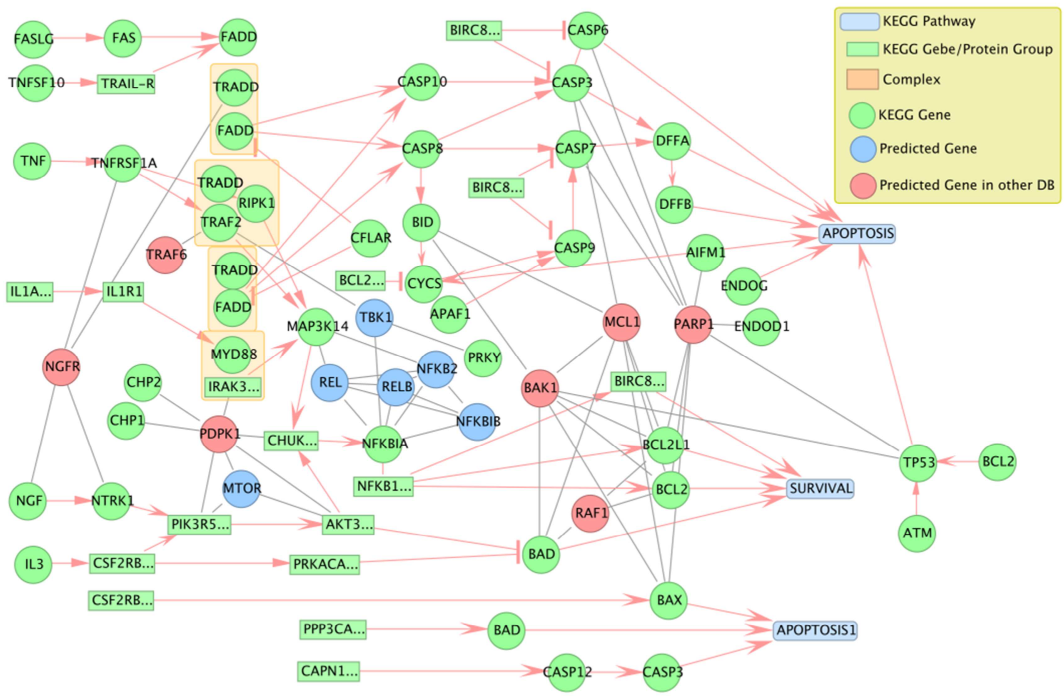

Figure 4-8 Expanded KEGG apoptosis pathway. Green nodes are the KEGG apoptosis pathway genes. Blue nodes are candidate apoptosis genes identified by our algorithm. Red nodes with bold text are apoptosis genes that identified by our algorithm with evidence from other pathway databases. Pink links are interactions in KEGG apoptosis pathway and Grey links are interactions retrieved from the FLN. In order to clearly show the interactions between genes, we remove weak links on the FLN (normalized linkage weight < 0.2).

PathExpand predicts 39 new genes to add to the KEGG apoptosis pathway. Only one gene, NFKB2, was found by both PathExpand and our method. Among the 39 genes predicted by PathExpand, only three genes have evidence from other pathway databases which supports their roles in apoptosis. The results show that applying a module based algorithm on a functional linkage network can strongly improve both sensitivity and specificity.

4.4 Conclusions

We adapted a module identification algorithm to the problem of expanding manually curated pathways. By starting from a small seed set of a particular pathway, our algorithm searched on the FLN and recruited genes that may contribute to that pathway. The cross-validation experiments show that our method can recover genes in the test data with high specificity and good sensitivity.

The results of the expansion of three KEGG pathways further show that our method can successfully recover genes with strong evidence supporting their roles in a particular pathway as well as provide novel additional candidates for that pathway. In summary, we demonstrate that (a) genes in the same pathway tend to be clustered on the FLN, and (b) our functional weighting based algorithm can detect those clusters on the FLN and has a better performance than the density based method. Our results can be used as an alternative option for gene set enrichment analysis, and it can help to improve the coverage of manually curated pathways.

We have shown that applying our module identification method in the FLN can successfully identify driver modules in cancer (Chapter 3) as well as functional modules (*e.g.* pathways or other ontologies). Because our module identification method is context-based, the context of the input seed set is strongly correlated to the outcome. Therefore, one can use the algorithm to identify functional modules in any genomic area by selecting seeds with related context.

Appendices

APPENDIX A

Regulator	FDR	p-value	ave. corr. (regulatees)	fraction	ave. corr. (regulator-regulatees)	# of coherent regulatees
MYC	0.0058	0.0004	0.166	0.0578	0.1161	375
NFE2L2	0.0032	0.0002	0.159	0.0614	0.2215	277
PPARG	0.0018	0.0001	0.167	0.0667	0.2582	254
IL13	0.0032	0.0002	0.1662	0.0636	0.0733	252
MYCN	0.0018	0.0001	0.175	0.0721	0.1339	239
IL10	0.0045	0.0003	0.1628	0.0643	0.2825	213
CDKN1A	0.0018	0.0001	0.1668	0.0695	0.1755	208
PPARA	0.0119	0.001	0.1514	0.0602	0.2197	206
CSF2	0.0018	0.0001	0.1783	0.0829	0.1437	194
IL5	0.0018	0.0001	0.172	0.0757	0.0615	190
IL2	0.0045	0.0003	0.1692	0.0678	0.0636	190
ETS1	0.0378	0.0045	0.163	0.0593	0.1618	175
CLOCK	0.0018	0.0001	0.1786	0.0777	0.102	164
VEGFA	0.0252	0.0027	0.1699	0.0627	0	164
NRIP1	0.0018	0.0001	0.2416	0.1812	0.2552	149
BRCA1	0.0359	0.0042	0.168	0.0626	0.1602	143
TNFSF11	0.0018	0.0001	0.1753	0.0853	0.0963	139
CXCL12	0.0421	0.0055	0.1639	0.0621	0.1294	139
LGALS3	0.0415	0.0053	0.1647	0.0616	0.1098	137
CDKN2A	0.0374	0.0044	0.1728	0.0649	0.139	131
DMD	0.0018	0.0001	0.1843	0.0853	0.1827	128
IRF7	0.0032	0.0002	0.1786	0.0729	0.2607	119
INS	0.0101	0.0008	0.1712	0.0691	0.1381	118
CSF1	0.0018	0.0001	0.1822	0.0892	0.2073	115
XBP1	0.0018	0.0001	0.2056	0.1134	0.2462	112
GK	0.0032	0.0002	0.1748	0.0802	0.1459	110
KLF1	0.0489	0.0068	0.1595	0.0652	0.2596	105
HHEX	0.0288	0.0032	0.1685	0.0672	0.193	104
PPARGC1A	0.0018	0.0001	0.2055	0.1348	0.1537	102
PDX1	0.0018	0.0001	0.1853	0.0977	0	98
NEUROD6	0.0415	0.0053	0.1668	0.0669	0.1756	92

SREBF1	0.0018	0.0001	0.1881	0.1059	0.2032	86
ESRRA	0.0018	0.0001	0.1855	0.1109	0.1303	77
IL3	0.0252	0.0027	0.1667	0.0729	0.0486	77
CSF3	0.0112	0.0009	0.1715	0.0769	0.1805	76
ACVR2A	0.0333	0.0038	0.1798	0.0724	0.122	75
IRF5	0.0045	0.0003	0.1805	0.0837	0.1079	74
KITLG	0.0402	0.005	0.1683	0.0705	0.1059	74
INSR	0.0018	0.0001	0.2271	0.151	0.0902	73
CCL3L3	0.0378	0.0045	0.1789	0.0724	0	73
CD40LG	0.0018	0.0001	0.2082	0.1311	0.2672	70
KDM5B	0.0252	0.0027	0.1767	0.0734	0	69
MEIS1	0.0479	0.0066	0.1532	0.0687	0.1427	68
STAT6	0.0395	0.0048	0.1675	0.0725	0.0899	67
IFNA1	0.0101	0.0008	0.1812	0.0843	0.0368	63
SPI1	0.0018	0.0001	0.2084	0.138	0.1183	61
THAP1	0.0018	0.0001	0.223	0.1629	0.2036	59
SREBF2	0.0032	0.0002	0.1895	0.1066	0.1921	54
IL15	0.0101	0.0008	0.1795	0.0881	0.2449	53
FGFR3	0.0499	0.007	0.1732	0.0746	0.0882	53
FAS	0.0018	0.0001	0.226	0.1645	0.1692	50
NRF1	0.0018	0.0001	0.2611	0.2256	0.2691	49
IL29	0.0018	0.0001	0.2341	0.1576	0.0663	44
IL18	0.0136	0.0012	0.179	0.0876	0.2699	43
IGF1R	0.0018	0.0001	0.2291	0.1655	0.2818	42
IL27	0.0018	0.0001	0.2077	0.1398	0.0808	41
ITGB6	0.0018	0.0001	0.1988	0.1219	0.0691	40
SCAP	0.0018	0.0001	0.212	0.1443	0	40
CCNT1	0.007	0.0005	0.1943	0.1122	0.226	40
CD34	0.0499	0.007	0.1692	0.083	0.1198	40
CD99	0.007	0.0005	0.1899	0.1027	0.1261	38
PLIN1	0.0018	0.0001	0.2611	0.203	0	37
C5	0.0112	0.0009	0.1786	0.0965	0.17	37
TNFSF10	0.0328	0.0037	0.1851	0.0928	0.1442	37
CCL5	0.0018	0.0001	0.2286	0.1531	0.2137	36
TNFSF12	0.0045	0.0003	0.1907	0.111	0.1305	36
FADD	0.0152	0.0014	0.1842	0.0929	0.1488	34
APOE	0.0018	0.0001	0.2096	0.1438	0.1055	33

RBL2	0.0333	0.0038	0.1844	0.0951	0.2399	33
VIP	0.0112	0.0009	0.167	0.098	0.1321	31
MAPK9	0.0304	0.0034	0.1746	0.0922	0.2084	31
CD5	0.0119	0.001	0.1989	0.1081	0.219	30
SERPINA1	0.0018	0.0001	0.2227	0.1477	0.2226	29
THRB	0.0119	0.001	0.1814	0.1126	0.089	29
MLX	0.0081	0.0006	0.1861	0.1181	0.232	27
MIF	0.0288	0.0032	0.1834	0.1026	0.1218	27
NPSR1	0.0032	0.0002	0.1789	0.1229	0	26
MAP3K7	0.007	0.0005	0.1938	0.1277	0.1361	26
FOXL2	0.018	0.0018	0.1573	0.0933	0.0801	26
ANGPTL4	0.0421	0.0055	0.2045	0.1032	0.1661	24
TLR2	0.0018	0.0001	0.2566	0.2036	0.3803	23
GNA12	0.0382	0.0046	0.1874	0.0949	0.128	22
ATM	0.0453	0.0061	0.1687	0.0989	0.1245	22
CCR3	0.0018	0.0001	0.2031	0.1586	0.1079	21
JAK2	0.0091	0.0007	0.1974	0.1305	0.1692	20
NRAS	0.0101	0.0008	0.2334	0.14	0.2698	20
CORT	0.0018	0.0001	0.2557	0.2428	0.0493	19
TAB1	0.0018	0.0001	0.2197	0.1957	0	19
CSF2RB	0.0018	0.0001	0.2615	0.2048	0.1597	18
PPARGC1B	0.0018	0.0001	0.2534	0.2398	0.1444	18
CYP19A1	0.0032	0.0002	0.2007	0.16	0.1712	18
ALDH1A2	0.0197	0.002	0.2194	0.1381	0.1368	18
IRS2	0.0045	0.0003	0.2174	0.183	0.1787	16
NFKBIA	0.0119	0.001	0.2109	0.1526	0.2231	16
GNRHR	0.0146	0.0013	0.2182	0.1282	0.0974	16
CCR5	0.0288	0.0032	0.2014	0.1225	0.3539	16
ZNF175	0.0018	0.0001	0.3295	0.3099	0.1765	15
F7	0.0018	0.0001	0.2639	0.249	0.1317	14
INSIG1	0.0045	0.0003	0.2478	0.2167	0.146	14
TBK1	0.0136	0.0012	0.2302	0.181	0.1869	14
SEMA7A	0.0257	0.0028	0.2385	0.1474	0.0939	14
NFE2	0.0408	0.0051	0.1851	0.1397	0.2695	14
CD86	0.0018	0.0001	0.3014	0.2762	0.4353	13
LPIN1	0.0018	0.0001	0.2394	0.2083	0.322	13
IL17F	0.0018	0.0001	0.3134	0.2667	0.0415	13

TFAM	0.0197	0.002	0.2502	0.1544	0.2087	13
NPPA	0.0128	0.0011	0.2222	0.1618	0.0455	12
GABPA	0.0152	0.0014	0.2162	0.1691	0.1811	12
NR2F6	0.0453	0.0061	0.1879	0.1417	0.2273	12
INSIG2	0.0018	0.0001	0.2677	0.2692	0.0981	11
PARP9	0.0018	0.0001	0.2946	0.2667	0.4277	11
CAT	0.0091	0.0007	0.2069	0.1917	0.2225	11
TRIM28	0.0168	0.0016	0.2528	0.1667	0.1197	11
XDH	0.0174	0.0017	0.1996	0.1618	0.0886	11
EHHADH	0.0018	0.0001	0.2864	0.3333	0.4193	10
GRP	0.0018	0.0001	0.2488	0.2418	0.0462	10
HSD17B4	0.0018	0.0001	0.2864	0.3333	0.3351	10
CAMP	0.0152	0.0014	0.2377	0.1667	0.1999	10
ELANE	0.0354	0.0041	0.197	0.1818	0	10
PIAS1	0.0018	0.0001	0.2712	0.2727	0.1313	9
PLAU	0.007	0.0005	0.2336	0.1838	0.1484	9
FOSL1	0.0128	0.0011	0.267	0.2444	0.1846	9
DDX58	0.0252	0.0027	0.2191	0.2	0.1541	9
TBX21	0.034	0.0039	0.171	0.1438	0.2831	9
CGB5	0.0402	0.0049	0.2074	0.15	0.2411	9
PF4	0.0018	0.0001	0.3008	0.3333	0.1183	8
PRKCB	0.0018	0.0001	0.2646	0.2692	0	8
PRKCZ	0.0018	0.0001	0.28	0.2909	0.0663	8
AIMP1	0.0018	0.0001	0.3571	0.3778	0	8
IGF2BP1	0.0018	0.0001	0.3533	0.3273	0.1303	8
CXCL2	0.0058	0.0004	0.3299	0.3111	0.512	8
GCK	0.0081	0.0006	0.2817	0.2778	0.2934	8
HRH2	0.0081	0.0006	0.2245	0.2444	0.1321	8
F10	0.0174	0.0017	0.2221	0.2182	0.1744	8
G6PD	0.0249	0.0026	0.2704	0.2222	0.1572	8
EREG	0.0252	0.0027	0.2494	0.25	0.3799	8
NLRP12	0.0421	0.0055	0.2203	0.1818	0	8
TPPP2	0.0439	0.0058	0.2459	0.1538	0	8
EPHB1	0.0018	0.0001	0.3842	0.5278	0.0787	7
HMGB1	0.0018	0.0001	0.2947	0.303	0.1405	7
TNNI3	0.0032	0.0002	0.2896	0.3611	0.0315	7
DSP	0.0119	0.001	0.3648	0.3333	0.1436	7

