UNIVERSITY OF CALIFORNIA
SANTA CRUZ

SAMPLE-SPECIFIC CANCER PATHWAY PREDICTION FROM GENOMIC, TRANSCRIPTOMIC AND PHOSPHOPROTEOMIC DATA

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

DOCTOR OF PHILOSOPHY

in

BIOMOLECULAR ENGINEERING AND BIOINFORMATICS

by

Evan O. Paull

March 2016

The Dissertation of Evan O. Paull
is approved:

_____
Professor Joshua M. Stuart, Chair

_____
Distinguished Professor David Haussler

_____
Professor Todd Lowe

_____
Dean Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Sample-specific cancer pathway prediction from genomic, transcriptomic and phosphoproteomic data

by

Evan O. Paull

Cancer phenotypes, such as invasion, evasion of programmed cell death, and rapid growth, arise from the complex interactions of genes, proteins and extra-cellular environments. Understanding how selective alterations in the genome convert healthy cells to cancer is a critical step in developing new targeted and combination therapies. Current technology allows for detailed measurement of both genomic state as well as phenotype, through measurement of gene expression, chromatin state and protein activation. I present a method to find pathways linking key genomic alterations to phenotypic effects, using high-throughput data collected from cohorts of cancer patients, and then apply this method to predict sample-specific network models in metastatic prostate cancer. The method presented here, Tied Diffusion through Interacting Events (TieDIE), uses a "heat diffusion" model of information transfer to simultaneously combine multiple types of biological data with prior knowledge to predict network models of disease. Applying this method to four large data sets developed by The Cancer Genome Atlas (TCGA), I found key genes and interactions linking mutations related to histone modification and protein kinase signaling to gene-expression signatures of growth and proliferation. I next applied TieDIE to a study of metastatic, lethal prostate cancer

that also included detailed measurements of the phosphpoproteome. Sample-specific network predictions were developed through an extension of the original algorithm, providing a hierarchy that reveals the top kinase targets for each patient analyzed, and the corresponding therapeutic intervention.

Acknowledgments

I would like to thank my adviser, Joshua Stuart, for seeding and helping develop nearly all of the ideas presented here, and for his motivating guidance during my graduate work. I would also like to thank my committee chairman, David Haussler, for his generous guidance and inspiration throughout my graduate career, and his contribution to my graduate education. I thank my committee member, Todd Lowe, for his generous guidance, encouragement and contribution to my graduate education. I would also like to thank my colleague and fellow graduate student, Daniel E. Carlin, for his collaboration and contribution to the formation of ideas used in my graduate work.

Introduction

Cancer has a major health impact on society: in the United States alone, an estimated 1.6 million new cases of cancer will be diagnosed and 600,000 will die from the disease in 2015, according to the National Institutes of Health (www.cancer.gov). Although a third of cancers are preventable, according to current best estimates (www.who.int), evidence published in numerous scientific journals as well as independent research studies suggest that a majority of cases can be attributed to heritable genetic variation and random effects associated with the normal process of cell division [206].

Aside from inherited genetic factors, most cancers are thought to result from acquired genetic and epigenetic changes (the somatic mutation theory of cancer [69]) identified by at least eight biological capabilities developed through the multistep development of tumors [85, 219]. These eight biological "hallmarks" currently include sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, reprogramming of energy metabolism, and evading immune destruction [85].

1.1 Detecting Impactful Genomic Alterations

While evidence for each of these biological programs can be detected by measurement of transcriptional activity, epigenetic state, and protein activity (See appendix sections A.1.4, A.1.5, A.1.6), relatively little is understood as to how genomic abberations affect this cancer phenotype. Current large-scale analyses of genomic "drivers" of tumor phenotype focus on statistical techniques to implicate point mutations by recurrence and sequence features [6, 59, 119] or more complex patterns of recurrence, such as mutual exclusion [44]. Similarly, analyses of copy-number data has been able to identify regions of recurrent copy-number amplification or deletion, using many of the same principles [139, 225]. Linking these putative genomic drivers to the phenotypic changes cataloged in the cancer "hallmarks" is more difficult, and often involves interrogation of multiple types of biological data.

In some cases, genomic analyses have revealed somatic mutations that predict constitutive activation of signaling circuits, often triggered by activated growth factor receptors [85]. In these cases, tumor cells are often able to exploit unmodified components of the cellular signaling machinery such as the mitogen-activated protein kinase (MAPK) [56] and AKT [104] pahtways. This observation has led to the study of targeted inhibitors of these canonical signaling pathways for use in cancer therapy [51,56,85] (See appendix A.1.2).

1.2 Pathway Methods

After first identifying candidate genomic drivers, a further step in finding clinically actionable drug targets is to use genomic, transcriptomic , and other high-throughput biological measurements to gain a better understanding of the molecular mechanisms driving each cancer. The goal of this research strategy is to find the most vulnerable points of each tumor's biology and disable it with targeted drug therapies. Biological pathways—representations of the known interactions between proteins, RNA and genes—can support this goal by adding *a priori* biological information that provides a context to integrate otherwise disparate types of high-throughput data (See appendix A.1.1.). In this way, pathways allow for information transfer between high-throughput datasets that capture very different aspects of cellular biology (e.g., RNA-Seq for transcriptional state, Mass Spectrometry for protein phosphorylation state), allowing one to construct a more complete model of cell state, one that is greater than the sum of its parts.

To illustrate this concept, for example, possessing data on copy number alterations, gene mutations, and gene-expression, does not clarify which of these genomic changes is responsible for the cancer phenotype, approximated by microarray gene expression. A pathway-based methodology, in contrast, would attempt to answer this question by finding a set of literature-supported biological interactions that connects the genes and protein identified by each dataset. This pathway, in turn, would provide a logical explanation of how the perturbations relate to the observed expression data,

and reveal proteins that might disrupt the pathway when targeted by small molecule inhibitors.

A small example of such a pathway is shown in figure 1.1. In this example, relationships between the epidermal growth factor receptor (EGFR), protein components of the PI3K pathway, and the tumor-suppressor PTEN are shown alongside genomic and expression data in glioblastoma (GBM). Mutual exclusivity between loss and gain of function events is observed between these genes, indicating equivalence of alteration between genes in this pathway. In this case, the pathway information includes the knowledge that PIK3CA and PIK3R1 are components in the same complex, explaining the equivalence between loss of function (truncating or missense mutations) in either of these protein sub components. Similarly, the pathway interactions between EGFR, PI3K components, and PTEN explain the equivalence between gain of function alterations in EGFR and loss of function alterations in PTEN, providing a context to interpret multiple types of genomic and gene expression data.

Several pathway models are available for combining experimental data with prior pathway knowledge. The Signaling Pathway Impact Analysis (SPIA) algorithm [201] was developed to discover specific pathway perturbations in colorectal cancer, using a model that propagates information over a gene interaction network while accounting for edge direction and type (see methods 2.2.2.3).

To identify transcriptional regulatory interactions and transcriptional "master regulators" that regulate the expression of a large number of genes, the Master Regulator Analysis (MRA) algorithm [124] uses a modified form of Gene Set Enrichment Analysis