PARP2	0.0128	0.0011	0.3133	0.3333	0.1502	7
HSF2	0.0152	0.0014	0.2902	0.2182	0.2684	7
DGAT1	0.0152	0.0014	0.2361	0.2222	0.3676	7
ELF3	0.0231	0.0024	0.1983	0.1758	0.1483	7
FOXC2	0.0265	0.0029	0.224	0.2222	0.1311	7
GFI1	0.0419	0.0054	0.2569	0.2143	0.1242	7
C3	0.0433	0.0057	0.2363	0.1667	0.1715	7
AKR1B1	0.0464	0.0063	0.2122	0.1944	0.2298	7
BACH1	0.0018	0.0001	0.4038	0.5238	0.2259	6
TNXB	0.0018	0.0001	0.5725	0.7333	0.568	6
BID	0.0032	0.0002	0.377	0.5238	0.3656	6
GHRH	0.0032	0.0002	0.3222	0.3929	0.438	6
PLAT	0.0168	0.0016	0.2629	0.2182	0.1346	6
MLXIPL	0.0168	0.0016	0.2796	0.3333	0.256	6
ZFPM2	0.0174	0.0017	0.203	0.197	0.2007	6
SPIC	0.0231	0.0024	0.2315	0.2222	0.0844	6
TIMP1	0.0453	0.0061	0.2858	0.2381	0.3974	6
ACACA	0.0018	0.0001	0.5054	0.6	0.4811	5
TNFRSF8	0.0018	0.0001	0.7131	1	0.1519	5
FCER2	0.0032	0.0002	0.4649	0.5333	0.1378	5
GHR	0.0032	0.0002	0.4344	0.6667	0.1591	5
DGAT2	0.0045	0.0003	0.3564	0.4667	0.5307	5
CASP8	0.0058	0.0004	0.342	0.381	0.1787	5
PCK1	0.0091	0.0007	0.3331	0.3333	0.3374	5
ABCA1	0.0136	0.0012	0.2524	0.2778	0.0853	5
FASLG	0.0168	0.0016	0.3134	0.3333	0.1243	5
DACH1	0.0168	0.0016	0.2285	0.2222	0.0937	5
CCR1	0.018	0.0018	0.327	0.4	0.45	5
CFB	0.0257	0.0028	0.2593	0.2857	0.4113	5
B2M	0.0378	0.0045	0.3088	0.3	0.1353	5
MBTPS1	0.0382	0.0046	0.2416	0.2143	0.1461	5
TARP	0.0413	0.0052	0.2341	0.3	0	5
ABCG1	0.0032	0.0002	0.4312	0.6	0.1138	4
C5AR1	0.0058	0.0004	0.4764	0.6667	0.5814	4
EIF2AK2	0.0081	0.0006	0.4058	0.4	0.1109	4
PON1	0.0168	0.0016	0.3573	0.4	0.0289	4
MT1F	0.0174	0.0017	0.3188	0.4	0.1936	4

NRTN	0.018	0.0018	0.2651	0.4	0.0678	4
PRKG1	0.0206	0.0021	0.3687	0.4	0.0555	4
TGFB1I1	0.0402	0.005	0.219	0.2143	0.1398	4
FLT3LG	0.0415	0.0053	0.2597	0.2381	0.3232	4
PLG	0.0419	0.0054	0.2079	0.2143	0.0538	4
PPBP	0.0444	0.0059	0.2464	0.2143	0.1414	4
EHF	0.0464	0.0063	0.2269	0.2143	0.0438	4
TIRAP	0.047	0.0064	0.3606	0.2143	0.0696	4
GCLC	0.0018	0.0001	0.8637	1	0.1916	3
GCLM	0.0018	0.0001	0.8637	1	0.128	3
ZBED1	0.0018	0.0001	0.6115	1	0.0354	3
PNPT1	0.0018	0.0001	0.667	1	0.0515	3
SLC12A2	0.0081	0.0006	0.4535	0.5	0.1619	3
SLC4A1	0.0091	0.0007	0.4418	0.5	0.2792	3
RTKN	0.0119	0.001	0.3728	0.5	0.1417	3
DRD1	0.0128	0.0011	0.5053	0.5	0.0747	3
CBFB	0.0146	0.0013	0.4325	0.5	0.1673	3
UXT	0.0189	0.0019	0.5082	0.6667	0.0433	3
ZIC1	0.0214	0.0022	0.3747	0.6667	0.0265	3
MT1E	0.0257	0.0028	0.4084	0.6667	0.152	3
ARRB2	0.032	0.0036	0.3149	0.3	0.3479	3
TYMP	0.0354	0.0041	0.1778	0.3	0	3
SERPIND1	0.0359	0.0042	0.2313	0.3	0.0612	3
TANK	0.0395	0.0048	0.319	0.3	0.0772	3
CASP10	0.0402	0.005	0.3267	0.3	0.114	3
KLF6	0.0402	0.005	0.2461	0.3	0.1744	3
MDK	0.0413	0.0052	0.2397	0.3	0.0611	3
ACTC1	0.0018	0.0001	0.8186	1	0	2
ACTG2	0.0018	0.0001	0.8186	1	0.9093	2
CKS1B	0.0018	0.0001	0.5419	1	0.0915	2
CKS2	0.0018	0.0001	0.5419	1	0.4229	2
CMA1	0.0018	0.0001	0.7911	1	0.2685	2
COL2A1	0.0018	0.0001	0.7911	1	0.0977	2
COL4A3	0.0018	0.0001	0.9219	1	0.0664	2
EMD	0.0018	0.0001	0.4409	1	0.1359	2
FCGR3A	0.0018	0.0001	0.6979	1	0.2433	2
G6PC	0.0018	0.0001	0.6594	1	0.3004	2

HIST1H1C	0.0018	0.0001	0.6185	1	0.2032	2
HTR2C	0.0018	0.0001	0.5514	1	0.0286	2
IK	0.0018	0.0001	0.6056	1	0.0581	2
IL12RB1	0.0018	0.0001	0.6583	1	0.1491	2
MB	0.0018	0.0001	0.563	1	0.0208	2
NMB	0.0018	0.0001	0.8704	1	0.1293	2
ODC1	0.0018	0.0001	0.5283	1	0.1498	2
OMG	0.0018	0.0001	0.8007	1	0.055	2
PLA2G2A	0.0018	0.0001	0.6166	1	0.1777	2
PLP1	0.0018	0.0001	0.5454	1	0.0966	2
TRPM2	0.0018	0.0001	0.8177	1	0.0566	2
TSC1	0.0018	0.0001	0.5283	1	0.1606	2
TXK	0.0018	0.0001	0.5334	1	0.1316	2
NPHS2	0.0018	0.0001	0.4347	1	0.0287	2
CFLAR	0.0018	0.0001	0.5688	1	0.1169	2
RPS6KA4	0.0018	0.0001	0.8704	1	0.1618	2
F2RL3	0.0018	0.0001	0.8605	1	0.0776	2
RPS6KA5	0.0018	0.0001	0.8704	1	0.0441	2
ZFYVE16	0.0018	0.0001	0.5054	1	0.0801	2
SH2B2	0.0018	0.0001	0.8007	1	0	2
CD160	0.0018	0.0001	0.8704	1	0.0258	2
TPSD1	0.0018	0.0001	0.4796	1	0.0404	2
TOR2A	0.0018	0.0001	0.8007	1	0.0532	2
ARID5B	0.0018	0.0001	0.6033	1	0.076	2
CREB3L1	0.0018	0.0001	0.5488	1	0.2695	2
VSX2	0.0018	0.0001	0.4514	1	0	2
KDM1A	0.0433	0.0057	0.2832	0.3333	0	2
USP11	0.0479	0.0066	0.2602	0.3333	0.3514	2
FGF18	0.0479	0.0066	0.2315	0.3333	0.0379	2
MSR1	0.0484	0.0067	0.3703	0.3333	0.591	2

APPENDIX B

Regulator	FDR	p-value	ave. corr. (regulatee)	fraction	ave. corr. (regulator-regulatee)	# of regulatees
TNF	0.0011	0.0001	0.1284	0.1206	0.1488	810
TGFB1	0.0011	0.0001	0.1333	0.129	0.2157	643
IFNG	0.0011	0.0001	0.1614	0.1927	0.1821	555
IL1B	0.0011	0.0001	0.1316	0.1279	0.2055	408
IL6	0.0011	0.0001	0.1422	0.1465	0.1612	325
IL13	0.0011	0.0001	0.1522	0.1681	0	278
IL4	0.0011	0.0001	0.1611	0.1918	0	274
IL1A	0.0011	0.0001	0.1426	0.1574	0.1224	257
IFNB1	0.0011	0.0001	0.1516	0.1697	0	249
MYCN	0.0018	0.0002	0.1362	0.1328	0.1015	248
F2	0.0128	0.002	0.1325	0.1262	0	234
CDKN1A	0.0011	0.0001	0.1746	0.2228	0.108	233
IL10	0.0011	0.0001	0.185	0.2508	0.2602	227
FGF2	0.0033	0.0004	0.1348	0.1329	0.1198	227
IL5	0.0275	0.0054	0.1334	0.124	0	201
CSF2	0.0011	0.0001	0.1648	0.2001	0	196
IL2	0.0011	0.0001	0.1684	0.1998	0	185
MAPK1	0.0011	0.0001	0.1723	0.215	0.1101	179
VEGFA	0.0059	0.0008	0.1356	0.1364	0.0877	177
STAT5A	0.0297	0.0061	0.134	0.1286	0.1326	163
TNFSF11	0.0011	0.0001	0.1593	0.1986	0.0757	148
DMD	0.0011	0.0001	0.1543	0.1795	0.0807	146
E2F1	0.0163	0.0027	0.1394	0.1413	0.2017	136
E2F4	0.0011	0.0001	0.1685	0.2067	0.08	130
IRF7	0.0011	0.0001	0.2097	0.2899	0.2619	130
RELA	0.0011	0.0001	0.1437	0.1589	0.1123	129
CSF1	0.0011	0.0001	0.1691	0.1871	0.234	120
IFNA2	0.0011	0.0001	0.2372	0.3505	0	118
IKBKB	0.0018	0.0002	0.1468	0.1686	0.0786	112
EP300	0.004	0.0005	0.1446	0.1594	0.0758	106
AGT	0.0011	0.0001	0.1577	0.1896	0.0816	105
SOCS3	0.0011	0.0001	0.1527	0.1781	0.1643	103
EDN1	0.0011	0.0001	0.1674	0.2124	0.1599	102
CHUK	0.0018	0.0002	0.1513	0.1769	0.072	100
FGF1	0.0471	0.0118	0.1368	0.1389	0.2085	97

SRC	0.0289	0.0058	0.1457	0.1493	0.1043	96
NCOR1	0.0088	0.0013	0.1492	0.1641	0.116	93
STAT1	0.0011	0.0001	0.2121	0.3137	0.3147	90
PTGS2	0.0364	0.0079	0.1356	0.1465	0.1487	83
CD40LG	0.0011	0.0001	0.2297	0.3547	0	80
CSF3	0.0018	0.0002	0.1643	0.1975	0	78
IL3	0.0026	0.0003	0.1514	0.1816	0	78
IFNA1	0.0011	0.0001	0.2187	0.326	0	76
STAT6	0.0011	0.0001	0.1781	0.2401	0.088	76
THAP1	0.0011	0.0001	0.2848	0.4717	0.0959	76
DYSF	0.0046	0.0006	0.1515	0.1764	0.1377	75
REL	0.0278	0.0055	0.1442	0.1576	0.1125	74
TGFB2	0.0033	0.0004	0.1574	0.1826	0.0978	72
SMAD3	0.0011	0.0001	0.1714	0.2074	0.0896	71
JUN	0.0011	0.0001	0.1682	0.2083	0.096	68
IKBKG	0.0053	0.0007	0.1501	0.1834	0.1075	68
TLR4	0.0011	0.0001	0.2073	0.3009	0.3535	67
IL1RN	0.0011	0.0001	0.2628	0.3863	0.28	66
OSM	0.0011	0.0001	0.1654	0.2221	0	65
PTH	0.0128	0.002	0.1556	0.1766	0	62
IL15	0.0011	0.0001	0.1864	0.2438	0.282	60
IL17A	0.0011	0.0001	0.1746	0.2458	0	59
BRMS1	0.0175	0.0029	0.1598	0.1702	0.0651	59
ERBB3	0.0011	0.0001	0.1781	0.2212	0.0998	57
IRF8	0.0011	0.0001	0.1697	0.214	0.3138	54
FAS	0.0011	0.0001	0.2742	0.4511	0.2914	53
SPI1	0.0011	0.0001	0.3227	0.4904	0.4947	53
SMAD7	0.0331	0.007	0.1607	0.1712	0.1411	52
IL18	0.0011	0.0001	0.2321	0.381	0.2549	50
TGFB3	0.0011	0.0001	0.2354	0.3491	0.2858	50
ERBB4	0.0018	0.0002	0.1808	0.2327	0.1124	50
NRF1	0.0066	0.0009	0.1647	0.2021	0	50
IL29	0.0011	0.0001	0.5344	0.8831	0	49
IL7	0.0046	0.0006	0.1776	0.2202	0.2266	49
IRF1	0.0011	0.0001	0.2634	0.4253	0.4006	48
FADD	0.0088	0.0013	0.1654	0.1983	0.0749	48
CD40	0.0128	0.002	0.1571	0.1873	0.1693	48
TNFRSF1A	0.0026	0.0003	0.1715	0.2264	0.1495	47
RB1	0.0011	0.0001	0.1808	0.2569	0.1588	46

MAP2K1	0.0053	0.0007	0.1712	0.2151	0.0557	46
IL27	0.0011	0.0001	0.3201	0.5347	0	45
KLF2	0.0011	0.0001	0.1796	0.2451	0.26	44
ITGB6	0.0011	0.0001	0.2578	0.4201	0.2582	43
TNFSF10	0.0011	0.0001	0.2345	0.3788	0.322	43
C5	0.0011	0.0001	0.184	0.2627	0.0681	42
BNIP3L	0.0018	0.0002	0.1846	0.2302	0.0797	41
APOE	0.0011	0.0001	0.2377	0.3426	0.3081	40
IFNA13	0.0011	0.0001	0.2582	0.4297	0	40
S100B	0.0341	0.0073	0.1626	0.185	0.1257	40
CTGF	0.0434	0.01	0.1675	0.1798	0.2495	40
HTT	0.0011	0.0001	0.1896	0.247	0	39
IRF3	0.0297	0.0061	0.1725	0.1884	0.0957	39
TNFSF12	0.0011	0.0001	0.2722	0.4679	0.2291	37
FOXP3	0.0011	0.0001	0.1903	0.2432	0	37
CCL5	0.0018	0.0002	0.1833	0.2525	0.2789	37
TCF3	0.0083	0.0012	0.1823	0.2106	0.0902	37
VIP	0.0011	0.0001	0.2019	0.2776	0	35
HOXD3	0.0018	0.0002	0.2067	0.2973	0.1129	35
MAP3K7	0.0011	0.0001	0.25	0.4082	0.085	33
SPDEF	0.0011	0.0001	0.2248	0.3131	0	33
SMAD4	0.0018	0.0002	0.1837	0.2709	0.0793	33
NFKB1	0.0033	0.0004	0.1941	0.2632	0.2394	33
RBL2	0.0078	0.0011	0.1746	0.2179	0.1133	33
LPL	0.0184	0.0031	0.1711	0.2033	0.1149	33
SATB1	0.0321	0.0067	0.1758	0.1991	0.092	32
NPSR1	0.0018	0.0002	0.1935	0.2648	0	31
CD5	0.0118	0.0018	0.1926	0.2302	0.2713	30
FOXL2	0.015	0.0024	0.1585	0.2176	0.1192	30
MYD88	0.0011	0.0001	0.2282	0.3747	0.2271	29
PTK2	0.0159	0.0026	0.1835	0.2298	0.0947	29
PDGFB	0.0011	0.0001	0.2122	0.3106	0.2156	28
FOXM1	0.0046	0.0006	0.197	0.2644	0.308	28
THRA	0.0409	0.0091	0.1439	0.195	0.1148	28
ATM	0.0011	0.0001	0.2577	0.3419	0.0838	27
SOCS1	0.0011	0.0001	0.2203	0.2989	0.1808	26
IL21	0.0011	0.0001	0.2433	0.381	0	26
LTA	0.0018	0.0002	0.2045	0.3228	0	26
GPX1	0.0094	0.0014	0.185	0.2506	0.1878	26

HSF1	0.0439	0.0102	0.1641	0.2116	0.0887	26
ADIPOQ	0.0184	0.0031	0.1772	0.228	0.1679	25
RBL1	0.0011	0.0001	0.2754	0.4954	0.3379	24
TLR2	0.0011	0.0001	0.3218	0.56	0.4882	24
CCR2	0.0011	0.0001	0.2698	0.4433	0.3448	24
LIPG	0.0263	0.0051	0.1729	0.2194	0.1466	24
CD34	0.0321	0.0067	0.1675	0.2222	0.1642	24
BCL6	0.0417	0.0094	0.1534	0.2046	0.0877	24
GAPDH	0.0011	0.0001	0.2597	0.4167	0.0853	23
TAB1	0.0011	0.0001	0.3017	0.5181	0	23
CCL13	0.0018	0.0002	0.2367	0.3518	0.1887	22
CORT	0.0011	0.0001	0.3158	0.5801	0	21
TLR3	0.0011	0.0001	0.2547	0.4372	0.3063	21
DKK1	0.0469	0.0117	0.1704	0.2138	0.1019	21
F7	0.0011	0.0001	0.2337	0.3939	0	20
ALDH1A2	0.0011	0.0001	0.2857	0.4286	0.1411	20
CCL11	0.01	0.0015	0.2002	0.2717	0.3139	20
GNRHR	0.0289	0.0058	0.1815	0.2355	0	20
CXCL10	0.0392	0.0086	0.1908	0.22	0.3279	20
CCR5	0.0018	0.0002	0.2285	0.3238	0.3969	19
BAX	0.0018	0.0002	0.2256	0.3571	0.0947	18
RORA	0.0206	0.0037	0.171	0.2632	0.1178	18
IL12A	0.0011	0.0001	0.2283	0.3987	0.1417	17
IL12B	0.0011	0.0001	0.2494	0.4	0	17
IL1R1	0.0059	0.0008	0.2152	0.3333	0.0693	16
CRH	0.0133	0.0021	0.1963	0.2865	0	16
IHH	0.0255	0.0049	0.1836	0.2745	0	16
SEMA7A	0.0266	0.0052	0.207	0.2474	0	16
PRKCD	0.0422	0.0096	0.1637	0.2294	0.1003	16
NFKBIA	0.0011	0.0001	0.2838	0.5515	0.3262	15
XDH	0.0011	0.0001	0.297	0.55	0.166	15
ABCB1	0.0018	0.0002	0.2629	0.3476	0.1998	15
CSF2RB	0.0046	0.0006	0.2207	0.3095	0	15
E2F2	0.0066	0.0009	0.2118	0.3268	0.2793	15
E2F3	0.0066	0.0009	0.209	0.3309	0.1528	15
BTC	0.0088	0.0013	0.2302	0.3333	0.0817	15
F2R	0.025	0.0048	0.1942	0.2632	0.2201	15
THBS1	0.0011	0.0001	0.2645	0.4667	0.3828	14
C5orf13	0.0011	0.0001	0.3452	0.5333	0.1319	14

IL17F	0.0011	0.0001	0.3312	0.5619	0	14
CEBPE	0.0046	0.0006	0.2137	0.3583	0	14
RUNX1	0.0187	0.0032	0.2011	0.3083	0.2246	14
HBP1	0.0431	0.0099	0.1911	0.2667	0.1022	14
CAMP	0.0011	0.0001	0.3524	0.7051	0	13
ZNF175	0.0011	0.0001	0.4129	0.5714	0.1094	13
IL17RA	0.0011	0.0001	0.2942	0.5641	0.2207	13
TBK1	0.0011	0.0001	0.3499	0.5275	0.1024	13
TBX21	0.0011	0.0001	0.3869	0.6571	0	13
PARP9	0.0011	0.0001	0.4825	0.7692	0.5707	13
CCL2	0.0018	0.0002	0.2731	0.5128	0.3962	13
ADA	0.0026	0.0003	0.2518	0.419	0.0915	13
CDH1	0.0163	0.0027	0.2264	0.3333	0.0915	13
MYOCD	0.0243	0.0046	0.1901	0.281	0	13
IL22	0.0297	0.0061	0.1948	0.2794	0	13
NPPA	0.0399	0.0088	0.1941	0.2667	0	13
AGTR1	0.0453	0.0108	0.1749	0.2398	0.2567	13
FBN1	0.0011	0.0001	0.4	0.6484	0.5883	12
HDAC6	0.0011	0.0001	0.2858	0.4872	0.1044	12
IL25	0.0011	0.0001	0.2872	0.5128	0	12
DDR1	0.0018	0.0002	0.3401	0.5909	0.1206	12
TREM1	0.0026	0.0003	0.2255	0.4066	0.2824	12
ITGB1	0.0033	0.0004	0.273	0.4359	0.2205	12
HMOX1	0.0083	0.0012	0.2265	0.3974	0.3442	12
GDF9	0.0192	0.0034	0.2036	0.2941	0.136	12
CAT	0.0215	0.0039	0.1834	0.2952	0.0382	12
HOXB7	0.0266	0.0052	0.221	0.3182	0.0449	12
CD86	0.0011	0.0001	0.3673	0.6818	0.5284	11
PTAFR	0.0033	0.0004	0.2631	0.4909	0.2883	11
ICAM1	0.004	0.0005	0.2448	0.4103	0.2991	11
HIVEP1	0.0066	0.0009	0.239	0.35	0.0584	11
NPPB	0.0228	0.0042	0.1772	0.3077	0	11
LDLR	0.025	0.0048	0.2386	0.3636	0.1969	11
TNFSF13B	0.0297	0.0061	0.1679	0.2952	0.3165	11
PIAS1	0.042	0.0095	0.2098	0.3091	0.0469	11
PF4	0.0011	0.0001	0.4154	0.7778	0	10
NLRP12	0.0011	0.0001	0.4782	0.5636	0	10
TICAM1	0.0011	0.0001	0.4281	0.6444	0.1982	10
HMGB1	0.0026	0.0003	0.282	0.4848	0.0788	10

C3	0.0072	0.001	0.2785	0.4182	0.2492	10
BAK1	0.0237	0.0044	0.2121	0.3205	0.2372	10
PRKCB	0.0239	0.0045	0.2223	0.3485	0	10
OLR1	0.0011	0.0001	0.4369	0.8056	0.5582	9
AIMP1	0.0011	0.0001	0.4375	0.75	0	9
IGF2BP1	0.0011	0.0001	0.5706	0.8	0.0748	9
NPR1	0.0018	0.0002	0.3367	0.5636	0.1101	9
NT5E	0.0018	0.0002	0.3068	0.5333	0.362	9
DDX58	0.0046	0.0006	0.3512	0.5278	0.4749	9
HRH2	0.0083	0.0012	0.2851	0.5	0	9
CXCL3	0.0159	0.0026	0.2454	0.4167	0.2109	9
TRAF6	0.0239	0.0045	0.2157	0.3889	0.0786	9
FOXF1	0.0286	0.0057	0.2249	0.3333	0.1901	9
SMAD1	0.0407	0.009	0.2633	0.3333	0.0894	9
S1PR1	0.0456	0.0112	0.1915	0.3091	0	9
PRKCZ	0.0493	0.0126	0.2288	0.3111	0.1228	9
TSC22D3	0.0011	0.0001	0.4354	0.75	0.094	8
PPBP	0.0011	0.0001	0.4081	0.7143	0.278	8
TNNI3	0.0011	0.0001	0.4634	0.8214	0.0579	8
TRAF3IP2	0.0011	0.0001	0.3141	0.6429	0.0739	8
SPIC	0.0011	0.0001	0.4684	0.75	0	8
CBL	0.0018	0.0002	0.4775	0.6222	0.0572	8
IFNAR1	0.0018	0.0002	0.3825	0.6786	0.2189	8
PLG	0.0059	0.0008	0.2737	0.5357	0	8
VCAN	0.0094	0.0014	0.2565	0.5	0.4753	8
TFDP1	0.0123	0.0019	0.2675	0.4444	0.2462	8
HSF2	0.0187	0.0032	0.2125	0.3778	0.2117	8
DUSP1	0.0278	0.0055	0.2218	0.3556	0.245	8
HEY1	0.0364	0.0079	0.2458	0.3611	0.118	8
ELF1	0.0011	0.0001	0.5876	0.9524	0.1427	7
ABCB4	0.0026	0.0003	0.3895	0.5714	0	7
FOSL1	0.0026	0.0003	0.3372	0.5556	0	7
EPHB1	0.0033	0.0004	0.3184	0.5278	0.0401	7
CXCL2	0.0088	0.0013	0.3083	0.5	0.3822	7
BID	0.0105	0.0016	0.2986	0.5714	0.1931	7
TIRAP	0.0123	0.0019	0.3934	0.5238	0	7
IL23A	0.0133	0.0021	0.2495	0.4643	0.1135	7
BARX1	0.0159	0.0026	0.2287	0.4167	0	7
FASLG	0.0192	0.0034	0.2736	0.4762	0.1772	7

IDO1	0.0228	0.0042	0.2009	0.4286	0	7
ILK	0.0259	0.005	0.2629	0.3455	0.1241	7
FCGR2B	0.0291	0.0059	0.2715	0.4286	0.3921	7
KNG1	0.0331	0.007	0.2186	0.3611	0	7
BDKRB1	0.0431	0.0099	0.2367	0.3571	0.2655	7
TNXB	0.0011	0.0001	0.8818	1	0.058	6
DSP	0.0018	0.0002	0.4585	0.8	0.0541	6
CASP8	0.0033	0.0004	0.3866	0.7333	0.0716	6
INHBB	0.019	0.0033	0.2693	0.4167	0.0725	6
RXRB	0.019	0.0033	0.2241	0.4286	0.0871	6
TXN	0.0278	0.0055	0.243	0.4286	0.072	6
RFX5	0.0282	0.0056	0.3977	0.4667	0.3275	6
LILRB4	0.0291	0.0059	0.3095	0.4667	0.466	6
S1PR2	0.03	0.0062	0.307	0.4667	0	6
ZBTB7B	0.03	0.0062	0.3386	0.4667	0	6
EREG	0.0417	0.0094	0.2208	0.381	0.2933	6
NPC1	0.0417	0.0094	0.2289	0.3571	0	6
IL10RA	0.0459	0.0114	0.255	0.381	0.4407	6
IRF9	0.0488	0.0124	0.2424	0.4	0.2781	6
PSMD4	0.0011	0.0001	0.4077	1	0.0567	5
MAPKAPK2	0.0011	0.0001	0.411	0.9	0.0191	5
IFNK	0.0011	0.0001	0.5843	0.9	0	5
SMOC2	0.0011	0.0001	0.5815	1	0.1801	5
PRKG1	0.0026	0.0003	0.4543	0.9	0.467	5
PTGIR	0.0026	0.0003	0.3117	0.8	0	5
PARP2	0.004	0.0005	0.369	0.6667	0.0627	5
MED28	0.004	0.0005	0.4004	0.6667	0	5
E2F8	0.0046	0.0006	0.3503	0.8	0.448	5
FCER2	0.0053	0.0007	0.4035	0.6667	0	5
TNFRSF8	0.0059	0.0008	0.3596	0.7	0	5
ANXA2	0.0072	0.001	0.3253	0.6	0.2871	5
RARRES3	0.0078	0.0011	0.4587	0.7	0.4302	5
SFTPD	0.0083	0.0012	0.3168	0.6	0.0803	5
MST1	0.0088	0.0013	0.324	0.7	0.2038	5
F2RL1	0.015	0.0024	0.3132	0.6	0.0917	5
SLPI	0.0156	0.0025	0.2585	0.6	0.1241	5
LTBP4	0.0184	0.0031	0.2472	0.5333	0.1344	5
TGFB1I1	0.0192	0.0034	0.2986	0.5333	0.3685	5
DLX5	0.0196	0.0035	0.2702	0.5333	0.1014	5

FCGR2A	0.0313	0.0065	0.2602	0.4667	0.4464	5
STAT2	0.0324	0.0068	0.3167	0.5	0.2525	5
MMP13	0.0338	0.0072	0.2907	0.5	0.3257	5
TRIM21	0.0359	0.0077	0.2568	0.5	0.2794	5
HP	0.0453	0.0108	0.2559	0.4	0.1692	5
GABPB1	0.0454	0.011	0.2047	0.4	0	5
ABCA1	0.0474	0.0119	0.2596	0.381	0.1622	5
ITGAV	0.0011	0.0001	0.6839	1	0.2752	4
MBP	0.0011	0.0001	0.6479	1	0.1123	4
DDR2	0.0011	0.0001	0.5943	1	0.4028	4
ABCG1	0.0078	0.0011	0.4424	0.8333	0.1188	4
TNFAIP3	0.0094	0.0014	0.3463	0.8333	0.4688	4
LGI1	0.0105	0.0016	0.4602	0.8333	0.1187	4
C5AR1	0.0123	0.0019	0.5863	0.8333	0.5999	4
PON1	0.0159	0.0026	0.3832	0.6	0.1179	4
AHRR	0.0159	0.0026	0.2525	0.6	0	4
LBP	0.0184	0.0031	0.3223	0.6667	0.0716	4
PLA2G10	0.0187	0.0032	0.299	0.6667	0.0612	4
PTN	0.019	0.0033	0.2579	0.6	0.0574	4
CBX4	0.019	0.0033	0.4649	0.6667	0	4
CD200R1	0.0192	0.0034	0.3108	0.6667	0	4
TANK	0.0196	0.0035	0.353	0.6667	0.0958	4
DNASE2	0.0206	0.0037	0.3843	0.6667	0.19	4
C1QA	0.021	0.0038	0.3002	0.6667	0.1886	4
CASP10	0.021	0.0038	0.3528	0.6667	0.1002	4
SELL	0.0218	0.004	0.3453	0.6667	0.3205	4
MAVS	0.0218	0.004	0.372	0.6667	0	4
B2M	0.0232	0.0043	0.444	0.6667	0.4819	4
WNT10B	0.0239	0.0045	0.2961	0.6667	0	4
LAT	0.0243	0.0046	0.3081	0.6667	0.1237	4
RCAN1	0.0247	0.0047	0.2791	0.6667	0.2421	4
ARRB2	0.0328	0.0069	0.301	0.5	0.1803	4
SFTPA1	0.0338	0.0072	0.2646	0.5	0	4
SCGB1A1	0.0362	0.0078	0.2891	0.5	0.0724	4
CCR1	0.0382	0.0083	0.3042	0.5	0.4395	4
EDNRA	0.0385	0.0084	0.3741	0.5	0.3249	4
SPIB	0.0388	0.0085	0.2963	0.5	0	4
NOD2	0.0409	0.0091	0.2624	0.5	0.2456	4
FLT3LG	0.0417	0.0094	0.2545	0.5	0	4