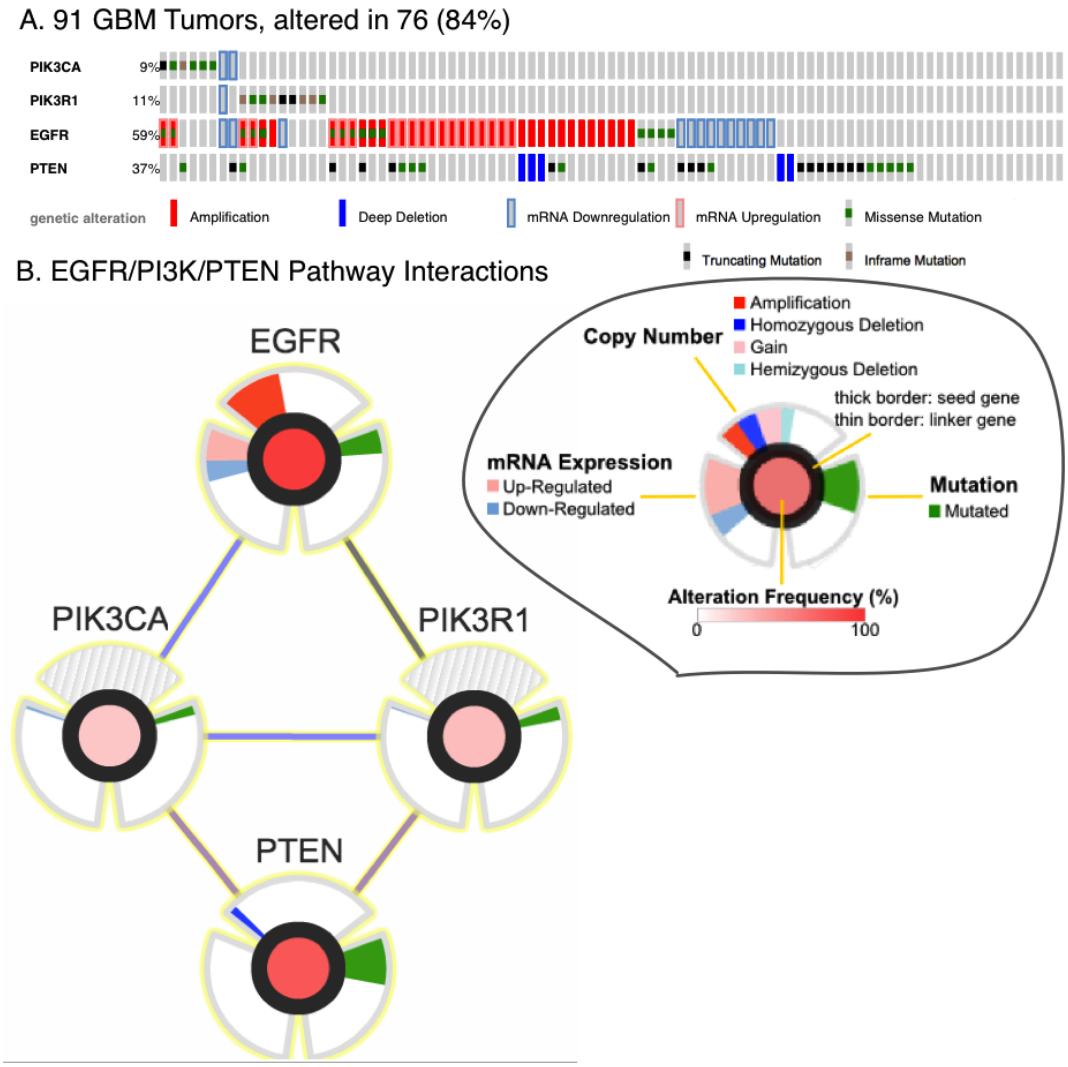

Figure 1.1: A) Oncoprint (www.cbioportal.org) summary of genomic and transcriptional alterations in GBM tumor samples: each sample is a column, and each row corresponds to one of 4 genes. Frequent amplification and mRNA over-expression of receptor tyrosine kinase Epidermal Growth Factor Receptor (EGFR) displays a tendancy towards mutual-exclusivity with activating mutations in PI3K pathway genes and loss-of-function mutation in the PTEN tumor-supressor gene [30]. B) Pathway interactions between frequently altered genes (reacts with: EGFR-PIK3CA,PIK3R1, PIK3CA-PIK3R1; same component: PTEN-PIK3CA,PIK3R1; other: EGFR-PIK3R1) illustrate different ways the cancer can alter key mechanisms of tumorogenesis.

(GSEA) [198], along with a set of transcription factor to target interactions inferred from the data [137]. A sample-specific version of this algorithm is provided as part of the VIPER software package [19] (see appendix A.3).

Another method, PARADIGM [217], uses a Bayesian factor graph model to perform inference across a gene interaction network and multiple data types, computing gene-specific inferences of predicted activity (see appendix A.4).

In addition to these inference algorithms, several methods can be applied to search for sub networks (a collection of pathway interactions that interconnect a set of genes of interest) that explain some aspect of a given data set. The jActive modules plug-in to cytoscape is one of the original algorithms developed to search for sub networks that connect a set of genes across a given network [102]. Following that, the problem of sub network selection was posed as a prize-collecting Steiner tree (PCST) linear optimization problem and used to find modules in a protein-protein interaction network, from microarray data of lymphoma patients [61]. This is one of several optimization-based methods that find connected sub networks around a given input set of genes, with a minimal number of connecting nodes (See appendix A.2.1).

A novel algorithm, HotNet [213], finds sub networks by searching for areas of influence for a given set of genes on a network. This method has been used in many large-scale studies to find mutated sub networks; in a large-scale integrative study of clear cell renal cell carcinoma by The Cancer Genome Atlas (TCGA), HotNet identified a key sub network containing von Hippel-Lindau tumor suppressor (VHL), a key factor in the development of kidney cancer, as well as a network involved in chromatin remodeling

[150] (see methods 2.2.2.1).

This approach has also been used to predict genetic interactions and co-complex membership in yeast [172] using a similar graph-diffusion methodology. In addition, new methods such as "Network Based Stratification (NBS)" [95] extend the heat-diffusion concept by transforming mutational data through a "network smoothing" process, which is then used to cluster and stratify patient samples. This method was applied to a study of thyroid carcinoma, finding three distinct subtypes that were both correlated with histological subtype, risk of recurrence, cancer stage, and multiple molecular characteristics [152].

1.3 Connecting Genomic Changes to Transcriptional State of Cancer Cells

Finding driver genomic events is a key step in identifying clinically actionable targets; a more specific goal, however, is to find the causal relationships between specific alterations to the genome and cancer phenotype. The importance of this question can be seen in the biology of phosphoproteomic signaling and how it processes extra cellular signals to patterns of growth, differentiation, apoptosis and migration in the nucleus of the cell. For example, in figure 1.2, an overview of the (MAP)-kinase signaling pathway diagrams some of the signaling pathways that relay extra cellular signals to the transcriptional machinery in the nucleus, resulting in phenotypic changes. These pathways are often essential, highly conserved units of biology that allow for rapid changes in

cellular state [60]; genomic alterations in these pathways, however, are often responsible for giving cancer cells the capacity to sustain proliferative signaling, often in the absence of extra cellular growth factors [85]. For instance, roughly 40% of human melanomas contain activating mutations that change the structure of the B-Raf protein, resulting in constitutive signaling through the RAF and MAPK pathways [56, 85]. Algorithms that can link these genomic aberations to their respective phenotypic consequences are needed not only to find the causal drivers of specific tumor biology, but also to identify the unmodified signaling components that relay these aberant signals and therefore represent unique points of vulerability.

The PCST framework is one of the first methods [117,189] developed to extract sub networks that connect genomic perturbation data and expression data through a small set of intermediate nodes. However, this approach does not take into account the type of interactions present and may produce network solutions that are not logically consistent with the data. For example, a protein that suppresses transcription of a gene paradoxically might be used to explain higher transcription of that gene. In addition, by using a linear optimization approach to minimize the number of edges in the biological network, one might miss key functional parts of a biological network that might be only a single step further than the "optimal" solution (See appendix A.2.1.). Flow algorithms attempt to avoid this situation by allowing for weights to be assigned to edges, but in reality it is impractical to compute relative probabilities over edges that are based on multiple types of biological assays [110], and the highly connected topology of graphs leads to frequent ties in edge length [24]. In spite of those limitations, these approaches

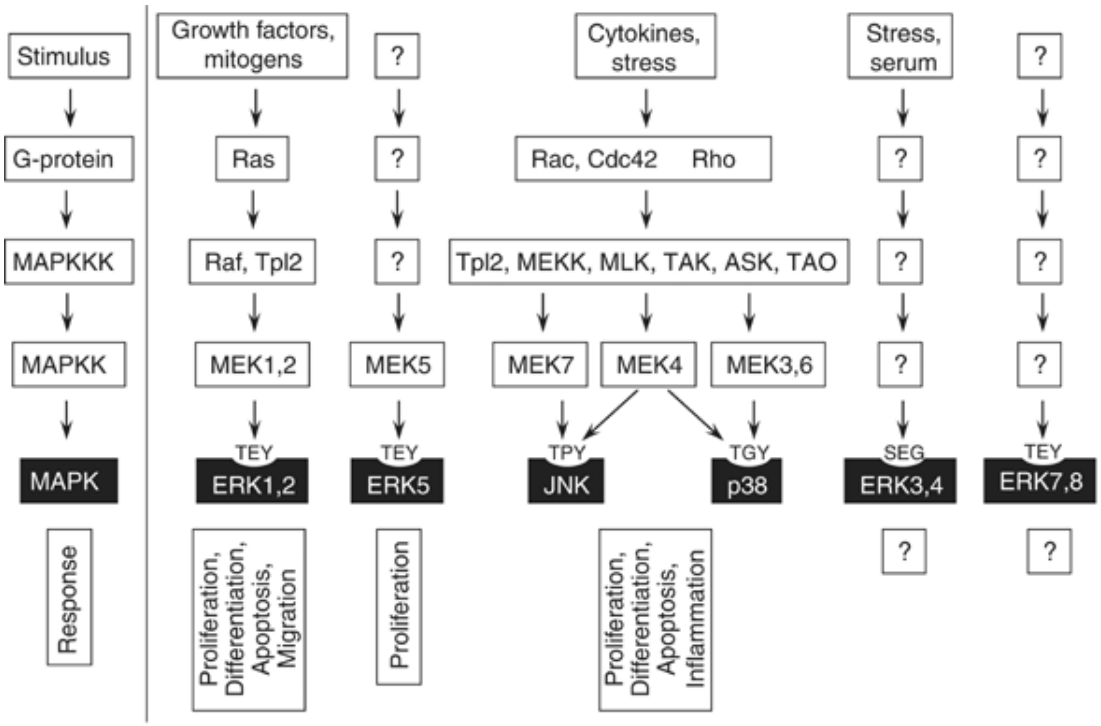

Figure 1.2: Schematic overview of MAPK pathways [60]. extra cellular signals are linked through a three-tier kinase pathway where MAPK is activated upon phosphorylation of mitogen activated protein kinase kinase (MAPKK), which is in turn activated by MAPKKK citedhillon2007map. The cellular processes listed below are essential "hallmarks" of tumorogenesis [85] and their state can be assessed by measurement of gene expression, using technologies such as RNA-Seq A.1.4.

have been successful in linking oncogene signaling to transcriptional activity through the proteome [117].

In this thesis, I propose a different approach to connect genomic perturbations to transcriptional changes over a pathway space, producing sub network solutions that are far more complete than linear optimization methods, while also making full use of the interaction logic provided in pathway databases. This method—Tied Diffusion Through Interacting Events (TieDIE)—uses a heat diffusion model of information flow,

finding sub networks of protein-protein interactions, predicted transcription factor to target connections, and curated interactions from literature. It then find paths connecting source to target nodes that are logically consistent with both the perturbation or expression status of each gene pair and the type of interaction between each edge in the path. Finally, permutation-based analysis is used to gauge the significance of the solutions resulting from the TieDIE network.