IL27RA	0.0448	0.0106	0.2917	0.5	0.0863	4
FABP4	0.0458	0.0113	0.2735	0.5	0.2879	4
FYN	0.0458	0.0113	0.3026	0.5	0.1694	4
LTB4R	0.0484	0.0122	0.2208	0.5	0	4
CFB	0.0487	0.0123	0.2938	0.4	0.3246	4
PBX1	0.0488	0.0124	0.2667	0.4	0.1543	4
TIMP1	0.05	0.0128	0.2648	0.4	0.2038	4
CBFB	0.0011	0.0001	0.9446	1	0.0961	3
DES	0.0011	0.0001	0.4602	1	0.1004	3
DRD1	0.0011	0.0001	0.7372	1	0	3
EPHX2	0.0011	0.0001	0.3956	1	0.1719	3
GABBR1	0.0011	0.0001	0.3313	1	0.1287	3
GCLC	0.0011	0.0001	0.5474	1	0.0791	3
GCLM	0.0011	0.0001	0.5474	1	0.0468	3
GNRH2	0.0011	0.0001	0.396	1	0	3
HSPA1A	0.0011	0.0001	0.495	1	0.1519	3
IL13RA2	0.0011	0.0001	0.4086	1	0.0599	3
JAG2	0.0011	0.0001	0.3767	1	0.2345	3
LTB	0.0011	0.0001	0.5664	1	0.4006	3
MSR1	0.0011	0.0001	0.3951	1	0.3965	3
GADD45B	0.0011	0.0001	0.5926	1	0.4372	3
PRTN3	0.0011	0.0001	0.4449	1	0	3
TCF7	0.0011	0.0001	0.3491	1	0.0571	3
UBA7	0.0011	0.0001	0.6578	1	0	3
UXT	0.0011	0.0001	0.4761	1	0.0956	3
RIPK2	0.0011	0.0001	0.6284	1	0.2921	3
TNFRSF6B	0.0011	0.0001	0.5085	1	0	3
ZBED1	0.0011	0.0001	0.3991	1	0	3
BACH2	0.0011	0.0001	0.3644	1	0	3
NLRP3	0.0011	0.0001	0.5498	1	0.6298	3
LSS	0.0422	0.0096	0.3034	0.6667	0.1865	3
RGS2	0.0431	0.0099	0.2664	0.5	0.2105	3
LRP1	0.0437	0.0101	0.3483	0.5	0	3
MSX1	0.0439	0.0102	0.2214	0.5	0.0819	3
CD69	0.0442	0.0103	0.2746	0.5	0.4503	3
HEYL	0.0445	0.0104	0.2454	0.6667	0	3
ITGA5	0.0448	0.0106	0.3309	0.6667	0.5833	3
MMP7	0.0448	0.0106	0.324	0.6667	0.1766	3
SULT1E1	0.0448	0.0106	0.2764	0.6667	0.0688	3

AGA	0.0453	0.0109	0.2718	0.5	0	3
TNIP3	0.0453	0.0109	0.3881	0.6667	0.3482	3
NCF1	0.0453	0.0109	0.2966	0.6667	0.3854	3
RTKN	0.0454	0.011	0.2728	0.5	0.0705	3
ING4	0.0454	0.011	0.2578	0.5	0.0801	3
HSPG2	0.0454	0.0111	0.2216	0.6667	0.1914	3
MMP9	0.0454	0.0111	0.3104	0.6667	0.3037	3
LRRC8C	0.0454	0.0111	0.2664	0.5	0	3
KDM1A	0.0459	0.0114	0.3209	0.6667	0	3
SLC4A1	0.0462	0.0115	0.4954	0.5	0	3
FGA	0.0493	0.0126	0.388	0.6667	0	3

APPENDIX C

Regulator	FDR	p-value	ave. corr. (regulatees)	fraction	ave. corr. (regulator-regulatees)	# of coherent regulatees
TGFB1	0.0014	0.0001	0.1242	0.0942	0.0885	723
IFNG	0.0014	0.0001	0.1365	0.1246	0.238	624
MYC	0.0014	0.0001	0.1475	0.1406	0.1623	439
IL6	0.0014	0.0001	0.1289	0.1084	0.1104	370
IL4	0.0014	0.0001	0.1313	0.112	0.0863	333
IL13	0.0084	0.0007	0.1255	0.1004	0.0724	302
MYCN	0.0014	0.0001	0.1749	0.1848	0.0972	272
IL1A	0.0466	0.0056	0.1244	0.097	0.0746	262
ERBB2	0.0027	0.0002	0.1343	0.1133	0.0535	253
IFNB1	0.0014	0.0001	0.1449	0.1394	0.0911	251
IL10	0.0014	0.0001	0.1351	0.1186	0.1336	249
CDKN1A	0.0014	0.0001	0.1896	0.2396	0.084	234
F2	0.004	0.0003	0.1351	0.1148	0.0704	233
IL2	0.0014	0.0001	0.1508	0.1561	0.0807	205
CCND1	0.0014	0.0001	0.1454	0.1388	0.2025	204
VEGFA	0.0014	0.0001	0.1427	0.1271	0.1572	189
MAPK1	0.0014	0.0001	0.1394	0.1291	0.1062	183
TNFSF11	0.0014	0.0001	0.1493	0.1511	0.076	166
BRCA1	0.0014	0.0001	0.1443	0.141	0.251	162
CDKN2A	0.0014	0.0001	0.1523	0.1541	0.0842	158
PRL	0.0084	0.0007	0.139	0.1202	0.0822	155
DMD	0.0014	0.0001	0.1483	0.1456	0.097	153
IRF7	0.0014	0.0001	0.172	0.2045	0.2636	142
FOS	0.0493	0.006	0.1331	0.1113	0.1802	133
E2F4	0.0014	0.0001	0.1801	0.2227	0.1615	130
IFNA2	0.0014	0.0001	0.1893	0.2487	0.0788	128
E2F1	0.0027	0.0002	0.1427	0.1382	0.2102	127
CSF1	0.0014	0.0001	0.1395	0.1307	0.1579	125
XBP1	0.0376	0.0043	0.1382	0.1176	0.1923	124
IKBKB	0.0014	0.0001	0.1434	0.1402	0.1011	121
EDN1	0.0014	0.0001	0.1581	0.1692	0.0985	116
SOCS3	0.0051	0.0004	0.1413	0.1355	0.1739	113
EP300	0.0073	0.0006	0.14	0.1291	0.0842	110

FGF1	0.0204	0.002	0.134	0.1223	0.0909	109
HMGA1	0.0499	0.0061	0.1349	0.1162	0.1676	109
CHUK	0.0014	0.0001	0.155	0.1661	0.1245	105
MAPK8	0.0169	0.0016	0.136	0.1241	0.0779	96
STAT1	0.0014	0.0001	0.1803	0.231	0.3404	93
NCOR1	0.0014	0.0001	0.1605	0.1717	0.0867	93
THAP1	0.0014	0.0001	0.2674	0.4077	0.0747	87
DYSF	0.0014	0.0001	0.1566	0.1667	0.2287	84
STAT6	0.0014	0.0001	0.1543	0.1583	0.1399	80
CD40LG	0.0051	0.0004	0.1482	0.148	0.1277	79
TLR4	0.0336	0.0037	0.1356	0.1247	0.2064	79
ZBTB16	0.0014	0.0001	0.1721	0.215	0.0855	78
ROCK1	0.026	0.0027	0.1442	0.1321	0.1128	77
JAK1	0.0062	0.0005	0.1467	0.1445	0.0989	75
JUN	0.032	0.0035	0.1383	0.1283	0.2185	74
IFNA1	0.0014	0.0001	0.1824	0.2283	0.0874	73
IKBKG	0.0027	0.0002	0.1495	0.156	0.0764	71
IL1RN	0.0014	0.0001	0.1812	0.2302	0.1523	67
SPI1	0.0014	0.0001	0.1579	0.1757	0.2008	66
PTH	0.0362	0.0041	0.1451	0.1326	0.071	65
IL15	0.0299	0.0032	0.1472	0.1396	0.2728	64
TGFBR1	0.0236	0.0024	0.1495	0.1425	0.1207	62
IRF8	0.0105	0.0009	0.1469	0.1497	0.1062	61
ERBB3	0.0178	0.0017	0.1496	0.1461	0.0669	57
RB1	0.0014	0.0001	0.1744	0.2112	0.1264	56
VHL	0.0094	0.0008	0.1474	0.158	0.0973	56
IL29	0.0014	0.0001	0.3501	0.5966	0.1094	53
IRF1	0.0051	0.0004	0.161	0.1782	0.3162	52
FAS	0.0014	0.0001	0.2081	0.277	0.1505	51
SMAD7	0.0134	0.0012	0.1631	0.1598	0.2424	51
CCL5	0.0073	0.0006	0.1621	0.1829	0.2596	50
IL27	0.0014	0.0001	0.1894	0.2348	0.1012	48
FADD	0.0362	0.0041	0.1539	0.1494	0.082	48
S100B	0.0014	0.0001	0.1797	0.2257	0.0838	47
ITGB6	0.0014	0.0001	0.1879	0.2498	0.0923	46
TGFB3	0.0014	0.0001	0.1992	0.2513	0.1224	45
BNIP3L	0.0014	0.0001	0.1722	0.2044	0.1035	44

TNFSF10	0.0014	0.0001	0.1775	0.2029	0.298	43
TNC	0.0027	0.0002	0.1905	0.2251	0.2042	43
APOE	0.0014	0.0001	0.1806	0.2199	0.3067	41
MAP2K1	0.0027	0.0002	0.1642	0.1967	0.0952	41
TNFSF12	0.0027	0.0002	0.1648	0.1998	0.2424	39
IFNA13	0.0014	0.0001	0.1888	0.2387	0	38
RBL2	0.0124	0.0011	0.1578	0.1839	0.13	35
CYP2C18	0.0294	0.0031	0.1592	0.1787	0.0623	33
MAP3K7	0.0014	0.0001	0.1898	0.2492	0.1012	30
SPDEF	0.0014	0.0001	0.2372	0.3455	0.0598	30
MYD88	0.0169	0.0016	0.1677	0.1849	0.1536	29
HOXD3	0.0312	0.0034	0.1505	0.1603	0.0679	29
MAP3K1	0.0204	0.002	0.1568	0.1857	0.1201	28
GPX1	0.0027	0.0002	0.1887	0.2359	0.274	27
HSF1	0.0062	0.0005	0.1755	0.2238	0.1897	27
ATM	0.0014	0.0001	0.217	0.2874	0.1025	26
SOCS1	0.0014	0.0001	0.1938	0.2828	0.0826	25
S100A6	0.0299	0.0032	0.1824	0.1994	0.1827	25
LTA	0.0394	0.0046	0.1576	0.1828	0.1517	25
FOXM1	0.0139	0.0013	0.1659	0.1978	0.3062	23
RBL1	0.0014	0.0001	0.247	0.4021	0.2387	22
TAB1	0.0014	0.0001	0.2218	0.3167	0.0741	22
CCL13	0.0196	0.0019	0.1768	0.2116	0.1846	21
CCR5	0.0245	0.0025	0.1755	0.2101	0.3323	21
GAPDH	0.0014	0.0001	0.222	0.3033	0.1692	19
IL12B	0.0094	0.0008	0.2066	0.249	0.0743	19
CORT	0.0113	0.001	0.2005	0.2391	0.097	19
F7	0.0305	0.0033	0.1748	0.2065	0.0784	18
PRKCD	0.0358	0.004	0.1596	0.1931	0.0875	18
BLK	0.0437	0.0052	0.1595	0.2016	0.1642	18
SEMA7A	0.0139	0.0013	0.2185	0.2579	0.0743	17
ALDH1A2	0.0014	0.0001	0.2892	0.3853	0.0975	16
IL17F	0.0051	0.0004	0.2474	0.3235	0.0628	16
PARP9	0.0014	0.0001	0.3443	0.6	0.3841	15
F2R	0.004	0.0003	0.2317	0.3041	0.24	15
GRP	0.0362	0.0041	0.1911	0.2353	0.0556	15
TBK1	0.0014	0.0001	0.2776	0.4667	0.0711	14

TBX21	0.0014	0.0001	0.2755	0.4035	0.3648	14
E2F2	0.0139	0.0013	0.2057	0.2579	0.2276	14
E2F3	0.0196	0.0019	0.2007	0.2573	0.0737	14
FBN1	0.0014	0.0001	0.3882	0.6154	0.5149	13
TRIM28	0.0236	0.0024	0.2146	0.2667	0.2096	13
HIVEP1	0.0051	0.0004	0.2308	0.3309	0.0559	12
FOXE1	0.0376	0.0043	0.1627	0.2476	0.0719	12
HDAC6	0.0014	0.0001	0.2692	0.4505	0.0799	11
TICAM1	0.0014	0.0001	0.3041	0.4848	0.0619	11
GABPA	0.0139	0.0013	0.1879	0.2647	0.1213	11
IL17RA	0.0299	0.0032	0.1945	0.2571	0.2073	11
PTAFR	0.0014	0.0001	0.3045	0.4909	0.298	10
NLRP12	0.0014	0.0001	0.291	0.4182	0.1362	10
CD86	0.0351	0.0039	0.1999	0.2821	0.1688	10
CBL	0.0014	0.0001	0.3898	0.6667	0.113	9
HSF2	0.0014	0.0001	0.2943	0.5636	0.1572	9
IGF2BP1	0.0014	0.0001	0.4454	0.6182	0.0708	9
MXI1	0.0027	0.0002	0.2921	0.5455	0.1856	9
CXCL2	0.0105	0.0009	0.2587	0.4222	0.1751	9
PEX5L	0.0113	0.001	0.269	0.4	0.0554	9
CEACAM1	0.0252	0.0026	0.2163	0.3091	0.0665	9
S1PR1	0.0391	0.0045	0.1872	0.2692	0.1001	9
ACVR1C	0.0394	0.0046	0.1885	0.2727	0.1475	9
NPC1	0.0213	0.0021	0.2556	0.3611	0.41	8
HRH2	0.0351	0.0039	0.2102	0.3111	0.0429	8
TFDP1	0.0431	0.0051	0.2108	0.3111	0.3236	8
INHBB	0.0014	0.0001	0.349	0.5556	0.2109	7
EPHB1	0.0027	0.0002	0.2979	0.5	0.1241	7
IFNAR1	0.0027	0.0002	0.3115	0.5278	0.2966	7
BID	0.0094	0.0008	0.3151	0.5238	0.13	7
IFNGR1	0.0236	0.0024	0.2176	0.3333	0.237	7
ZBTB7B	0.0236	0.0024	0.312	0.3929	0.1176	7
IL10RA	0.0394	0.0046	0.2364	0.381	0.1994	7
FOSL1	0.0409	0.0048	0.2183	0.3333	0.1899	7
SMOC2	0.0014	0.0001	0.5275	1	0.2699	6
ACVRL1	0.0113	0.001	0.3285	0.5238	0.3632	6
TNNI3	0.0134	0.0012	0.295	0.3889	0.13	6

DSP	0.0188	0.0018	0.3519	0.4762	0.1475	6
RFX5	0.0232	0.0023	0.3405	0.4667	0.3185	6
DDX58	0.0278	0.0029	0.2534	0.3333	0.384	6
ABCG1	0.0443	0.0053	0.2576	0.4	0.1398	6
LRP1	0.0014	0.0001	0.5121	1	0.0483	5
GDF15	0.0014	0.0001	0.531	1	0.1111	5
IFNK	0.0014	0.0001	0.4998	1	0.0474	5
TNXB	0.0051	0.0004	0.5586	0.6667	0.0398	5
FLT1	0.0139	0.0013	0.351	0.5333	0.06	5
USP18	0.0139	0.0013	0.2906	0.5333	0.3677	5
FCGR2B	0.0286	0.003	0.2672	0.4667	0.0915	5
ANXA2	0.0305	0.0033	0.2935	0.4667	0.3118	5
CBFB	0.0014	0.0001	0.584	1	0.1214	4
C1QA	0.0073	0.0006	0.3026	0.8333	0.1332	4
ITGAV	0.0113	0.001	0.4178	0.8333	0.5506	4
E2F8	0.0139	0.0013	0.2867	0.6	0.38	4
RCAN1	0.0252	0.0026	0.3187	0.6667	0.1061	4
ANPEP	0.0014	0.0001	0.4959	1	0.0727	3
CKS1B	0.0014	0.0001	0.644	1	0.5824	3
CKS2	0.0014	0.0001	0.644	1	0.5994	3
GCLC	0.0014	0.0001	0.5574	1	0.0698	3
GCLM	0.0014	0.0001	0.5574	1	0.134	3
GADD45B	0.0014	0.0001	0.3908	1	0.0054	3
PTPRC	0.0014	0.0001	0.5627	1	0.078	3
TCF7	0.0014	0.0001	0.4494	1	0.43	3
UBA7	0.0014	0.0001	0.445	1	0.4173	3
ZBED1	0.0014	0.0001	0.8627	1	0.1592	3
PLAA	0.0014	0.0001	0.3908	1	0.0423	3
MSC	0.0431	0.0051	0.3339	0.5	0.0852	3
UXT	0.0472	0.0057	0.3355	0.6667	0.111	3
FGA	0.0487	0.0059	0.3133	0.6667	0.0271	3
ADORA2B	0.0014	0.0001	0.3253	1	0.0202	2
AMELX	0.0014	0.0001	0.3022	1	0.0789	2
CASP3	0.0014	0.0001	0.6509	1	0.1038	2
CASQ2	0.0014	0.0001	0.2899	1	0.3478	2
CMA1	0.0014	0.0001	0.8021	1	0.0651	2
COL2A1	0.0014	0.0001	0.8021	1	0.281	2

COL4A3	0.0014	0.0001	0.8042	1	0.0216	2
CSK	0.0014	0.0001	0.3253	1	0.1208	2
CYP7A1	0.0014	0.0001	0.3342	1	0.0382	2
LPAR1	0.0014	0.0001	0.3305	1	0.3737	2
EMD	0.0014	0.0001	0.3614	1	0.0415	2
FCGR3A	0.0014	0.0001	0.6148	1	0.508	2
G6PC	0.0014	0.0001	0.4559	1	0.1199	2
HIST1H1C	0.0014	0.0001	0.905	1	0.0473	2
HAS2	0.0014	0.0001	0.3028	1	0.2549	2
CFH	0.0014	0.0001	0.2783	1	0.1831	2
IK	0.0014	0.0001	0.3344	1	0.0554	2
IL12RB1	0.0014	0.0001	0.4116	1	0.1899	2
NMB	0.0014	0.0001	0.5762	1	0.0703	2
PCBP2	0.0014	0.0001	0.6483	1	0.0498	2
PLP1	0.0014	0.0001	0.4066	1	0.1574	2
HTRA1	0.0014	0.0001	0.4329	1	0.1983	2
PTPRG	0.0014	0.0001	0.306	1	0.1091	2
CXCL6	0.0014	0.0001	0.3253	1	0.236	2
TRPM2	0.0014	0.0001	0.5403	1	0.1315	2
TXK	0.0014	0.0001	0.8001	1	0.295	2
EOMES	0.0014	0.0001	0.6204	1	0.6593	2
IRS4	0.0014	0.0001	0.3355	1	0.1116	2
RPS6KA4	0.0014	0.0001	0.5762	1	0.0689	2
F2RL3	0.0014	0.0001	0.6558	1	0.051	2
RPS6KA5	0.0014	0.0001	0.5762	1	0.0487	2
BCL2L11	0.0014	0.0001	0.5626	1	0.1119	2
AGPAT2	0.0014	0.0001	0.277	1	0.2332	2
WDR5	0.0014	0.0001	0.2838	1	0.0597	2
CD160	0.0014	0.0001	0.5762	1	0.154	2
TPSD1	0.0014	0.0001	0.497	1	0.0743	2
MYOZ2	0.0014	0.0001	0.2872	1	0.1113	2
IKZF4	0.0014	0.0001	0.618	1	0.1766	2
MARCKSL1	0.0014	0.0001	0.2899	1	0.1275	2
MCPH1	0.0014	0.0001	0.7106	1	0.1196	2
CMIP	0.0014	0.0001	0.3129	1	0.3021	2
TWIST2	0.0014	0.0001	0.3253	1	0.3092	2
VSX2	0.0014	0.0001	0.5959	1	0	2

APPENDIX D

Regulator	FDR	p-value	ave. corr. (regulatees)	fraction	ave. corr. (regulator - regulatees)	# of coherent regulatees	Found with whole KB
HIF1A	0.0013	0.0001	0.1407	0.1442	0.1381	149	N
TGFB1	0.0013	0.0001	0.1529	0.1697	0.2343	111	Y
IL13	0.0013	0.0001	0.1729	0.1954	0	80	Y
EP300	0.0013	0.0001	0.1538	0.1681	0.0781	75	Y
THAP1	0.0013	0.0001	0.2848	0.4625	0.0959	75	Y
E2F4	0.0013	0.0001	0.1618	0.1852	0.0733	73	Y
CCND1	0.0013	0.0001	0.1671	0.1999	0.1699	58	N
MYCN	0.0013	0.0001	0.1916	0.2482	0.1009	43	Y
TLR4	0.0013	0.0001	0.1721	0.1976	0.3036	42	Y
STAT1	0.0013	0.0001	0.262	0.4228	0.3932	41	Y
IFNA13	0.0013	0.0001	0.2582	0.4153	0	38	Y
JUN	0.0013	0.0001	0.1989	0.2613	0.0942	32	Y
CHUK	0.0013	0.0001	0.2197	0.339	0.0724	31	Y
SMAD3	0.0013	0.0001	0.183	0.2222	0.0853	31	Y
CD40LG	0.0013	0.0001	0.2379	0.3977	0	29	Y
TGFB2	0.0013	0.0001	0.2395	0.3563	0.0659	26	Y
SPI1	0.0013	0.0001	0.3665	0.547	0.5267	24	Y
IL10	0.0013	0.0001	0.259	0.3953	0.3306	19	Y
IFNB1	0.0013	0.0001	0.2366	0.368	0	17	Y
LPL	0.0013	0.0001	0.2468	0.3762	0.1209	16	Y
CSF2RB	0.0013	0.0001	0.2533	0.3743	0	15	Y
ABCB1	0.0013	0.0001	0.2812	0.3737	0.1905	15	Y
IL12B	0.0013	0.0001	0.2469	0.3971	0	13	Y
IL18	0.0013	0.0001	0.2503	0.4381	0.2372	13	Y
ZNF175	0.0013	0.0001	0.4129	0.5604	0.1094	13	Y
E2F3	0.0013	0.0001	0.2646	0.4909	0.1863	11	Y
CTGF	0.0013	0.0001	0.4258	0.6444	0.4182	10	Y
E2F2	0.0013	0.0001	0.2548	0.4545	0.3182	10	Y
IFNA1	0.0013	0.0001	0.2559	0.4545	0	10	Y
TBX21	0.0013	0.0001	0.3759	0.6	0	9	Y
THBS1	0.0013	0.0001	0.3112	0.6429	0.3813	8	Y
TSC22D3	0.0013	0.0001	0.397	0.7143	0.1041	7	Y
IL25	0.0013	0.0001	0.3314	0.6667	0	7	Y

MYOCD	0.0013	0.0001	0.5027	0.9048	0	7	Y
TNXB	0.0013	0.0001	0.8818	1	0.058	6	Y
PRKG1	0.0013	0.0001	0.4543	0.9	0.467	5	Y
ELF1	0.0013	0.0001	0.701	1	0.1731	4	Y
CCL13	0.0013	0.0001	0.4227	1	0.1349	4	Y
TLR2	0.0013	0.0001	0.494	1	0.5222	4	Y
SCARB1	0.0013	0.0001	0.3654	1	0.0638	3	N
IRF8	0.0013	0.0001	0.516	1	0.6405	3	Y
MBP	0.0013	0.0001	0.7215	1	0.1289	3	Y
MYD88	0.0013	0.0001	0.4956	1	0.199	3	Y
PTN	0.0013	0.0001	0.4264	1	0.0659	3	Y
XDH	0.0013	0.0001	0.3586	1	0.1566	3	Y
CALCB	0.0013	0.0001	0.7026	1	0	2	N
ERBB3	0.0013	0.0001	0.67	1	0.0291	2	Y
ERBB4	0.0013	0.0001	0.67	1	0.1921	2	Y
F7	0.0013	0.0001	0.8118	1	0	2	Y
FLT1	0.0013	0.0001	0.3567	1	0.2615	2	N
HOXB7	0.0013	0.0001	0.6509	1	0.0349	2	Y
ITGA5	0.0013	0.0001	0.2736	1	0.6176	2	Y
LDLR	0.0013	0.0001	0.4349	1	0.2112	2	Y
NFKBIA	0.0013	0.0001	0.3253	1	0.348	2	Y
NRAS	0.0013	0.0001	0.4514	1	0.1129	2	N
SERPINE1	0.0013	0.0001	0.3655	1	0.6957	2	N
PTGIR	0.0013	0.0001	0.312	1	0	2	Y
THRB	0.0013	0.0001	0.4044	1	0.0509	2	N
FOSL1	0.0013	0.0001	0.2736	1	0	2	Y
CDKN1A	0.0023	0.0002	0.1492	0.155	0.1159	104	Y
SRC	0.0023	0.0002	0.1719	0.2074	0.1101	45	Y
IL6	0.0023	0.0002	0.1646	0.2185	0.2013	42	Y
NRF1	0.0023	0.0002	0.1762	0.2141	0	40	Y
IL4	0.0023	0.0002	0.1802	0.2282	0	33	Y
LTA	0.0023	0.0002	0.2255	0.3684	0	15	Y
RB1	0.0023	0.0002	0.2011	0.3203	0.1474	13	Y
CD40	0.0023	0.0002	0.3043	0.5556	0.2403	9	Y
FAS	0.0023	0.0002	0.2984	0.5333	0.3199	8	Y
SPIC	0.0023	0.0002	0.4684	0.75	0	8	Y
HOXD3	0.0032	0.0003	0.1856	0.2371	0.1139	31	Y
HIVEP1	0.0032	0.0003	0.257	0.4	0.0605	11	Y
HMOX1	0.0032	0.0003	0.2907	0.5556	0.3727	8	Y

FCGR2A	0.0032	0.0003	0.3683	0.7	0.5194	5	Y
DYSF	0.0041	0.0004	0.1515	0.1644	0.1377	73	Y
TAF4	0.0041	0.0004	0.2142	0.3333	0.1358	11	N
PIAS1	0.0041	0.0004	0.3114	0.6667	0.0665	6	Y
S100B	0.0049	0.0005	0.1737	0.2011	0.1289	34	Y
CASP8	0.0049	0.0005	0.3896	0.8	0.0726	5	Y
IL12A	0.0049	0.0005	0.3065	0.7	0.2182	5	Y
E2F8	0.0049	0.0005	0.3305	0.8333	0.4369	4	Y
IL5	0.0058	0.0006	0.3222	0.6	0	4	Y
IFNA2	0.0065	0.0007	0.2511	0.2917	0	11	Y
IL21	0.0065	0.0007	0.2935	0.5	0	8	Y
FN1	0.0065	0.0007	0.3018	0.5	0.5166	7	N
TLR3	0.0082	0.0009	0.3189	0.5238	0.3669	6	Y
LIPG	0.0089	0.001	0.1905	0.2619	0.1586	16	Y
ANXA2	0.0089	0.001	0.3253	0.6	0.2871	5	Y
BARX1	0.0096	0.0011	0.2287	0.4167	0	7	Y
CD200	0.0096	0.0011	0.3108	0.6667	0.1463	4	N
RUNX2	0.0103	0.0012	0.2803	0.4643	0.4012	6	N
FADD	0.0103	0.0012	0.3211	0.6	0.0721	5	Y
AHR	0.011	0.0013	0.3503	0.4762	0.1746	5	N
IL1B	0.0117	0.0014	0.1467	0.1584	0.2313	53	Y
LEP	0.0131	0.0016	0.1413	0.1393	0	74	N
SMAD4	0.0131	0.0016	0.1955	0.2745	0.0893	16	Y
PIM1	0.0138	0.0017	0.2639	0.4762	0.2291	5	N
GABPA	0.0143	0.0018	0.2288	0.3889	0.0842	9	N
PTAFR	0.0143	0.0018	0.3062	0.5333	0.2799	5	Y
CSF2	0.0157	0.0021	0.1484	0.1454	0	63	Y
CD34	0.0157	0.0021	0.1675	0.209	0.1642	24	Y
ILK	0.0157	0.0021	0.2715	0.3889	0.128	6	Y
DDX58	0.0157	0.002	0.3862	0.6667	0.469	4	Y
OLR1	0.0157	0.0021	0.3848	0.6667	0.4439	4	Y
PRRX1	0.0157	0.0021	0.3381	0.6667	0.522	4	N
CASP10	0.0167	0.0023	0.3528	0.6667	0.1002	4	Y
HMGA1	0.0167	0.0023	0.2474	0.6667	0.081	4	N
CD200R1	0.0167	0.0023	0.3108	0.6667	0	4	Y
DLX5	0.0173	0.0024	0.2702	0.4667	0.1014	5	Y
MSX2	0.0178	0.0025	0.2515	0.3929	0.2244	6	N
RHOA	0.0189	0.0027	0.1442	0.1497	0.1293	53	N
APOE	0.0189	0.0027	0.3483	0.4286	0.3621	7	Y