As I show with a rigorous computational evaluation on synthetic data from human breast tumors, sub network solutions produced by TieDIE are both accurate and far more comprehensive than previous methods because of two deliberate choices: to separate the search and pathway integration (the heat-diffusion step) from the modeling of biology (the search for logically consistent paths and to treat genomic and transcriptional data as distinct entities, effectively adding information to the algorithm.

The clinical relevance of TieDIE is first validated on a panel of breast cancer cell lines, using RNA-interference (RNAi; See appendix A.1.7.) knock-down data to measure the quality of network models produced by the algorithm. The algorithm is then demonstrated on four large-scale sequencing projects coordinated by The Cancer Genome Atlas (TCGA), in each case finding tissue-specific network models that connect cancer-driving genomic events with the respective gene-expression signatures of carcinogenesis, tumor progression, and maintenance. The identified networks and corresponding "linking" genes and interactions that are implicated by the surrounding pathway context rather than by direct genomic evidence, are validated through extensive literature review and outside data in coordination with TCGA project members,

and are published as part of three separate TCGA project papers. In each case, the TieDIE algorithm highlighted genes, interactions, and cellular processes that might be useful for follow-up studies, in particular to assess the clinical utility of inhibiting these respective targets.

I continue to explore tissue-specific network and gene prediction next, with an outline of results related to my participation in the Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenges 8 and 9. These results provide further support for the use of prior pathway knowledge in predicting important genes and interactions that are specific to a given tissue or genomic context. Finally, I combine all the previous concepts detailed here and apply them to a study of lethal, metastatic prostate cancer for which a dataset measuring phosphophoteomic signaling has been created A.1.6. Patient-specific network models are developed using an extension of the TieDIE algorithm for samples with phosphoproteomic, gene expression, and genomic data, and a methodology for finding clinically actionable protein targets in these patients is developed, along with a new visualization paradigm to provide deeper insight into the disease.

TieDIE: An Algorithm for Sub Network Search and Discovery

2.1 Introduction

Integrating multiple data sources with a pathway model can greatly improve the ability to find active sub networks in disease, compared with using just a single source of information such as gene expression or mutational data. Several methods for identifying sub-network models are available to combine genomic, gene expression data, or both with pathway databases, including the HotNet, SPIA, and PCST algorithms discussed in the previous chapter.

Here, I introduce the TieDIE algorithm, which provides a way to combine an arbitrary number of data inputs with pathway knowledge simultaneously, applying a heat diffusion model to find "linker genes" that are strongly implicated by each data input. In this chapter, linker genes are then used to find sub networks that connect genomic perturbations to transcriptional changes, while later chapters demonstrate the method on more than two data sets. These sub networks are critical in finding genes that

may be lacking in cis-level data but are implicated by other genes in the surrounding pathway and the logic of the corresponding interactions. TieDIE also is a critical step in the computational pipeline developed to predict sample-specific, $N = 1$ networks in the final chapters, providing a framework to relate patient-level data to the statistically robust sub networks generated with cohort-wide data.

I benchmarked TieDIE against multiple competing methods on both synthetic datasets and those derived from patient tissues, and found that it greatly outperforms previous methods by several metrics, on these data. All the methodological description and results presented here are also available in the TieDIE *Bioinformatics* publication [165] and supplemental text.

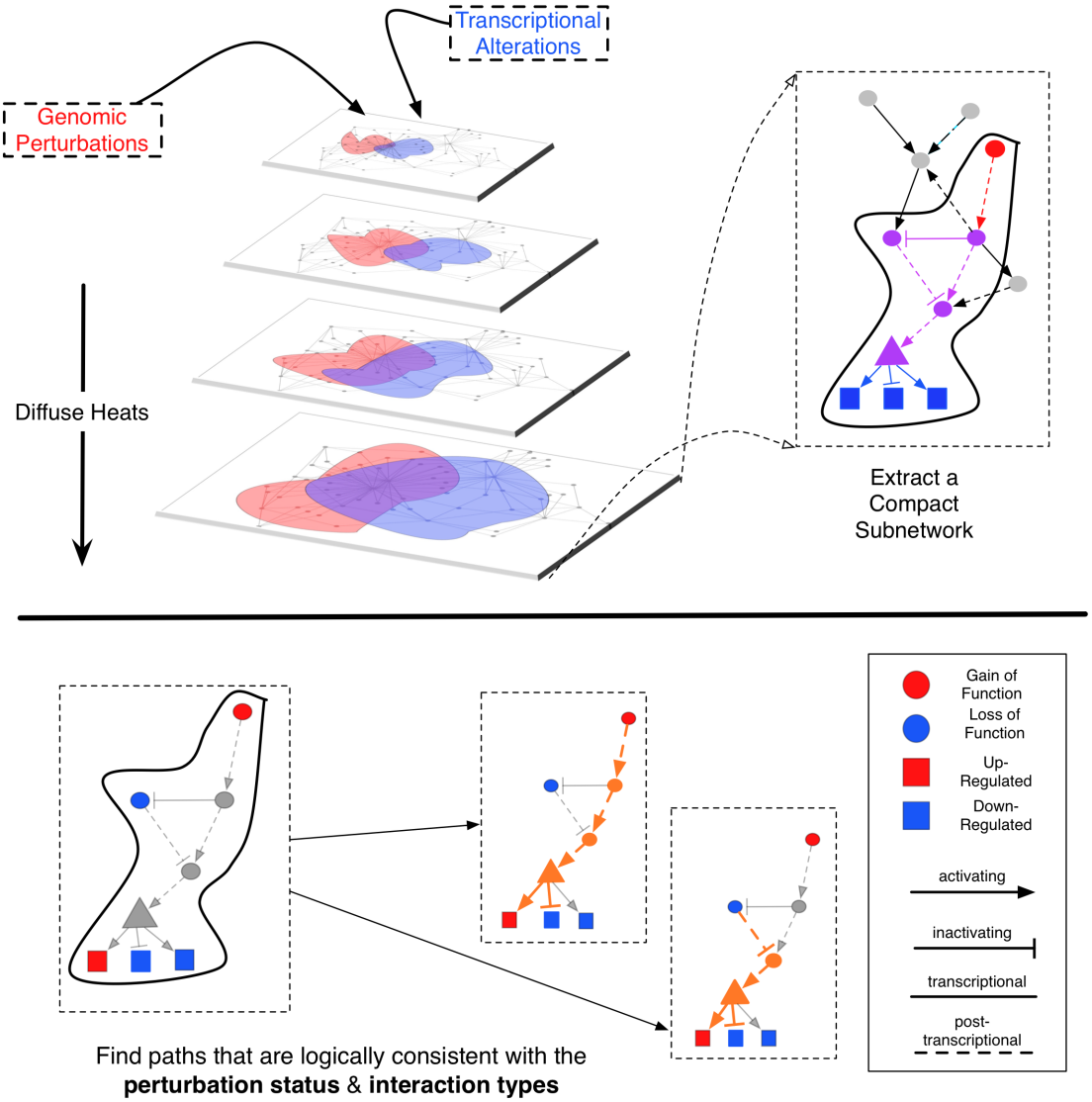

Figure 2.3: TieDIE: heats are diffused from source and target sets. "Linker" genes that include heat from both source and target nodes are drawn in purple as the intersection between diffused sets. TieDIE attempts to maximize this intersection while maintaining specificity of the diffused source and target sets (see methods 2.2.5). (Below) Linker genes are used to find a compact sub network solution that connects source to target nodes. I then find paths connecting source to target nodes that are logically consistent with both the perturbation or expression status of each gene pair, and the type of interaction between each edge in the path.

2.2 The TieDIE Algorithm

Building on a "heat kernel" model developed for physics, the TieDIE method [165] takes the output of a differential gene expression analysis genered from transcriptional (RNA-Seq, microarray) data and performs heat diffusion over a supplied biological pathway to find a pathway neighborhood strongly supported by the transcriptional data. Separately, the algorithm takes a set of significantly perturbed genes found with genomic sequencing data (mutations, copy number alterations), performs heat diffusion on these inputs and then merges the result with the diffused transcriptional data via a "linker" function, producing an overall score for each gene in a given input pathway. The algorithm uses high-scoring "linker genes" to find sub networks connecting genomic alterations ("source" genes) to the observed transcriptional changes ("target" genes) in a patient cohort, providing a probable explanation for dysregulated transcriptional profiles (see Figure 2.2.2.4).

2.2.1 Problem Statement

The TieDIE approach searches for relevant interconnecting genes on a background network using a diffusion strategy. The method is given as input an interaction network graph G containing N vertices $V = v_1..v_N$ that represent genes, proteins or other biological pathway features such as gene products, protein complexes and cellular abstract processes. The nodes in G are interlinked by I edges $E = e_1..e_I$ representing both directed interactions as well as undirected relations such as proteinprotein interac-

tions. The interactions can be derived from curated sources such as the National Cancer Institute's Pathway Interaction Database [185], from functional genomics predictions, such as undirected high-throughput proteinprotein assays, or directed TF to target interactions such as from genome-wide chromatin-immunoprecipitation experiments, or from a mixture of both sources such as Reactomes Functional Interaction Network [105]. TieDIE makes use of the adjacency matrix A of the graph G, where $A_{ij} = 1$ if node i activates node j, $A_{ij} = -1$ if node i represses or inactives node j and 0 otherwise.