HNRNPK	0.0193	0.0028	0.317	0.4286	0	5	N
NOS2	0.0193	0.0028	0.2793	0.5	0	4	N
PRKCB	0.0196	0.0029	0.2284	0.4286	0	6	Y
CCR2	0.0196	0.0029	0.2653	0.4667	0.331	5	Y
JAK1	0.0213	0.0032	0.1483	0.1433	0.1727	61	N
PDGFB	0.0213	0.0032	0.1971	0.2574	0.2283	13	Y
RFX5	0.0224	0.0034	0.3977	0.4667	0.3275	6	Y
PARP1	0.0242	0.0037	0.2497	0.5	0.1925	4	N
PTGS2	0.0253	0.004	0.1376	0.1408	0.1512	57	Y
S1PR1	0.0253	0.004	0.1915	0.3091	0	9	Y
FCGR1A	0.0253	0.0039	0.2761	0.5	0.4286	4	N
IRF1	0.0253	0.004	0.3039	0.5	0.4653	4	Y
HRH2	0.0257	0.0041	0.2575	0.4667	0	6	Y
ATF2	0.0261	0.0042	0.1853	0.2571	0.0779	11	N
TNFSF10	0.0263	0.0043	0.1846	0.2692	0.2691	9	Y
HSF2	0.0263	0.0043	0.2125	0.3333	0.2117	8	Y
KLF2	0.0267	0.0044	0.2006	0.2821	0.2907	10	Y
E2F1	0.0275	0.0046	0.1392	0.1344	0.198	74	Y
HOXA13	0.0275	0.0046	0.1418	0.1513	0	48	N
IL10RA	0.0279	0.0047	0.3094	0.5	0.4879	4	Y
PPARGC1A	0.0304	0.0052	0.1534	0.165	0.0635	33	N
GLI2	0.0304	0.0052	0.2832	0.5	0.164	4	N
FOXP3	0.0309	0.0054	0.1676	0.2035	0	13	Y
EDNRA	0.0309	0.0054	0.3741	0.5	0.3249	4	Y
AHRR	0.0309	0.0054	0.2525	0.5	0	4	Y
IRF9	0.0312	0.0055	0.2424	0.4	0.2781	6	Y
CXCL2	0.0313	0.0056	0.2688	0.381	0.3839	6	Y
CCL2	0.0313	0.0056	0.2608	0.5	0.3994	4	Y
CBS	0.0322	0.0058	0.2461	0.381	0.0381	6	N
ING4	0.0331	0.006	0.2578	0.5	0.0801	3	Y
CXCL3	0.0334	0.0061	0.2034	0.3571	0.2002	8	Y
FHL2	0.0343	0.0063	0.2648	0.5	0.1757	4	N
AIF1	0.0351	0.0065	0.3262	0.6667	0.0657	3	N
FGFR1	0.0362	0.0068	0.2557	0.4	0.1183	4	N
MAPK10	0.0362	0.0068	0.2406	0.5	0.0968	4	N
VEGFA	0.0365	0.0069	0.1507	0.1628	0.0968	37	Y
TRIM28	0.0373	0.0071	0.205	0.2476	0.1972	12	N
HBP1	0.0376	0.0072	0.1994	0.2527	0.1089	13	Y

RAC1	0.0382	0.0077	0.1562	0.1714	0.0464	25	N
PTH	0.0382	0.0076	0.2063	0.2727	0	8	Y
CGB5	0.0382	0.0077	0.2488	0.3214	0	5	N
GFI1	0.0382	0.0076	0.3233	0.6667	0	3	N
TGIF1	0.0382	0.0076	0.3088	0.6667	0.0952	3	N
ETV4	0.0382	0.0077	0.2617	0.5	0.0706	3	N
SMAD1	0.0382	0.0077	0.218	0.5	0.1117	3	Y
TFDP1	0.0382	0.0077	0.3476	0.6667	0.3058	3	Y
ADRB1	0.0384	0.0078	0.1558	0.1651	0	31	N
FGF1	0.0396	0.0081	0.1453	0.1585	0.2101	36	Y
PXMP3	0.0399	0.0082	0.1707	0.2418	0	11	N
FTH1	0.0406	0.0084	0.2025	0.2105	0.2294	12	N
GABPB1	0.0432	0.009	0.1959	0.4	0	4	Y
HP	0.0444	0.0093	0.2544	0.4	0.1587	4	Y
BACH1	0.0446	0.0094	0.2104	0.3333	0.1254	5	N
CEBPD	0.0448	0.0095	0.1724	0.2576	0.2326	10	N
CSF1	0.045	0.0096	0.1649	0.1664	0.2421	28	Y
ADRB2	0.0452	0.0097	0.1484	0.1564	0.1592	34	N
SMARCA4	0.0482	0.0104	0.1343	0.1223	0.1351	70	N
NRIP1	0.0483	0.0105	0.1318	0.1123	0.0551	121	N
ETS2	0.0485	0.0106	0.1737	0.2308	0.1097	10	N

APPENDIX E

Gene Symbol	Entrez ID	Known drivers in breast cancer	Drivers in other cancer types	Candidate drivers identified by GWAS	Involved in cancer realted pathways	TCGA
AATF	26574					
ABL1	25		Y		Y	
AKT1*	207	Y	Y		Y	Y
AKT2	208		Y		Y	
ARAF	369				Y	
BAK1	578					
BCL2	596		Y		Y	
BCL2L11	10018					
BRAF	673		Y		Y	
BRCA1	672	Y	Y	Y		Y
BRCA2	675	Y	Y	Y		
BRIP1*	83990	Y				
CDK5	1020					
CDKN1B	1027			Y	Y	Y
CDKN2A	1029		Y	Y		Y
CHEK1	1111					Y
CHUK	1147				Y	
DAXX	1616		Y		Y	
DMAP1	55929					
EP300	2033	Y	Y	Y	Y	
FOXO3	2309		Y			Y
GSK3B	2932				Y	
HGS	9146					
HSP90AB1	3326		Y		Y	
HSPB1	3315				Y	
IKBKB	3551				Y	Y
ING5	84289					
IRS1	3667			Y		
JAK2	3717		Y			
KAT2B	8850					
KRAS*	3845	Y	Y	Y	Y	
MAP2K1	5604				Y	
MAP2K2	5605				Y	
MAP2K4	6416	Y	Y		Y	Y

MAP3K8	1326		Y		Y	
MAPK1	5594				Y	
MAPK10	5602				Y	
MAPK11	5600				Y	
MAPK3	5595				Y	
MAPK9	5601				Y	
MDM2	4193		Y	Y		
MDM4	4194		Y			
MLL	4297		Y			
MSH2	4436		Y			
MTOR	2475				Y	Y
NF2	4771		Y			
NFKBIA	4792				Y	Y
NR3C1	2908			Y		
PAK1	5058				Y	Y
PDGFRB	5159		Y		Y	
PIAS4	51588					
PIK3CA*	5290	Y	Y	Y	Y	Y
PIK3R1	5295		Y		Y	Y
PML	5371		Y			
PPM1D*	8493	Y				
PTGS2	5743			Y	Y	
PTK2	5747				Y	
PTTG1	9232					
RAF1	5894		Y		Y	
RALGDS	5900		Y			
RASA1	5921		Y		Y	
RB1	5925	Y	Y	Y		Y
RB1CC1*	9821	Y				
RPS6KB1	6198				Y	Y
S100A4	6275					
S100B	6285					
SLC22A18*	5002	Y	Y			
SOS1	6654				Y	
SRC	6714		Y		Y	
TCL1A	8115		Y			
THRB	7068					
TP53*	7157	Y	Y	Y	Y	
TP53BP1	7158					

TP53INP1	94241					
TSC2	7249		Y			
TSG101*	7251	Y				
VEGFA	7422			Y	Y	
VPS28	51160					
VPS36	51028					
VRK1	7443					
WWOX	51741		Y			Y
YBX1	4904					

APPENDIX F

KEGG pathway	Number of genes in breast cancer module	Genes	p-value	FDR
ErbB signaling pathway	24	GSK3B, RPS6KB1, PIK3R1, AKT1, PIK3CA, MAP2K2, RAF1, BRAF, ABL1, SOS1, MTOR, MAPK1, MAPK9, PAK1, KRAS, ARAF, MAP2K1, AKT2, MAP2K4, CDKN1B, MAPK10, SRC, PTK2, MAPK3	< 1.00E-10	< 1.00E-10
MAPK signaling pathway	23	HSPB1, TP53, AKT1, MAP2K2, RAF1, BRAF, SOS1, PDGFRB, MAPK1, MAPK9, MAPK11, PAK1, KRAS, MAP2K1, IKBKB, MAP3K8, AKT2, RASA1, MAP2K4, CHUK, MAPK10, DAXX, MAPK3	< 1.00E-10	< 1.00E-10
Focal adhesion	19	GSK3B, PIK3R1, VEGFA, AKT1, PIK3CA, RAF1, BRAF, SOS1, PDGFRB, MAPK1, MAPK9, PAK1, MAP2K1, AKT2, MAPK10, SRC, PTK2, BCL2, MAPK3	< 1.00E-10	< 1.00E-10
VEGF signaling pathway	16	HSPB1, PIK3R1, VEGFA, AKT1, PTGS2, PIK3CA, MAP2K2, RAF1, MAPK1, MAPK11, KRAS, MAP2K1, AKT2, SRC, PTK2, MAPK3	< 1.00E-10	< 1.00E-10
Regulation of actin cytoskeleton	15	PIK3R1, PIK3CA, MAP2K2, RAF1, BRAF, SOS1, PDGFRB, MAPK1, PAK1, KRAS, ARAF, MAP2K1, SRC, PTK2, MAPK3	< 1.00E-10	< 1.00E-10
Progesterone-mediated oocyte maturation	15	HSP90AB1, PIK3R1, AKT1, PIK3CA, RAF1, BRAF, MAPK1, MAPK9, MAPK11, KRAS, ARAF, MAP2K1, AKT2, MAPK10, MAPK3	< 1.00E-10	< 1.00E-10
GnRH signaling pathway	12	MAP2K2, RAF1, SOS1, MAPK1, MAPK9, MAPK11, KRAS, MAP2K1, MAP2K4, MAPK10, SRC, MAPK3	< 1.00E-10	< 1.00E-10
Apoptosis	9	PIK3R1, TP53, AKT1, PIK3CA, IKBKB, AKT2, CHUK, BCL2, NFKBIA	< 1.00E-10	< 1.00E-10
Adherens junction	4	EP300, MAPK1, SRC, MAPK3	0.000170277	0.000779691

APPENDIX G

Gene Symbol	Entrez ID	Known drivers in ovarian cancer	Drivers in other cancer types	Candidate drivers identified by GWAS	Involved in cancer realted pathways	TCGA
ABL1	25		Y		Y	Y
AKT1*	207	Y	Y		Y	Y
APC	324		Y		Y	
BRAF	673		Y		Y	
CBLL1	79872					
CDH1*	999	Y	Y		Y	Y
CDK5	1020					Y
CSK	1445					
CTNNB1*	1499	Y	Y		Y	
EGFR	1956		Y		Y	
ERBB2*	2064	Y	Y		Y	
ERBB2IP	55914					
ERBB3	2065				Y	
ERBB4	2066				Y	
ESR1	2099		Y	Y		
FYN	2534				Y	
GSK3B	2932				Y	
IGF1R	3480				Y	Y
IKBKB	3551				Y	
INSR	3643				Y	
JAK2	3717		Y			
KLRG1	10219					
LCK	3932		Y			
MAP2K1	5604				Y	
MAP2K2	5605				Y	
MAP2K4	6416		Y		Y	
MAP3K1	4214				Y	
MAPK1	5594				Y	
MAPK14	1432				Y	
MAPK3	5595				Y	
MDM4	4194		Y			
MLST8	64223					
MRTO4	51154					
MTOR	2475				Y	

NCOA7	135112					
NOC2L	26155					
PDE3B	5140				Y	
PIK3CA*	5290	Y	Y		Y	
PIK3R1	5295	Y	Y		Y	Y
PRKCA	5578				Y	
PTK2	5747				Y	
PTK2B	2185				Y	
PTK6	5753					Y
RAF1	5894		Y		Y	Y
RRAS2*	22800	Y			Y	
SRC	6714		Y		Y	
ST5	6764					
TCF7L1	83439				Y	
TCF7L2	6934		Y		Y	Y
THOC3	84321					
TJP1	7082				Y	
TRAF3IP2	10758					
WDR12	55759					
WDR31	114987					
WDR51A	25886					Y
WDR51B	282809					
WDR69	164781					
WNK1	65125					Y

APPENDIX H

KEGG pathway	Number of genes in breast cancer module	Genes	p-value	FDR
ErbB signaling pathway	20	GSK3B, PIK3R1, AKT1, PRKCA, PIK3CA, MAP2K2, RAF1, BRAF, ABL1, ERBB3, MTOR, MAPK1, ERBB4, MAP2K1, ERBB2, MAP2K4, SRC, PTK2, EGFR, MAPK3	< 1.00E-10	< 1.00E-10
Focal adhesion	17	GSK3B, PIK3R1, CTNNB1, AKT1, PRKCA, PIK3CA, RAF1, BRAF, MAPK1, IGF1R, MAP2K1, ERBB2, FYN, SRC, PTK2, EGFR, MAPK3	< 1.00E-10	< 1.00E-10
MAPK signaling pathway	14	AKT1, PRKCA, MAP2K2, RAF1, BRAF, MAPK1, MAP3K1, MAP2K1, IKBKB, MAP2K4, MAPK14, EGFR, RRAS2, MAPK3	< 1.00E-10	< 1.00E-10
Regulation of actin cytoskeleton	13	PIK3R1, PIK3CA, MAP2K2, RAF1, BRAF, MAPK1, APC, MAP2K1, SRC, PTK2, EGFR, RRAS2, MAPK3	< 1.00E-10	< 1.00E-10
Adherens junction	13	TJP1, CDH1, CTNNB1, TCF7L1, TCF7L2, INSR, MAPK1, IGF1R, ERBB2, FYN, SRC, EGFR, MAPK3	< 1.00E-10	< 1.00E-10
GnRH signaling pathway	12	PRKCA, MAP2K2, RAF1, PTK2B, MAPK1, MAP3K1, MAP2K1, MAP2K4, SRC, MAPK14, EGFR, MAPK3	< 1.00E-10	< 1.00E-10
VEGF signaling pathway	12	PIK3R1, AKT1, PRKCA, PIK3CA, MAP2K2, RAF1, MAPK1, MAP2K1, SRC, PTK2, MAPK14, MAPK3	< 1.00E-10	< 1.00E-10
Progesterone-mediated oocyte maturation	11	PIK3R1, AKT1, PIK3CA, RAF1, BRAF, MAPK1, PDE3B, IGF1R, MAP2K1, MAPK14, MAPK3	< 1.00E-10	< 1.00E-10
Apoptosis	4	PIK3R1, AKT1, PIK3CA, IKBKB	8.35E-05	0.00041493

Bibliography

1. Shen-Orr SS, Milo R, Mangan S, Alon U: **Network motifs in the transcriptional regulation network of Escherichia coli.** *Nature genetics* 2002, **31**:64–8.

2. Alon U: **Network motifs: theory and experimental approaches.** *Nature reviews. Genetics* 2007, **8**:450–61.

3. Newman MEJ: **The Structure and Function of Complex Networks**. *SIAM Review* 2003, **45**:167–256.

4. Clauset A, Shalizi CR, Newman MEJ: **Power-Law Distributions in Empirical Data**. *SIAM Review* 2009, **51**:661–703.

5. Barabási A-L, Gulbahce N, Loscalzo J: **Network medicine: a network-based approach to human disease.** *Nature reviews. Genetics* 2011, **12**:56–68.

6. Snuderl M, Fazlollahi L, Le LP, Nitta M, Zhelyazkova BH, Davidson CJ, Akhavanfard S, Cahill DP, Aldape KD, Betensky RA, Louis DN, Iafrate AJ: **Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma.** *Cancer cell* 2011, **20**:810–7.

7. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406**:747–52.

8. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lønning PE, Børresen-Dale AL: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**:10869–74.

9. HANAHAN D, Weinberg RA: **The Hallmarks of Cancer**. *Cell* 2000, **100**:57–70.

10. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation.** *Cell* 2011, **144**:646–74.

11. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW: **Cancer genome landscapes.** *Science (New York, N.Y.)* 2013, **339**:1546–58.

12. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK, Ding L: **MuSiC: identifying mutational significance in cancer genomes.** *Genome research* 2012, **22**:1589–98.

13. Oscar M. Rueda RD-U: **Finding recurrent copy number alteration regions: a review of methods.** *Current Bioinformatics* 2010, **5**:1–17.

14. Leary RJ, Lin JC, Cummins J, Boca S, Wood LD, Parsons DW, Jones S, Sjöblom T, Park B-H, Parsons R, Willis J, Dawson D, Willson JK V, Nikolskaya T, Nikolsky Y, Kopelovich L, Papadopoulos N, Pennacchio LA, Wang T-L, Markowitz SD, Parmigiani G, Kinzler KW, Vogelstein B, Velculescu VE: **Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105**:16224–9.

15. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G: **GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.** *Genome biology* 2011, **12**:R41.

16. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R: **Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations.** *Cancer research* 2009, **69**:6660–7.

17. Gonzalez-Perez A, Lopez-Bigas N: **Functional impact bias reveals cancer drivers.** *Nucleic acids research* 2012, **40**:e169.

18. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew Y-E, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan M-H, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR: **Patterns of somatic mutation in human cancer genomes.** *Nature* 2007, **446**:153–8.

19. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N: **OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes.** *Bioinformatics (Oxford, England)* 2013, **29**:2238–44.

20. Reimand J, Bader GD: **Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers.** *Molecular systems biology* 2013, **9**:637.

21. Ciriello G, Cerami E, Sander C, Schultz N: **Mutual exclusivity analysis identifies oncogenic network modules.** *Genome research* 2012, **22**:398–406.

22. Ng S, Collisson EA, Sokolov A, Goldstein T, Gonzalez-Perez A, Lopez-Bigas N, Benz C, Haussler D, Stuart JM: **PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis.** *Bioinformatics (Oxford, England)* 2012, **28**:i640–i646.

23. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**:15545–50.

24. Hung J-H, Whitfield TW, Yang T-H, Hu Z, Weng Z, Delisi C: **Identification of functional modules that correlate with phenotypic difference: the influence of network topology.** *Genome biology* 2010, **11**:R23.

25. Naeem H, Küffner R, Zimmer R: **MIRTFnet: analysis of miRNA regulated transcription factors.** *PloS one* 2011, **6**:e22519.

26. Liu Q, Tan Y, Huang T, Ding G, Tu Z, Liu L, Li Y, Dai H, Xie L: **TF-centered downstream gene set enrichment analysis: Inference of causal regulators by integrating TF-DNA interactions and protein post-translational modifications information.** *BMC bioinformatics* 2010, **11 Suppl 1**:S5.

27. Essaghir A, Toffalini F, Knoops L, Kallin A, van Helden J, Demoulin J-B: **Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data.** *Nucleic acids research* 2010, **38**:e120.

28. Sherman BT, Huang DW, Tan Q, Guo Y, Bour S, Liu D, Stephens R, Baseler MW, Lane HC, Lempicki RA: **DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis.** *BMC bioinformatics* 2007, **8**:426.

29. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH: **PID: the Pathway Interaction Database.** *Nucleic acids research* 2009, **37**:D674–9.

30. Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story M-A, Smith B: **The National Center for Biomedical Ontology.** *Journal of the American Medical Informatics Association : JAMIA* , **19**:190–5.

31. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone S-A, Scheuermann RH, Shah N, Whetzel PL, Lewis S: **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.** *Nature biotechnology* 2007, **25**:1251–5.

32. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, May B, Schmidt E, Vastrik I, Wu G, Birney E, Stein L, D’Eustachio P: **Reactome knowledgebase of human biological pathways and processes.** *Nucleic acids research* 2009, **37**:D619–22.

33. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic acids research* 2000, **28**:27–30.

34. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M: **KEGG for integration and interpretation of large-scale molecular data sets.** *Nucleic acids research* 2012, **40**:D109–14.

35. **The Cancer Genome Atlas** [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2821689&tool=pmcentrez&rendertype=abstract].

36. Zhao W, Langfelder P, Fuller T, Dong J, Li A, Hovarth S: **Weighted gene coexpression network analysis: state of the art.** *Journal of biopharmaceutical statistics* 2010, **20**:281–300.

37. Margolin A a, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A: **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.** *BMC bioinformatics* 2006, **7 Suppl 1**:S7.

38. Friedman N, Linial M, Nachman I, Pe’er D: **Using Bayesian networks to analyze expression data.** *Journal of computational biology : a journal of computational molecular cell biology* 2000, **7**:601–20.

39. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data.** *Science (New York, N.Y.)* 2003, **302**:449–53.

40. Cosgrove EJ, Zhou Y, Gardner TS, Kolaczyk ED: **Predicting gene targets of perturbations via network-based filtering of mRNA expression compendia.** *Bioinformatics (Oxford, England)* 2008, **24**:2482–90.

41. De Smet R, Marchal K: **Advantages and limitations of current network inference methods.** *Nature reviews. Microbiology* 2010, **8**:717–29.

42. Yu H, Luscombe NM, Qian J, Gerstein M: **Genomic analysis of gene expression relationships in transcriptional regulatory networks.** *Trends in genetics : TIG* 2003, **19**:422–7.

43. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ: **An integrative genomics approach to infer causal associations between gene expression and disease.** *Nature genetics* 2005, **37**:710–7.

44. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway L a, Pe’er D: **An Integrated Approach to Uncover Drivers of Cancer.** *Cell* 2010, **143**:1005–1017.

45. **National Library of Medicine** [http://www.nlm.nih.gov/].

46. Mani KM, Lefebvre C, Wang K, Lim WK, Basso K, Dalla-Favera R, Califano A: **A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas.** *Molecular systems biology* 2008, **4**:169.

47. Wang K, Saito M, Bisikirska BC, Alvarez MJ, Lim WK, Rajbhandari P, Shen Q, Nemenman I, Basso K, Margolin A a, Klein U, Dalla-Favera R, Califano A: **Genome-wide identification of post-translational modulators of transcription factor activity in human B cells.** *Nature biotechnology* 2009, **27**:829–39.

48. Hanisch D, Zien A, Zimmer R, Lengauer T: **Co-clustering of biological networks and gene expression data.** *Bioinformatics (Oxford, England)* 2002, **18 Suppl 1**:S145–54.

49. Ulitsky I, Shamir R: **Identification of functional modules using network topology and high-throughput data.** *BMC systems biology* 2007, **1**:8.

50. Bell D, Berchuck A, Birrer M, et al.: **Integrated genomic analyses of ovarian carcinoma**. *Nature* 2011, **474**:609–615.

51. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiriksdottir GH, Bjornsdottir G, Reynisdottir I, Gudbjartsson D, Helgadottir A, Jonasdottir AA, Styrkarsdottir U, Gretarsdottir S, Magnusson KP, Stefansson H, Fossdal R, Kristjansson K, Gislason HG, Stefansson T, Leifsson BG, Thorsteinsdottir U, Lamb JR, Gulcher JR, Reitman ML, Kong A, Schadt EE, Stefansson K: **Genetics of gene expression and its effect on disease.** *Nature* 2008, **452**:423–8.

52. Coiffier B, Lepage E, Briere J, Herbrecht R, Tilly H, Bouabdallah R, Morel P, Van Den Neste E, Salles G, Gaulard P, Reyes F, Lederlin P, Gisselbrecht C: **CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma.** *The New England journal of medicine* 2002, **346**:235–42.

53. **Ingenuity Inc.** [www.ingenuity.com].

54. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic acids research* 2005, **33**:D54–8.

55. **Selventa Inc.** [http://selventa.com/].

56. Viger F, Latapy M: **Fast generation of random connected graphs with prescribed degrees**. In *Computing and Combinatorics Conference*. Springer; 2005, **3595**:1–21.

57. Milo R, Kashtan N, Itzkovitz S, Newman MEJ, Alon U: **On the uniform generation of random graphs with prescribed degree sequences**. 2003.

58. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I: **Controlling the false discovery rate in behavior genetics research.** *Behavioural brain research* 2001, **125**:279–84.

59. Iman RL, Conover WJ: **A distribution-free approach to inducing rank correlation among input variables**. *Communications in Statistics - Simulation and Computation* 1982, **11**:311–334.

60. **Gene Expression Omnibus** [http://www.ncbi.nlm.nih.gov/geo/].

61. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O'Kelly M, Tamayo P, Weir BA, Gabriel S, Winckler W, Gupta S, Jakkula L, Feiler HS, Hodgson JG, James CD, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, Getz G, Perou CM, Hayes DN: **Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.** *Cancer cell* 2010, **17**:98–110.

62. Bø TH, Dysvik B, Jonassen I: **LSimpute: accurate estimation of missing values in microarray data with least squares methods.** *Nucleic acids research* 2004, **32**:e34.

63. Chen L, Xuan J, Riggins RB, Clarke R, Wang Y: **Identifying cancer biomarkers by network-constrained support vector machines.** *BMC systems biology* 2011, **5**:161.

64. Zhang J, Yang Y, Wang Y, Zhang J, Wang Z, Yin M, Shen X: **Identification of hub genes related to the recovery phase of irradiation injury by microarray and integrated gene network analysis.** *PloS one* 2011, **6**:e24680.

65. Kim S-J, Lee K-H, Lee Y-S, Mun E-G, Kwon D-Y, Cha Y-S: **Transcriptome analysis and promoter sequence studies on early adipogenesis in 3T3-L1 cells.** *Nutrition research and practice* 2007, **1**:19–28.

66. Pi J, Leung L, Xue P, Wang W, Hou Y, Liu D, Yehuda-Shnaidman E, Lee C, Lau J, Kurtz TW, Chan JY: **Deficiency in the nuclear factor E2-related factor-2 transcription factor results in impaired adipogenesis and protects against diet-induced obesity.** *The Journal of biological chemistry* 2010, **285**:9292–300.

67. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, Leonardson A, Castellini LW, Wang S, Champy M-F, Zhang B, Emilsson V, Doss S, Ghazalpour A, Horvath S, Drake TA, Lusis AJ, Schadt EE: **Variations in DNA elucidate molecular networks that cause disease.** *Nature* 2008, **452**:429–35.

68. Madani R, Karastergiou K, Ogston NC, Miheisi N, Bhome R, Haloob N, Tan GD, Karpe F, Malone-Lee J, Hashemi M, Jahangiri M, Mohamed-Ali V: **RANTES release by**

human adipose tissue in vivo and evidence for depot-specific differences. *American journal of physiology. Endocrinology and metabolism* 2009, **296**:E1262–8.

69. Kulbe H, Thompson R, Wilson JL, Robinson S, Hagemann T, Fatah R, Gould D, Ayhan A, Balkwill F: **The inflammatory cytokine tumor necrosis factor-alpha generates an autocrine tumor-promoting network in epithelial ovarian cancer cells.** *Cancer research* 2007, **67**:585–92.

70. Qian B-Z, Li J, Zhang H, Kitamura T, Zhang J, Campion LR, Kaiser EA, Snyder LA, Pollard JW: **CCL2 recruits inflammatory monocytes to facilitate breast-tumour metastasis.** *Nature* 2011.

71. Stein B, Kung Sutherland MS: **IL-6 as a drug discovery target**. *Drug Discovery Today* 1998, **3**:202–213.

72. Lejeune FJ, Liénard D, Matter M, Rüegg C: **Efficiency of recombinant human TNF in human cancer therapy.** *Cancer immunity : a journal of the Academy of Cancer Immunology* 2006, **6**:6.

73. Ferrara N: **VEGF as a therapeutic target in cancer.** *Oncology* 2005, **69 Suppl 3**:11–6.

74. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R: **Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations.** *Cancer research* 2009, **69**:6660–7.

75. Cerami E, Demir E, Schultz N, Taylor BS, Sander C: **Automated network analysis identifies core pathways in glioblastoma.** *PloS one* 2010, **5**:e8918.

76. Torkamani A, Schork NJ: **Identification of rare cancer driver mutations by network reconstruction.** *Genome research* 2009, **19**:1570–8.

77. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R: **Associating genes and protein complexes with disease via network propagation.** *PLoS computational biology* 2010, **6**:e1000641.

78. Oti M, Snel B, Huynen MA, Brunner HG: **Predicting disease genes using protein-protein interactions.** *Journal of medical genetics* 2006, **43**:691–8.

79. Linghu B, Snitkin ES, Hu Z, Xia Y, Delisi C: **Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network.** *Genome biology* 2009, **10**:R91.

80. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C: **STRING 8--a global view on proteins and**

their functional interactions in 630 organisms. *Nucleic acids research* 2009, **37**:D412–6.

81. Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR, Kasif S: **Whole-genome annotation by using evidence integration in functional-linkage networks.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**:2888–93.

82. Wang L, Tu Z, Sun F: **A network-based integrative approach to prioritize reliable hits from multiple genome-wide RNAi screens in Drosophila.** *BMC genomics* 2009, **10**:220.

83. Huttenhower C, Haley EM, Hibbs MA, Dumeaux V, Barrett DR, Coller HA, Troyanskaya OG: **Exploring the human genome with functional maps.** *Genome research* 2009, **19**:1093–106.

84. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L: **The human disease network.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:8685–90.

85. Abadie V, Sollid LM, Barreiro LB, Jabri B: **Integration of genetic and immunological insights into a model of celiac disease pathogenesis.** *Annual review of immunology* 2011, **29**:493–525.

86. 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 systems biology* 2012, **6**:64.

87. Mitra K, Carvunis A-R, Ramesh SK, Ideker T: **Integrative approaches for finding modular structure in biological networks**. *Nature Reviews Genetics* 2013, **14**:719–732.

88. Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C: **Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes.** *American journal of human genetics* 2006, **78**:1011–25.

89. **Online Mendelian Inheritance in Man** [http://www.ncbi.nlm.nih.gov/omim].

90. Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR: **The Catalogue of Somatic Mutations in Cancer (COSMIC).** *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]* 2008, **Chapter 10**:Unit 10.11.

91. **Genetic association database** [http://geneticassociationdb.nih.gov/].

92. Visscher PM, Brown MA, McCarthy MI, Yang J: **Five years of GWAS discovery.** *American journal of human genetics* 2012, **90**:7–24.

93. Pearson TA, Manolio TA: **How to interpret a genome-wide association study.** *JAMA : the journal of the American Medical Association* 2008, **299**:1335–44.

94. Newman MEJ: **Modularity and community structure in networks.** *Proceedings of the National Academy of Sciences of the United States of America* 2006, **103**:8577–82.

95. **KEGG PATHWAY Database** [http://www.genome.jp/kegg/pathway.html].

96. Hu Z, Hung J-H, Wang Y, Chang Y-C, Huang C-L, Huyck M, DeLisi C: **VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology.** *Nucleic acids research* 2009, **37**:W115–21.

97. Huang DW, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic acids research* 2009, **37**:1–13.

98. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nature protocols* 2009, **4**:44–57.

99. Krippl P, Langsenlehner U, Renner W, Yazdani-Biuki B, Wolf G, Wascher TC, Paulweber B, Haas J, Samonigg H: **A common 936 C/T gene polymorphism of vascular endothelial growth factor is associated with decreased breast cancer risk.** *International journal of cancer. Journal international du cancer* 2003, **106**:468–71.