In addition to the graph, the method is given a set of scores for each node in G. Let $x = [x1, x2, , xN]$ be a vector of scores assigned to the nodes in the graph. Typically, only a limited set of genes (e.g. in the range 1050) will have known involvement in the disease process being studied. For example, this involved set may consist of genes for which a minimum number of mutations or copy number changes or DNA methylation silencing events have been detected in a given cohort of patient samples. Nodes corresponding to these involved genes are assigned scores between 1 and +1 to reflect a positive or negative association of activity with the disease state under study. Nodes associated with genes not known to be involved in the disease process are assigned a value of 0 to reflect an *a priori* belief in no involvement.

The values in x can represent different types of measurements on the genes. For instance, the scores might reflect how often a gene is mutated in one subtype of patients compared with another (e.g. from a *log* of the P-value computed from a Fishers exact test to detect differential mutation frequency). Alternatively, the scores can reflect a genes differential expression in tumor versus normal. Statistical techniques, such as

significance analysis of microarrays (SAM) [41], or edgeR for RNA-Seq data [180] may be used to compute the significantly differentially expressed genes and the SAM score (d-statistic) for each gene normalized to this range. Transcription factors this may represent the score from a Gene Set Enrichment Analysis (GSEA) [198] test on gene expression data, taking d-statistic scores as input. For genomics data, mutation or copy-number events may be prefiltered with algorithms such as MutSig [40] and MEMo [45] that use sample statistics to find events that are likely to be 'driving the cancer phenotype. Note here that all mutations are assumed to lower a genes activity (even though a minority, oncogenic mutations in particular, increases gene activity) and to take precedence over copy number alterations (e.g. amplifications) and expression events (e.g. overexpression), which will inflate the false-negative rate of the algorithms tested here as some true paths will not be counted. Relaxing this assumption is an important topic discussed later in this document. The input vector of positive scores is scaled to match an intuition that the scores reflect a stationary probability distribution of occupancy on the nodes in the network obtained from a random walk process: $\sum_{i-1}^{N} x_i = 1$.

2.2.2 Relevance Functions

To simplify the comparison between TieDIE and preceding methods, I abstract the idea of diffusion on a biological network and describe a *relevance function* as an update of the state of all of the nodes in G by some function $x = r(x, A)$, that is a function of the input set of scores x and of the full adjacency matrix A.

The TieDIE method and methods that are compared to it here all make use of a derivative of the graphs adjacency matrix. Let the full signed adjacency matrix be A where $A_{ij} = 1$ if node i activates node j, $A_{ij} = -1$ if node i represses or inactivates node j, and 0 otherwise. Undirected edges, such as protein-protein interactions, are encoded as $A_{ij} = A_{ji} = 1$ in this work. Denote the directed, but unsigned version of the adjacency matrix as D; the entries are then set to $D_{ij} = |Dij|$, i.e. $D_{ij} = 1$ if there is a directed edge from gene i to gene j and 0 otherwise. Such an unsigned representation of A retains the directionality but throws away information about repression versus activation in regulatory interactions. Some approaches make use only of the undirected structure of G without considering how information flows upstream or downstream. Let the undirected version of the adjacency matrix be U, a symmetric matrix containing entries $U_{ij} = U_{ji} = (D_{ij}|D_{ji})$, where "|" denotes the logical OR operation [165].

2.2.2.1 HotNet

The HotNet method uses a heat diffusion process to derive an update for all of the nodes in G, but does not consider the directionality of an edge in G, making use of the U version of the adjacency matrix. Let B be a diagonal matrix with B_{ii} equal to the number of other genes that gene i interacts with, and $B_{ij} = 0$ for all ij. The heat diffusion-based update for HotNet is then:

$$r_{HotNet}(x, A) = x * e^{-(B-U)t}$$

where BU is known as the Laplacian matrix L of G, its exponentiated form

is known as the diffusion kernel, and t is an arbitrary time step. I fixed t=0.1, which Vandin *et al.* ([213] and Qi *et al.* [172]) found to be most useful from a number of simulation tests. The dynamics of this continuous-time process are governed by the vector equation $\frac{d\vec{f}(t)}{dt} = L * f(t)$. The solution to this system of equations is $\vec{f}(t) = \vec{f}(0) * e^{-Lt}$, with e representing the matrix exponential. The matrix exponential is estimated using the implementation of the Pad approximation in Matlab (Golub and Van Loan, 1989). Once the diffusion kernel is available, the relevance scores x can be computed in one linear operation. Thus, an advantage of this approach is that the kernel can be pre-computed and then applied efficiently to as many input sets as desired.

2.2.2.2 Personalized PageRank

Personalized PageRank is a variation of the PageRank algorithm, originally developed for Google search, that allows for bias of a particular topic to be introduced when calculating PageRank vectors [86]. The Personalized PageRank recurrence can be approximated using the following iterative computation in which an initial starting value x_0 is set to the input scores x and an updated version of relevance scores x_t is calculated from the previously calculated relevance scores x_{t-1} [43]:

$$x^t = \delta * x^0 + (1 - \delta) * x^{t-1} * (B^{-1}D),$$

where δ is the probability that a completely new node is chosen irrespective of the current node (teleport step), $(1 - \delta)$ is the probability a neighbor in G is chosen at time t (random walk step), and $B^{-1}D$ is the random walk matrix and uses the

directed, but unsigned connections in G [43]. The recurrence is repeated using the power method [86] until time T such that $(x^t - x^t - 1) < \epsilon$, where ϵ here was set to 0.001 to represent a 0.1% change between consecutive iterates. The final relevance scores x then are then set to x^T.

2.2.2.3 Signaling Pathway Impact Analysis (SPIA)

I also compare TieDIE to SPIA, a method that incorporates activation and inactivation of the regulatory links encoded in G and computes the quantity:

$$r_{SPIA}(x, A) = W * (I - W)^{-1} * x,$$

where W represents a normalized version of the full directed, signed adjacency matrix: $W_{ij} = A_{ij}/Nds(j)$, where $Nds(j)$ is the number of downstream neighbors of node j encoded by G.

All the methods discussed above can take advantage of a simple transformation to infer the activities of transcription factors (TFs) with expression data (see appendix A.3).

2.2.2.4 TieDIE

In contrast to previous approaches that implicate sub networks by running the relevance function with a single input set of genes, our algorithm extends the strategy by using multiple diffusion processes and then identifying overlapping regions in G, to find genes in a network that are proximal to multiple input sets. I develop the approach

here for the special case of two input sets but the method generalizes to any arbitrary number of input sets.

Suppose we are given a source set S and target set T where S acts upstream of T. In the cancer setting, S may correspond to genes involved in genomic alterations-mutations, deletions and amplificationswhereas the target set may correspond to genes involved in transcriptional and post-transcriptional activation or deactivation. However, any features in the pathway diagram in addition to proteins can be included in the input sets including protein complexes, small molecules or metabolites and cellular processes. In the same spirit as the spanning tree approach (Huang and Fraenkel, 2009), we are interested in identifying parsimonious networks that connect S to T. I now let x represent the vector of scores for the source set and y the scores for the target set [165].

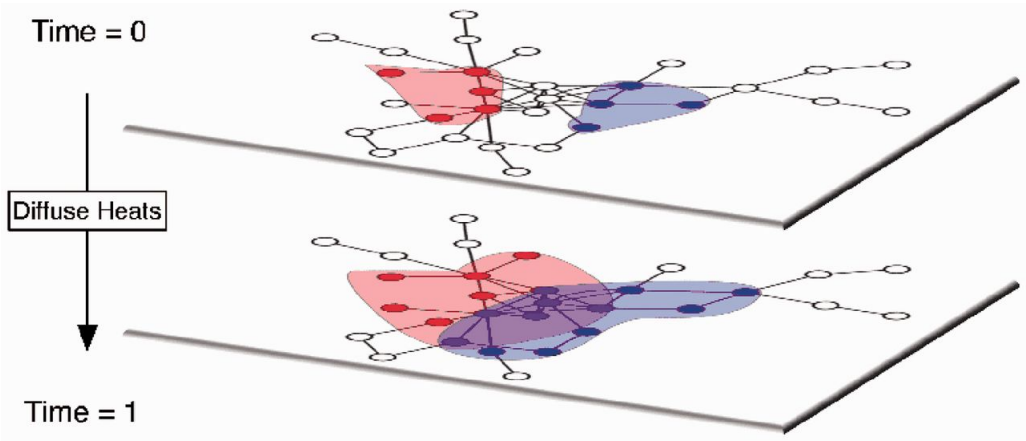

Figure 2.4: Visualization of the TieDIE Diffusion process. Relevant genes from two separate sets are shown as nodes colored by dyes diffusing on a network (e.g. mutated genes; red nodes) and target set (e.g. TFs; blue nodes). Linker genes (purple nodes) residing between the source and target sets are revealed through a diffusion process evolved over time; two time slices are shown as stacked layers of the same network [165]

We can visualize the process for two input sets by imagining the intermixing of

different color dyes, diffusing from two sources through a lattice representing G (figure 2.4. The intensity and hue of the dye reveals whether a particular node is close to either the source, target or to both sets. To disambiguate these cases and identify points that reside only between the sets, we determine a score for all nodes based on the relevance scores already computed [165]:

$$z = f(r(x, A), r(y, A^T)),$$

where A^T is the transpose of the full adjacency matrix, the function $f()$ is chosen to assign high relevance scores to nodes where both $r(x, A)$ and $r(y, A^T)$ are high and lower scores when either of the two are low. Note that the transpose of the adjacency matrix is used to force the diffusion to proceed upward from the targets by supplying a graph containing reversed edges. When applied to directional diffusion approaches like PageRank and SPIA, this has the effect of running the algorithm backward. Of course the transpose makes no difference for undirected approaches like HotNets heat diffusion, as $r_{HotNet}(y, A) = r_{HotNet}(y, A^T)$. I refer to z_i as the linking score for node i.

This form decomposes over the separate relevance calculations, which can be "plugged in" to the tied-diffusion calculation. In this document, I chose $f()$ equal to the $min()$ operator to extract genes that have intersecting evidence from each input datasets.

A set of linking genes is obtained by thresholding the linking scores using a chosen value α selected to guarantee a desired level of specificity as a fixed multiple of

the input set size. Specificially, the linker set generated at this threshold L_α is set so that the unique linker nodes (not in the source or target sets) are of size s:

$$\|L_\alpha \neg\{S \cup T\}\| \; s * \|S \cup T\|$$

This "network size" parameter provides a single user-tunable parameter to control the size of the resulting sub network, and, as discussed below, the null model developed to test the significance of the resulting sub networks is specific to this size control parameter. By default the size control factor is set to 1, on the assumption that we expect to find a number of novel genes comparable to the number of genes that are already known.

A set of connected sub networks C is obtained by taking the union of all nodes in the S, T, and linker sets L, and adding an edge between any two nodes in this subset that contain an edge in G: $C = \{(u, v) \; \forall u, v \in \{S \cup T \cup L_{alpha}\}\}$.

An alternative method, uses the BioNet heuristic to approximately solve the Prize Collecting Steiner Tree (PCST) problem ([61]) that includes all linker genes and input gene sets. This approach has the advantage of being able to connect smaller connected sub components into a single connected graph, when the number of required intermediate links is small. A fast approximation algorithm that solves the PCST algorithm is available as part of the BioNet package [27, 165].

2.2.3 Extracting Logically Coherent Paths

The network solutions that interconnect sources and targets can be large, with hundreds of genes, even though we attempt to control the specificity and size of the

output as described above. In addition, some diffusion approaches like HotNet ignore the logical consistency of the interconnections between the set of genes identified to be relevant. While it may be advantageous to maintain contradictory influences (e.g. to highlight discrepancies between pathway interactions from different databases), it is generally difficult to extract meaningful information from large networks. Therefore, I introduced a filter to specifically select for consistent regions of the network, which focuses attention on the best defined and most interpretable regions of the solution space.

To do this, I first assign an activity score $z(i)$ between -1 and $+1$ to all genes i involved in genomic, epigenomic, or transcriptional events. $z = -1$ is assigned to events such as mutations, methylated promoters, or focal deletions. Note this encoding makes the assumption that mutations are deleterious, which can be wrong because many mutations are known to activate important oncogenes, but in the TieDIE software this can be bypassed by a user with expert knowledge of mutation effects (or the help of a mutation impact assesment algorithm). $AS = +1$ are assigned to focal copy number gains, and finally, genes involved in only transcriptional events are assigned their expression or differential expression levels normalized to the range [-1,+1] [165].

Figure 2.5 illustrates the intuition behind the idea of extracting logical paths. To find logically coherent paths, I search for all possible paths of length k, or smaller between any single source and any single target. I use $k = 3$ in this work for mathematical tractability and simplification of visualization, to test various approaches. I then use the activation scores to compute a measure of consistency on each path: first,

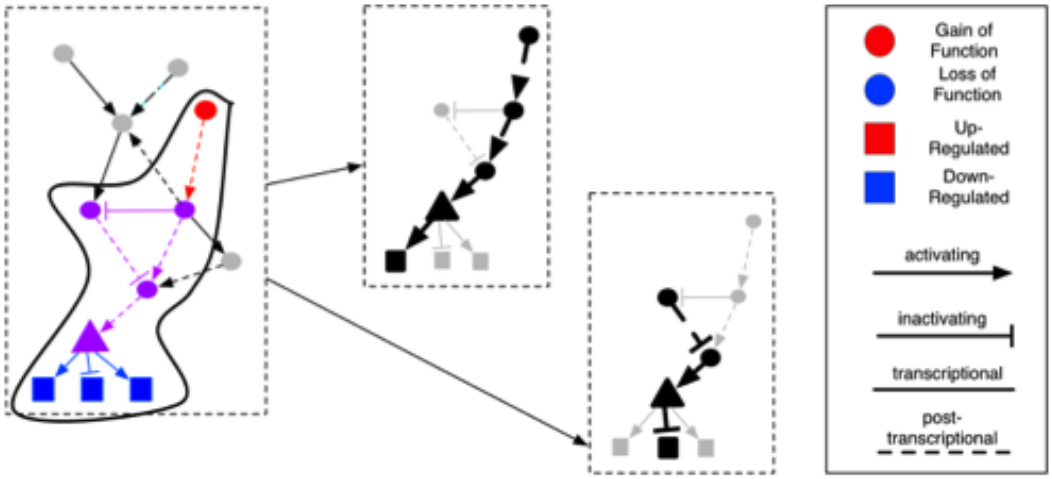

Figure 2.5: Extracting sub networks with a logical consistency filter. Interlinking sub-networks are extracted and logically consistent paths are identified.

edges are assigned a coherency score to reflect whether the source and target activities connected by the edge are in agreement given the interaction type encoded by the edge. Each edge i to j has an interaction type encoded in the full adjacency matrix, $A_{ij} = +1, -1$. Thus the product $AS(i) * AS(j) * Aij$ produces a measure of whether the activities of two interacting nodes in the pathway have activities consistent with their known interaction–positive scores are associated with consistent assignments and negative scores with inconsistent assignments. A logically consistent path in G is one that is composed entirely of (i, j) pairs for which this product is positive. Let ω be the ordered set of K nodes $(i_1, i_2..i_K)$ representing a path that connects a source node $i_1 \in S$ to a target node $i_k \in T$ through linker genes identified from tied diffusion. The coherency score of path ω is:

$$C(\omega) = AS(i_1) * AS(i_K) * \prod_{k=1}^{K-1} A_{k,k+1}$$

Note that only the activation scores for the source and target can be used here since linker genes by definition have no observed values. Paths found to have positive $C(\omega)$ scores are kept since the annotated up and down-regulation of the intervening path is consistent with the starting and ending activation logic [165].

2.2.4 Algorithm Summary

The following pseudocode summarizes the steps to generate the sub-graph g_c, given input source and target sets $\hat{X}_{source}, \hat{X}_{target}$ a user-supplied network size threshold S, graph G and linker function $f()$:

procedure TIEDIE($X_{source}, X_{target}, f(), S, G$)

$\hat{X}_{source} \leftarrow X_{source} * e^{-Lt}$

$\hat{X}_{target} \leftarrow X_{target} * e^{-Lt}$

$threshold \leftarrow max(f(\hat{X}_{source}, \hat{X}_{target}))$

while $\left| f(\hat{X}_{source}, \hat{X}_{target}) \geq threshold \right| \leq S$ **do**

$decrement(threshold)$

end while

$linkers \leftarrow \left\{ f(\hat{X}_{source}, \hat{X}_{target}) \geq threshold \right\}$

$g \leftarrow \{\}$

for $edge \in G$ **do**

if $edge \in linkers$ **then**

$g \leftarrow \{g \cup edge\}$

end if

end for

$g_c \leftarrow \{\}$

for all (paths $s \rightarrow t \in g$,where $\{s \in X_{source} : t \in X_{target}\}$) **do**

if $consistent(path)$ **then**

$g_c \leftarrow \{g_{consistent} \cup \{edge \in path\}\}$

end if

end for

return g_c

end procedure

2.2.5 Null Model/Significance Test

Gene sets that are biologically related are likely to exert a high degree of mutual influence over a pathway. I constructed a permutation test to determine the significance of the influence score by selecting one of the input data sets, in the case of two sets either x or y, and swap each gene value with a randomly chosen gene in the interaction network. To maintain the topology of the permuted set, genes of similar degrees (determined by binning all nodes according to degree in G) are swapped.

To do this, I use the relevance scores to measure the closeness in pathway space (G) of two or more input sets. I first propagate the scores in S through the network–

exactly as done in the core TieDIE algorithm–to obtain scores for all nodes starting from the source set, recorded in $x = r(x, A)$. Similarly, I obtain scores starting only from T and record these in $y = r(y, AT)$. A relevance cutoff α is then used to determine which nodes in G have high scores from both source and target set diffusion, using the same size factor s as used for the core algorithm. Let the relevant neighborhood for the sources, relative to a given cutoff α, be $S^*(\alpha) = i : xI > \alpha$ and for the targets be $T^*(\alpha) = i : yi > \alpha$, where the sets are written as functions to explicitly show their dependence on the threshold. I then compute a measure reflecting the degree to which the two input sets are related given our network and our choice of the relevance score. I then overlap the α-level relevance neighborhoods and compute the ratio of overlap between the two sets as:

$$RO = \frac{S^*(\alpha) \cap T^*(\alpha)}{S^*(\alpha) \cup T^*(\alpha)}$$

The ratio measures the proportion of mutually proximal nodes in G to both the source and target sets, which formalizes the simple intuition that genes related gene sets should be close in pathway space (connected by paths that are shorter than average). One could sweep through various choices of α and obtain a series of ratios, each time computing the significance of the overlap by a permutation test like the one described below. Instead, to match the core TieDIE algorithm as closely as possible, the null model sets α to a level that produces networks of similar size to the core result,

by using the same size control factor s and sweeping through α with the same sub-algorithm. In a typical experiment, 1000 permutations of the data are performed and the *RO* score generated with the real network is compared with the permuted scores to generate an empirical p-value [165].