100. Luo M, Guan J-L: **Focal adhesion kinase: a prominent determinant in breast cancer initiation, progression and metastasis.** *Cancer letters* 2010, **289**:127–39.

101. Chen C-T, Du Y, Yamaguchi H, Hsu J-M, Kuo H-P, Hortobagyi GN, Hung M-C: **Targeting the IKKβ/mTOR/VEGF signaling pathway as a potential therapeutic strategy for obesity-related breast cancer.** *Molecular cancer therapeutics* 2012, **11**:2212–21.

102. Lo H-W, Wang S-C, Hung M-C: **Novel Signaling Pathways in Breast Cancer**. In *Breast Cancer and Molecular Medicine*. edited by Piccart MJ, Hung M-C, Solin LJ, Cardoso F, Wood WC Springer Berlin Heidelberg; 2006:823–839.

103. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, Nik-Zainal S, Martin S, Varela I, Bignell GR, Yates LR, Papaemmanuil E, Beare D, Butler A, Cheverton A, Gamble J, Hinton J, Jia M, Jayakumar A, Jones D, Latimer C, Lau KW, McLaren S, McBride DJ, Menzies A, Mudie L, Raine K, Rad R, Chapman MS, Teague J, Easton D, Langerød A, Lee MTM, Shen C-Y, Tee BTK, Huimin BW, Broeks A, Vargas AC, Turashvili G, Martens J, Fatima A, Miron P, Chin S-F, Thomas G, Boyault S, Mariani O, Lakhani SR, van de Vijver M, van 't Veer L, Foekens J, Desmedt C, Sotiriou C, Tutt A,

Caldas C, Reis-Filho JS, Aparicio S a JR, Salomon AV, Børresen-Dale A-L, Richardson AL, Campbell PJ, Futreal PA, Stratton MR: **The landscape of cancer genes and mutational processes in breast cancer.** *Nature* 2012, **486**:400–4.

104. Shrestha Y, Schafer EJ, Boehm JS, Thomas SR, He F, Du J, Wang S, Barretina J, Weir BA, Zhao JJ, Polyak K, Golub TR, Beroukhim R, Hahn WC: **PAK1 is a breast cancer oncogene that coordinately activates MAPK and MET signaling.** *Oncogene* 2012, **31**:3397–408.

105. Hernandez-Aya LF, Gonzalez-Angulo AM: **Targeting the phosphatidylinositol 3-kinase signaling pathway in breast cancer.** *The oncologist* 2011, **16**:404–14.

106. Toker A, Yoeli-Lerner M: **Akt signaling and cancer: surviving but not moving on.** *Cancer research* 2006, **66**:3963–6.

107. Rajagopalan H, Bardelli A, Lengauer C, Kinzler KW, Vogelstein B, Velculescu VE: **Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status.** *Nature* 2002, **418**:934.

108. Naoki K, Chen T-H, Richards WG, Sugarbaker DJ, Meyerson M: **Missense mutations of the BRAF gene in human lung adenocarcinoma.** *Cancer research* 2002, **62**:7001–3.

109. Pao W, Girard N: **New driver mutations in non-small-cell lung cancer.** *The lancet oncology* 2011, **12**:175–80.

110. Hodis E, Watson IR, Kryukov G V, Arold ST, Imielinski M, Theurillat J-P, Nickerson E, Auclair D, Li L, Place C, Dicara D, Ramos AH, Lawrence MS, Cibulskis K, Sivachenko A, Voet D, Saksena G, Stransky N, Onofrio RC, Winckler W, Ardlie K, Wagle N, Wargo J, Chong K, Morton DL, Stemke-Hale K, Chen G, Noble M, Meyerson M, Ladbury JE, Davies MA, Gershenwald JE, Wagner SN, Hoon DSB, Schadendorf D, Lander ES, Gabriel SB, Getz G, Garraway LA, Chin L: **A landscape of driver mutations in melanoma.** *Cell* 2012, **150**:251–63.

111. Wan PTC, Garnett MJ, Roe SM, Lee S, Niculescu-Duvaz D, Good VM, Jones CM, Marshall CJ, Springer CJ, Barford D, Marais R: **Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF.** *Cell* 2004, **116**:855–67.

112. Bast RC, Hennessy B, Mills GB: **The biology of ovarian cancer: new opportunities for translation.** *Nature reviews. Cancer* 2009, **9**:415–28.

113. Hynes NE, Lane HA: **ERBB receptors and cancer: the complexity of targeted inhibitors.** *Nature reviews. Cancer* 2005, **5**:341–54.

114. Sheng Q, Liu X, Fleming E, Yuan K, Piao H, Chen J, Moustafa Z, Thomas RK, Greulich H, Schinzel A, Zaghlul S, Batt D, Ettenberg S, Meyerson M, Schoeberl B, Kung AL, Hahn WC, Drapkin R, Livingston DM, Liu JF: **An activated ErbB3/NRG1 autocrine**

loop supports in vivo proliferation in ovarian cancer cells. *Cancer cell* 2010, **17**:298–310.

115. Rajkumar T, Stamp GW, Hughes CM, Gullick WJ: **c-erbB3 protein expression in ovarian cancer.** *Clinical molecular pathology* 1996, **49**:M199–202.

116. Sithanandam G, Anderson LM: **The ERBB3 receptor in cancer and cancer gene therapy.** *Cancer gene therapy* 2008, **15**:413–48.

117. Ward KK, Tancioni I, Lawson C, Miller NLG, Jean C, Chen XL, Uryu S, Kim J, Tarin D, Stupack DG, Plaxe SC, Schlaepfer DD: **Inhibition of focal adhesion kinase (FAK) activity prevents anchorage-independent ovarian carcinoma cell growth and tumor progression.** *Clinical & experimental metastasis* 2013, **30**:579–94.

118. Sood AK, Coffin JE, Schneider GB, Fletcher MS, DeYoung BR, Gruman LM, Gershenson DM, Schaller MD, Hendrix MJC: **Biological significance of focal adhesion kinase in ovarian cancer: role in migration and invasion.** *The American journal of pathology* 2004, **165**:1087–95.

119. Halder J, Lin YG, Merritt WM, Spannuth WA, Nick AM, Honda T, Kamat AA, Han LY, Kim TJ, Lu C, Tari AM, Bornmann W, Fernandez A, Lopez-Berestein G, Sood AK: **Therapeutic efficacy of a novel focal adhesion kinase inhibitor TAE226 in ovarian carcinoma.** *Cancer research* 2007, **67**:10976–83.

120. Yarden Y, Sliwkowski MX: **Untangling the ErbB signalling network.** *Nature reviews. Molecular cell biology* 2001, **2**:127–37.

121. Nicholson R., Gee JM., Harper M.: **EGFR and cancer prognosis**. *European Journal of Cancer* 2001, **37**:9–15.

122. Maihle NJ, Baron AT, Barrette BA, Boardman CH, Christensen TA, Cora EM, Faupel-Badger JM, Greenwood T, Juneja SC, Lafky JM, Lee H, Reiter JL, Podratz KC: **EGF/ErbB receptor family in ovarian cancer.** *Cancer treatment and research* 2002, **107**:247–58.

123. Gotlieb WH, Bruchim I, Gu J, Shi Y, Camirand A, Blouin M-J, Zhao Y, Pollak MN: **Insulin-like growth factor receptor I targeting in epithelial ovarian cancer.** *Gynecologic oncology* 2006, **100**:389–96.

124. Gest C, Mirshahi P, Li H, Pritchard L-L, Joimel U, Blot E, Chidiac J, Poletto B, Vannier J-P, Varin R, Mirshahi M, Cazin L, Pujade-Lauraine E, Soria J, Soria C: **Ovarian cancer: Stat3, RhoA and IGF-IR as therapeutic targets.** *Cancer letters* 2012, **317**:207–17.

125. King ER, Zu Z, Tsang YTM, Deavers MT, Malpica A, Mok SC, Gershenson DM, Wong K-K: **The insulin-like growth factor 1 pathway is a potential therapeutic**

target for low-grade serous ovarian carcinoma. *Gynecologic oncology* 2011, **123**:13–8.

126. Ludyga N, Anastasov N, Gonzalez-Vasconcellos I, Ram M, Höfler H, Aubele M: **Impact of protein tyrosine kinase 6 (PTK6) on human epidermal growth factor receptor (HER) signalling in breast cancer.** *Molecular bioSystems* 2011, **7**:1603–12.

127. Irie HY, Shrestha Y, Selfors LM, Frye F, Iida N, Wang Z, Zou L, Yao J, Lu Y, Epstein CB, Natesan S, Richardson AL, Polyak K, Mills GB, Hahn WC, Brugge JS: **PTK6 regulates IGF-1-induced anchorage-independent survival.** *PloS one* 2010, **5**:e11729.

128. Carpten JD, Faber AL, Horn C, Donoho GP, Briggs SL, Robbins CM, Hostetter G, Boguslawski S, Moses TY, Savage S, Uhlik M, Lin A, Du J, Qian Y-W, Zeckner DJ, Tucker-Kellogg G, Touchman J, Patel K, Mousses S, Bittner M, Schevitz R, Lai M-HT, Blanchard KL, Thomas JE: **A transforming mutation in the pleckstrin homology domain of AKT1 in cancer.** *Nature* 2007, **448**:439–44.

129. Le X-F, Mao W, He G, Claret F-X, Xia W, Ahmed AA, Hung M-C, Siddik ZH, Bast RC: **The role of p27(Kip1) in dasatinib-enhanced paclitaxel cytotoxicity in human ovarian cancer cells.** *Journal of the National Cancer Institute* 2011, **103**:1403–22.

130. Risinger JI, Berchuck A, Kohler MF, Boyd J: **Mutations of the E-cadherin gene in human gynecologic cancers.** *Nature genetics* 1994, **7**:98–102.

131. Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki R a: **The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists.** *Genome biology* 2007, **8**:R183.

132. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM: **Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM.** *Bioinformatics (Oxford, England)* 2010, **26**:i237–45.

133. **A distribution-free approach to rank correlation.pdf**.

134. Shih Y-K, Parthasarathy S: **Identifying functional modules in interaction networks through overlapping Markov clustering.** *Bioinformatics (Oxford, England)* 2012, **28**:i473–i479.

135. Chin C-H, Chen S-H, Ho C-W, Ko M-T, Lin C-Y: **A hub-attachment based method to detect functional modules from confidence-scored protein interactions and expression profiles.** *BMC bioinformatics* 2010, **11 Suppl 1**:S25.

136. Maraziotis I a, Dimitrakopoulou K, Bezerianos A: **An in silico method for detecting overlapping functional modules from composite biological networks.** *BMC systems biology* 2008, **2**:93.

137. Brohée S, van Helden J: **Evaluation of clustering algorithms for protein-protein interaction networks.** *BMC bioinformatics* 2006, **7**:488.

138. Dutkowski J, Kramer M, Surma MA, Balakrishnan R, Cherry JM, Krogan NJ, Ideker T: **A gene ontology inferred from molecular networks.** *Nature biotechnology* 2013, **31**:38–45.

139. King AD, Przulj N, Jurisica I: **Protein complex prediction via cost-based clustering.** *Bioinformatics (Oxford, England)* 2004, **20**:3013–20.

140. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic acids research* 2002, **30**:1575–84.

141. Bader GD, Hogue CW V: **An automated method for finding molecular complexes in large protein interaction networks.** *BMC bioinformatics* 2003, **4**:2.

142. Adamcsek B, Palla G, Farkas IJ, Derényi I, Vicsek T: **CFinder: locating cliques and overlapping modules in biological networks.** *Bioinformatics (Oxford, England)* 2006, **22**:1021–3.

143. Chin C-H, Chen S-H, Ho C-W, Ko M-T, Lin C-Y: **A hub-attachment based method to detect functional modules from confidence-scored protein interactions and expression profiles.** *BMC bioinformatics* 2010, **11 Suppl 1**:S25.

144. Palla G, Derényi I, Farkas I, Vicsek T: **Uncovering the overlapping community structure of complex networks in nature and society.** *Nature* 2005, **435**:814–8.

145. Kovács I a, Palotai R, Szalay MS, Csermely P: **Community landscapes: an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics.** *PloS one* 2010, **5**.

146. Nepusz T, Yu H, Paccanaro A: **Detecting overlapping protein complexes in protein-protein interaction networks.** *Nature methods* 2012, **9**:471–2.

147. Linghu B, Snitkin ES, Holloway DT, Gustafson AM, Xia Y, DeLisi C: **High-precision high-coverage functional inference from integrated data sources.** *BMC bioinformatics* 2008, **9**:119.

148. Qi Y, Bar-Joseph Z, Klein-Seetharaman J: **Evaluation of different biological data and computational classification methods for use in protein interaction prediction.** *Proteins* 2006, **63**:490–500.

149. Jiang T, Keating AE: **AVID: an integrative framework for discovering functional relationships among proteins.** *BMC bioinformatics* 2005, **6**:136.

150. KETCHEN Jr. DJ, SHOOK CL: **THE APPLICATION OF CLUSTER ANALYSIS IN STRATEGIC MANAGEMENT RESEARCH: AN ANALYSIS AND CRITIQUE**. *Strategic Management Journal* 1996, **17**:441–458.

151. Glaab E, Baudot A, Krasnogor N, Valencia A: **Extending pathways and processes using molecular interaction networks to analyse cancer genome data.** *BMC bioinformatics* 2010, **11**:597.

152. Berg J, Tymoczko J, Stryer L: **the Pentose Phosphate Pathway Generates NADPH and Synthesizes Five-Carbon Sugars**. In *Biochemistry. 5th edition*. 2002.

153. Bogorad IW, Lin T-S, Liao JC: **Synthetic non-oxidative glycolysis enables complete carbon conservation.** *Nature* 2013.

154. **HumanCyc** [http://humancyc.org/].

155. **Pathway Interaction Database** [pid.nci.nih.gov].

156. **BioCarta** [www.biocarta.com].

157. **Pathway Commons**.

158. **Reactome**.

159. Vivanco I, Sawyers CL: **The phosphatidylinositol 3-Kinase AKT pathway in human cancer.** *Nature reviews. Cancer* 2002, **2**:489–501.

160. Song MS, Salmena L, Pandolfi PP: **The functions and regulation of the PTEN tumour suppressor.** *Nature reviews. Molecular cell biology* 2012, **13**:283–96.

161. Strömberg T, Dimberg A, Hammarberg A, Carlson K, Osterborg A, Nilsson K, Jernberg-Wiklund H: **Rapamycin sensitizes multiple myeloma cells to apoptosis induced by dexamethasone.** *Blood* 2004, **103**:3138–47.

162. Castedo M, Ferri KF, Kroemer G: **Mammalian Target of Rapamycin (mTOR): Pro- and Anti-Apoptotic**. *Cell Death and Differentiation* 2002, **9**:99–100.

163. Ballou LM, Lin RZ: **Rapamycin and mTOR kinase inhibitors.** *Journal of chemical biology* 2008, **1**:27–36.

164. Faivre S, Kroemer G, Raymond E: **Current development of mTOR inhibitors as anticancer agents.** *Nature reviews. Drug discovery* 2006, **5**:671–88.

165. Harada H, Andersen JS, Mann M, Terada N, Korsmeyer SJ: **p70S6 kinase signals cell survival as well as growth, inactivating the pro-apoptotic molecule BAD.**

Proceedings of the National Academy of Sciences of the United States of America 2001, **98**:9666–70.

Curriculum Vitae

[redacted]

[redacted]

[redacted]