2.2.5.1 A Parameter-Invariant Significance Test

In addition to the null model above, I have developed a generalized test that evaluates the closeness of two input vectors in pathway space, independent of any α parameter. Rejection of this null model can provide useful evidence when a user does not have enough prior knowledge to estimate the size of the sub pathway being searched, or when using TieDIE just to compute 'linker' scores to provide a ranking of genes.

Scores are obtained through diffusion over a pathway, as before, and recorded in $x = r(x, A)$, and $y = r(y, AT)$ vectors. I then use the observation that these pre-normalized vectors have a natural interpretation as a probability vector over the universe of genes in our pathway [42], and use the symmetric Kolmogorov-Smirnov test (KS-test) to quantify the distance between these distributions. Note that no parameteric form is required for this test, allowing us to use the emprical heat distributions of each vector safely. The symmetric version of this test is obtained by averaging the two KS-test scores computed by treating each vector as the reference distribution and the other as the empirical distribution to test against.

Permutations of the gene scores are obtained with the same procedure as described above, using a binning strategy to swap genes with others of similar degree in the

network. The symmetric KS scores produced with permuted data are then compared with the real KS scores to produce empirical p-values.

2.3 Validation on Synthetic Data

To validate TieDIE on a synthetic dataset, I created a toy network in which a pre-defined "core sub-network is embedded, and alterations to the state of expression of genes in the core network are assumed to contribute or "drive a disease process, while alterations outside of the core are assumed to have little effect.

A scale-free network topology was simulated using the NetworkX (Version 1.7) algorithm [83], and a core sub-network of twelve genes containing four "source genes and two "target genes was then embedded into the NetworkX network. For sources, the scores reflect the degree of alteration such as mutation frequency across a cohort or its predicted impact from sequence-based analyses. For targets, the scores represent the degree of differential expression observed in the targets of a transcription factor. I also simulated six false-positive sources representing genes with neutral hitchhiking events or passenger mutations. Six linking genes were simulated to connect the sources to targets, and I assessed the precision of TieDIE and a single-source diffusion (SDS) process in picking up the majority (4 out of 6) of these linker genes. In the case of single source diffusion, the algorithm was given the same mixture of true- and false-positive sources as TieDIE, including the two true-positive targets, considered as a single set.

The results of the single diffusion source (SDS) algorithm and the TieDIE

algorithm in discovering the six linking genes is shown in Figure 2.7. As expected, the single source process more frequently picked up nodes outside of the core network, which represents a lower precision. The TieDIE method, in comparison, captured fewer off-core false positive nodes, and the 3 false-positive nodes that it did capture are all directly connected to both a true source and target node.

I also tested to see if diffusion strategies used in this analysis outperformed the popular guilt-by-association-based approaches in which genes are implicated as involved in the disease process according to whether they are neighbors to involved genes in the pathway diagram. To do this, I implemented a k-nearest neighbor approach in which a gene was promoted to the positive set if a certain proportion of its direct neighbors were contained in either the source or target sets. The same simulation described above was used to introduce false-positives into the source set to gauge the ability of the methods to handle noise. As before, the parameters for all of the approaches were varied until they achieved a level of recall of 4 out of 6. In the case of the k-NN approach, the fraction of the neighborhood used as a cutoff to call a gene positive or not was the parameter varied. The results of the simulation are shown in Figure 2.6

I found that the diffusion approaches had significantly better precision compared to the k-NN approach at the 66% (4 out of 6 genes recovered) recall level imposed (a realistic compromise of sensitivity in some settings.) While single diffusion did achieve a higher proportion of precise predictions than the tied diffusion strategy, TieDIE had a higher precision on average with many fewer poor-quality predictions possible with the single diffusion approach (visible by the low-precision outliers in the SDS column in

Figure 2.7) [165].

It is important to note that this test represents an idealized scenario where both the network topology is simulated to match a scale free network, and the input sets are close enough to form a fully connected network when several "linker" genes are added. However, it displays the large gain in accuracy that can be acheived when we add knowledge of set membership to the input sets, along with the appropriate diffusion strategy.

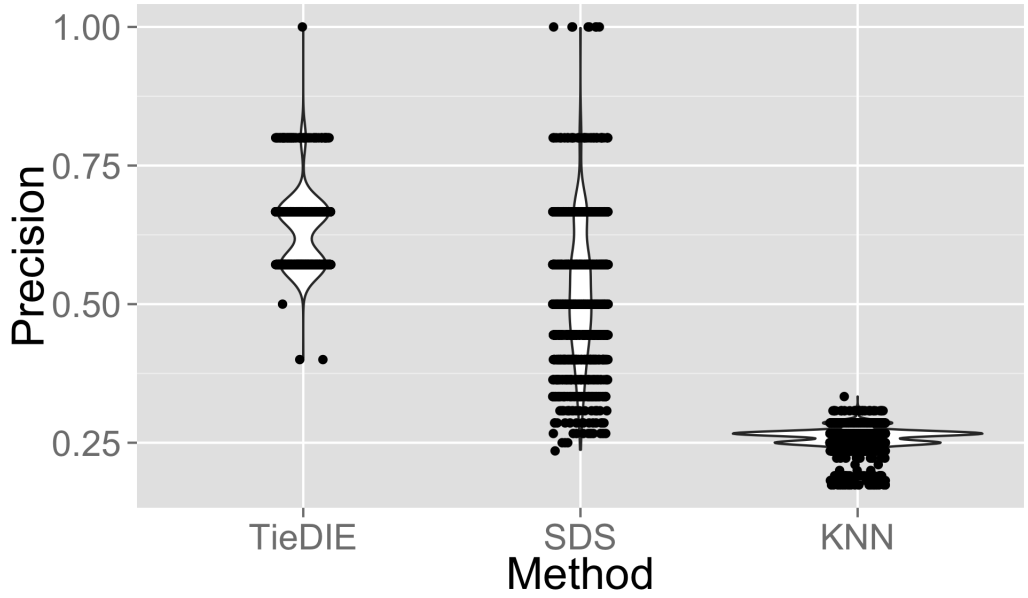

Figure 2.6: Boxplot of the precision of Single Source Diffusion (SDS) and TieDIE at finding linking genes in the core network on 20,000 simulated data sets, compared with a simple k nearest neighbors (KNN) classifier. Recall is fixed at 4/6 signaling genes in simulated trials, and linking genes outside of the core set of 6 are considered false positives. Jitter was added to the x-axis of the points as well as a small amount to the y-axis to allow viewing the quantized values of precision obtained across solutions. The width of the overlaid violin plot silhouettes is drawn proportional to the density of points for the corresponding value of precision. Drawn using an R script calling the geom_violin function available in the ggplot2 library version 0.93 (Wickham, 2009) [165].

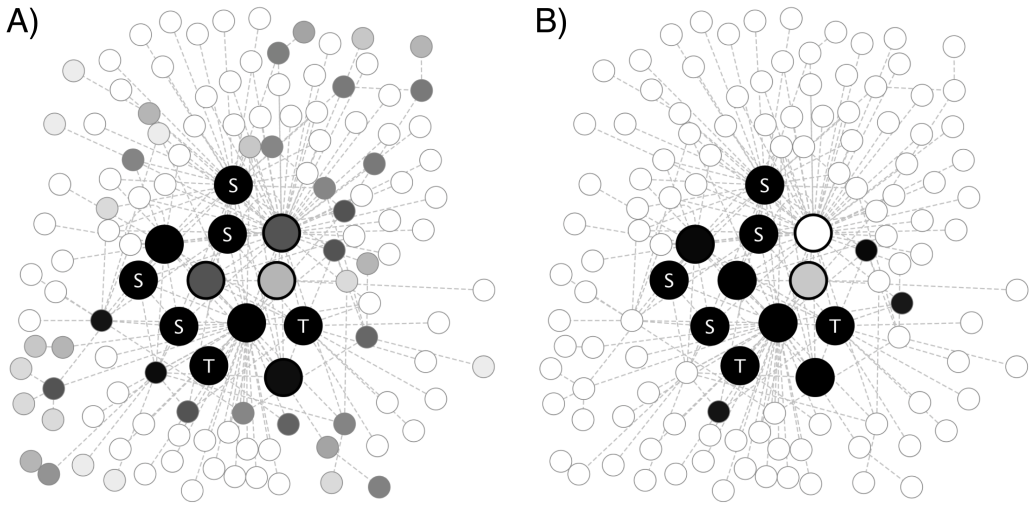

Figure 2.7: Frequency of a discovered core and off-core genes in single-source and tied-diffusion in a simulated network. (A) Single-source diffusion over the synthetic network. Darker colors indicate genes in a larger fraction of network solutions in repeated simulated trials at a fixed recall of 4 of 6 signaling genes. (B) The corresponding tied-diffusion frequencies at identical recall and test conditions [165]

2.4 Validation on genomic and transcriptomic data from patient breast-cancer samples:

2.4.1 Tied diffusion predicts breast-cancer related genes with higher precision than single diffusion approaches

The main assumption behind TieDIE is that more accurate pathways are obtained by directing diffusion processes to connect genes involved in somatic mutations to observed transcriptional effects. Therefore, I tested whether this approach could achieve higher accuracy than the comparable single-source approaches like HotNet that only consider the genomic perturbations without regard to the state of the transcriptome. I chose breast cancer as a test system because many mutations have been identified and

diverse types of data are available.

To determine if TieDIE networks were better able to find genes implicated in breast cancer, I collected gene lists from both WikiPathways-WP1984 [107] and retained a list of genes with 100 more known breast carcinoma mutations from the Catalogue of Somatic Mutations in Cancer (COSMIC) [3] version 57 [70]. Genomics and gene expression datasets were collected from the breast cancer analysis working-group of the Cancer Genome Atlas (TCGA) project [149]. At the time of acquisition, the breast cancer dataset included patient tumor samples for 533 patients and matched normal samples for a smaller subset, each with genomic sequencing and microarray expression data. To incorporate a diverse set of genomic and epigenomic alterations specific to a subtype, genes within regions of predicted copy number gain or loss based on the Genomic Identification of Significant Targets in Cancer algorithm (GISTIC) [139] were identified as having at least five samples with either high-copy amplifications or homozygous deletions [149]. I also included as sources the frequently mutated genes published by the TCGA Network (i.e. genes mutated in 410 tumors). Altogether, 110 genes were collected with significant numbers of events involving 41 amplified genes in 1258 samples, 14 deleted genes in 147 samples and 54 mutated genes in 1115 samples. One gene, BRCA1, was methylated in 15 samples and had gene expression inversely correlated in these tumors and so was also included. The background network contained a collection of curated transcriptional, protein-level and complex interactions for 4737 genes, proteins and abstract concepts, with 101,526 interactions [89]. To enrich for nodes with measured data, I removed complexes from the published SuperPathway after

transferring their interactions to the incident constituent members. For the purpose of the experiments presented in this document, I considered a protein a transcription factor if it had an outgoing regulatory interaction to at least one other gene in the SuperPathway.

TieDIE and single source diffusion were run multiple times, to produce a wide range of network sizes and the precision and recall of each validation set of genes was measured for each. The precision as a function of recall for identifying the positive control genes, either from COSMIC or from WikiPathways is shown in Figure 2.8

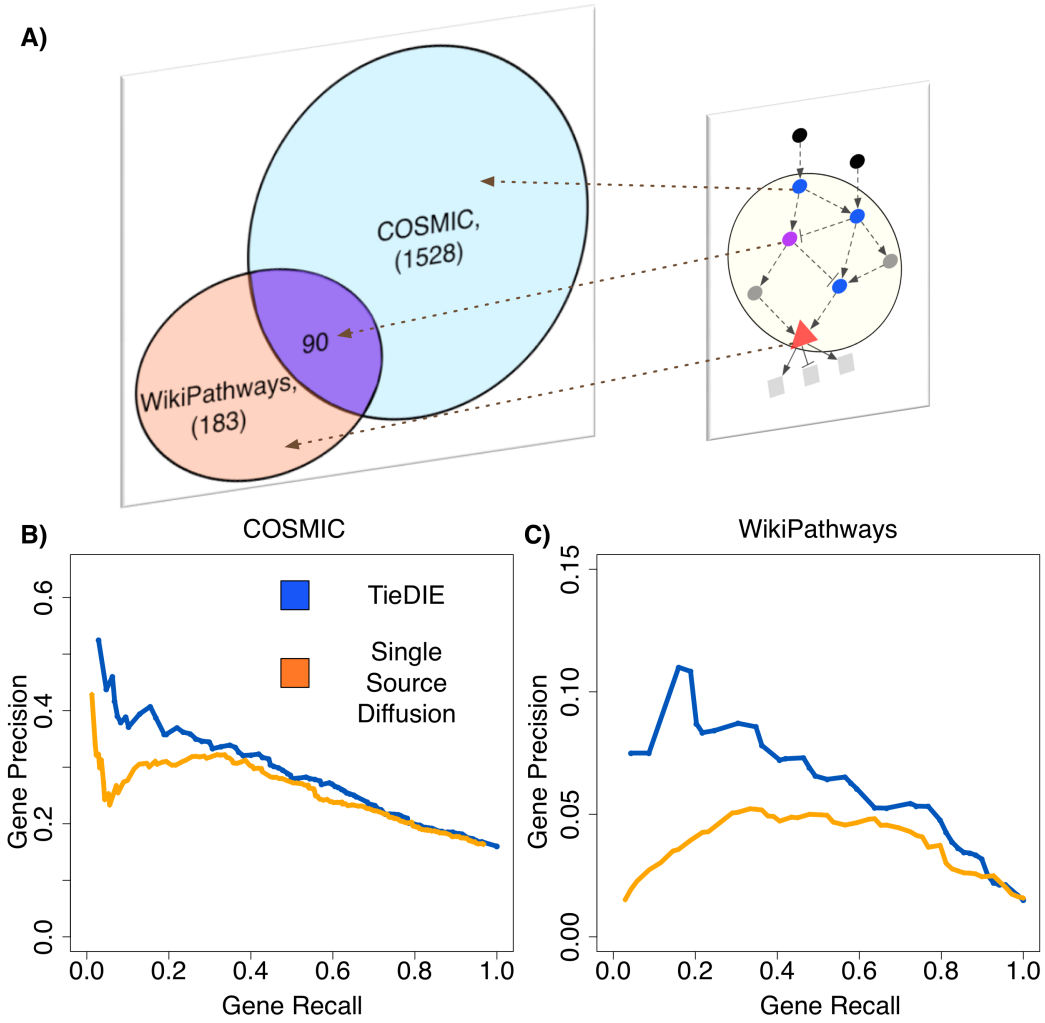

Figure 2.8: Accuracy of single and tied-diffusion approaches to recover known cancer genes. A) Known cancer genes used as positive controls were derived from COSMIC and WikiPathways collections. Any linker gene not in the input sets that was recovered in a diffusion solution was considered to be a true positive, if present in one of these validation sets. B) Accuracy of tied-diffusion (blue) compared to single-source diffusion (orange) in recovering breast cancer-associated genes as recorded in COSMIC measured in terms of precision (y-axis) plotted against recall (x-axis). C) Same as in B but using WikiPathway genes [165].

Compared with single-source diffusion, TieDIE demonstrated higher precision over an appreciable range of recall (by varying the size parameter) when used to predict breast cancer-implicated genes from the Catalogue of Somatic Mutations in Cancer and WikiPathways gene sets 2.8. Because WikiPathways includes genes related to breast cancer without any documented mutations, these results imply that the method may increase the ability to identify cancer essential genes as new drug targets. In realistic settings with 20% recall, TieDIE approximately doubles the precision for finding the genes. At recall levels higher than 3040%, the performance of TieDIE relative to single-source diffusion decreases but TieDIE is always equal or better than single diffusion methods. Including more sets may help increase the precision further (e.g. proteins with perturbed kinase activity). However, the linear decline in precision past these levels of recall indicates that any diffusion process searching sufficiently far from the input sets may be nomore effective than randomly drawing candidate genes [165].

2.4.2 Basal-luminal breast cancer networks on tumor samples.

To demonstrate the utility of the TieDIE method, I next used it to elucidate breast cancer-related networks that distinguish the major breast cancer subtypesthe so-called basal and luminal A subtypes. I chose to compare these two because they have a clear transcriptional signature compared with other subtypes that have a more intermediate or heterogeneous signature, such as the luminal B and HER2- amplified breast cancers. Basal breast cancer is known to be a more proliferative form of cancer compared with the luminal type, and the subtypes are thought to result from muta-

genic transformation of different originating cell types [144]. Because these subtypes respond differently to both general cytotoxic and targeted therapies [149], it is important to identify pathway mechanisms that differentiate the two subtypes to discover new treatments. The subtypes are easily identifiable from transcriptome signature analysis; however, different genomic alterations within each subtype can lead to the same transcriptional profile. For example, PIK3CA mutations or PTEN deletions are both correlated with the luminal A expression subtype [149]. On the other hand, it is possible that seemingly synonymous genomic alterations can lead to different transcriptional subtypes. For example, whereas TP53 mutations are enriched in basal cancers, some luminal cancers also harbor TP53 mutations. Thus, network-diffusion approaches may reveal how genomic alterations correlate with transcriptional signatures in a subtype-specific manner.

I performed a differential analysis between 99 Basal and 235 Luminal-A samples from the TCGA dataset. I used a set of 110 genomic perturbations published by the Cancer Genome Atlas Network [149] and applied a chi-square proportions test to find those that occur with significantly different frequency in one subtype as compared to the other. This uncovered 12 genes with mutations significantly associated with either the basal or luminal subtypes at the $p = 0.05$ level, used as the source gene set S. The target T set was defined by inspecting the differential expression of each TFs predicted target genes. Intuitively, TFs with activity more associated with basal tumors should have a set of targets with higher or lower expression than Luminal tumors, compared with random chance. To quantify this, I used a simplified version of the Califano

laboratorys MARINa algorithm [120] to find TFs with targets differentially expressed in basals compared with luminals. Predicted target genes were collected from the SuperPathway and SAM [41] was run to derive 'delta scores representing the degree of differential expression in basal tumors compared with luminal A tumors for each target gene. Gene set enrichment analysis (GSEA) [199] analysis was then used to identify TFs having targets with a non-random distribution of SAM scores. Rather than apply a strict multiple-hypothesis correction at the level of significance, I instead retained TFs with scores at a relaxed cutoff of $p = 0.05$ and then associated a relevance score to each retained TF by dividing its absolute GSEA score by the maximum absolute score. Similarly, source genes were retained that had a Fishers exact of $p = 0.05$ and were log-transformed P-value. These data were used to test various relevance measures for the following sections.

2.4.2.1 Evaluation of relevance scores:

I evaluated the precision of different methods for their ability to find true network paths in the presence of randomly generated false-positive 'decoy links, shown in figure 2.9. To generate these links, I added one random edge to the SuperPathway for every two true edges, leaving 33% of the edges false. Any paths in the final solutions that contained even a single decoy link were considered to be false positive. By computing the number of such paths that exist in the entire SuperPathway network, I calculated the total number of true-positive paths and used this to compute the precision and recall of the solutions. I compared tied-diffusion with its single-source counterpart and with

competing approaches such as prize-collecting Steiner trees as implemented by Dittrich et al. (2008), and Dijkstras allpairs- shortest-paths algorithm as a baseline approach. Three plug-in relevance score functions were included: heat-diffusion, Googles personalized PageRank and SPIA (Methods 2.2.2.3 2.2.2.2). Diffusion strategies performed comparably with the three competing methods at the specific levels of recall obtained with each approach (green, blue and red points in Figure 2.9). In addition, I found that tied-diffusion had higher precision over varying recall compared with the single-source equivalents. Also, the heat-diffusion relevance function performs comparably with personalized PageRank over moderate levels of recall, after which personalized PageRank outperforms the heat-diffusion kernel. Thus, given the significant computational benefits of using a precomputed kernel, I opted to use the heat-diffusion approach as the principal algorithm for my analysis as it allowed for a greater amount of flexibility and experimentation [165]. However, PageRank and other random-walk processes should be considered as alternatives, and future research might also consider the performance of various random-walk variations relative to the underlying network topology.

2.5 Visualization of Molecular Mechanisms Distinguishing Basal and Luminal Breast Cancer Samples:

The initial TieDIE evaluation dataset, including 99 Basal and 235 Luminal-A samples from the TCGA dataset discussed in the previous chapter (see methods 2.4.2), can be visualized to provide a pathway based view of the differential biology between

samples. The TieDIE network connecting basal-enriched genomic events, such as TP53, to basal-associated activated TFs, such as MYC, was found to be significant ($P < 0.002$) according to the background model (methods 2.2.5, and figure 2.10).

To make the visualization manageable, a relatively small size-factor of 0.85 was given to the algorithm (rather than the default of 1.0). As can be seen in figure 2.11, the paths in this network involve DNA damage checkpoint genes such as ATM, RB1, CCNE1, and CDKN2A, while luminal pathways include genes such as the estrogen receptor protein (ESR1), frequently mutated kinase PIK3CA and E-cadherin (CDH1) [165]. The basal and luminal subtypes also differ in copy number profile, with basal tumors characterized by amplifications in MYC, CCNE1, and deletions in CDKN2A, while luminal A tumors often have amplifications in CCND1, which acts in opposite fashion on the RB1 protein compared with basal samples [165]. This suggests that these luminal A tumors may have either flipped or lost the functional interaction between Cyclin D1 and RB1, or that the increase in transcriptional activity of RB1 can be explained by other upstream nodes such as TP53 or E2F1 [165].

Linking genes in the map may represent breast cancer "essential" genes whose functions are required for altered signaling logic in tumors. In support of this idea, inhibitors to PLK3 and HDAC1 were found to sensitize breast cancer cell lines [89], and targeting chromatin remodelers, such as the HDACs, are currently the focus of clinical trial work [134, 165]. Therefore, the breast cancer network found by TieDIE represents a data-driven visual summary of testable hypotheses that can explain the simultaneous protein activation, transcriptional activity and edge interactions found between many

of the key genes involved in breast cancer.

2.6 Application to Therapeutic Targets

I next tested the ability of the algorithm to generalize the breast-cancer model to a panel of breast cancer cell lines. This test has important implications for the clinical utility of TieDIE and related algorithms: if network predictions remain relatively consistent when taken from patient samples to cell lines, the potential for followup testing of network predictions on cell-line "surrogates", that have similar expression and genomic signatures to a given sample, is promising. If this is not the case, then alternatives such as patient-derived xenografts [203] are available.

I used TieDIE to infer basal-luminal networks generated from a completely independent data set collected in breast cancer cell lines. Data for a panel of 36 breast cancer cell lines (17 basal, 19 luminal) was obtained from the Gray lab at OHSU [89]. I used microarray gene expression data from these lines to get the "downstream" input heats and repeated the TieDIE analysis describe in section 2.4.2, using the original set of TCGA derived genomic perturbations as the source set. The latter "source" set was chosen because of both the sparse genomic data in cell lines as well as the lack of sample size needed to get a broad view of the genomic diversity of this disease. The resulting network was found to have a high degree of overlap with the Basal vs. Luminal-A network derived from TCGA data 2.12. This suggests that the transcription factor inputs are close in pathway space despite the relatively low specific overlap of

genes, allowing TieDIE to find a similar set of linker genes that represent the connecting topology between each pair of input sets.

To further assess the overlap of these networks I used DAVID [96] to test the enrichment of the intersection of these networks, and found it to be highly enriched in the expected Gene Ontology (GO) [18] biological processes related to metabolic regulation, anti-apoptosis, cellular signaling, differentiation, cell cycle and other known cancer processes. Interestingly, I found many of the most enriched terms for genes, that were present only in the TCGA network, were related to metabolism and biosynthetic processes. The gene lists for each network can be found in the supplemental data of the TieDIE Bioinformatics paper [165]. I hypothesize that this contrast in metabolism is due to the exceptional growth characteristics of cell lines, combined with the medium of growth provided in these lines. Patient tumors, on the other hand, would have fewer nutrients available in the extra-cellular environment and therefore be forced to increase internal production of macromolecules to support growth. It would therefore be interesting to determine whether growth conditions or the microenvironment could be altered to make cell lines more similar to tumors [165].

Overall, the results are promising: in spite of the relatively low overlap in transcription factors, the algorithm was able to find a highly similar network model and set of "linking" genes between the genomic background derived from TCGA patient samples, in spite of the relatively low overlap between TFs infered from patient and cell-line datasets, respectively. In each case, TieDIE was able to find a set of cancer related genes that link the genomic background to transcriptional changes in each data

set. This highlights the ability of pathway-based methods to generate generalizable models, through the use of prior knowledge.

In the directed networks used here, some of the proteins reside at the logical top of the pathway while others reside at the bottom. Based on this setup, I would expect the ones toward the top, the "master regulators, to produce more overall changes when their activities are modulated. To test this intuition, I selected 8 basal and 8 luminal breast cancer cell lines available in an RNAi knockout dataset [209] and selected all genes with a mean growth-inhibition z-score of -1 or less, or with a significantly increased growth-defect in basal lines ($p < 0.05$, SAM) to serve as a positive control. Ideally, these genes should be selected more often as master regulators if the pathway model is correct.

For each master-regulator algorithm, I found the intersection between the top k predictions and the validation set and compared it with the non-intersecting predictions to compute the precision-recall value for each. SPIA and PageRank were again run on the networks to identify proteins at the "top."

Each ranked "master regulator" was tested against a set of siRNA knockout data for breast cancer cell lines [209], and I used the rankings to compute the precision and recall in each case. We found that both methods (SPIA, PageRank) produced a significantly better precision than random orderings, at a given level of recall, both when used with the TieDIE network and the entire pathway (Figure 2.13). Predictions made with the TieDIE network achieved much higher levels of precision than those made with the entire network, which is largely a function of the methods ability to target key pathways and not necessarily due to any enhanced topological accuracy of

the TieDIE network. However, the fact that TieDIE was able to find enough of the correct topological features of the core cancer pathways to achieve significantly better precision than random suggests that the network solutions here should be useful for identifying key regulators that might be targeted therapeutically.

2.7 Sample-specific networks in Breast Cancer

2.7.1 Computational Analysis

I next applied diffusion approaches to characterize the specific pathways of individual samples. For each sample, I identified which of the TFs in the "downstream' setidentified by GSEA on the cohort-wide expression datahad at least one differentially expressed target for that sample. I then connected these tumor-specific active TFs to the genomic perturbations in that sample present in a scaffold network, a background network derived from a TieDIE solution from the cohort similar to the one shown in Figure 4 except using a smaller parameter to obtain a larger starting network of 106 nodes and 423 edges.

To test if TieDIE provides an accurate scaffold, I performed a search for sample-specific networks over TieDIE networks of multiple sizes as I as the entire SuperPathway. Using the procedure of adding random decoy links as described earlier, I measured the ratio of 'true and 'false paths in each sample-specific network solution, and plotted them for each choice of network "scaffold" (See figure 2.14). The precision of the sample-specific networks was plotted, and the mean precision was significantly higher when

using the TieDIE summary networks compared with using the entire SuperPathway. This precision declines gradually as the TieDIE summary networks increase in size, which is expected, given that they capture a larger fraction of the possible edges in the starting network and in the limit are identical to the SuperPathway. This shows that the improved precision achieved by the TieDIE summary networks is transferred to sample-specific networks, thereby using data from the entire cancer cohort to inform the network predictions for individual samples.

I evaluated the ability of TieDIE to connect differentially expressed genes to at least one of the genomic perturbations represented in the sample-specific network. I tested differentially expressed genes that were downstream of the input set of TFs, in the cohorts TieDIE network. For the majority of samples, these sample-specific networks explained a significant fraction of differential expression (20-80%), although for a subset of samples none of the expression could be explained 2.15. This may be because of missing links in the starting network, which greatly affects the TieDIE solutions, because the method cannot infer missing links. In addition, genomic perturbations that only occur in a small minority of samples may not be represented in the TieDIE network because the method, by design, will filter out pathway elements that cannot be supported by a sufficient number of samples (though by abstracting to pathway space it can often find groups of mutually exclusive, low frequency events). This trade-off increases the overall quality of the networks, at the expense of missing potentially novel, but rare, molecular mechanisms that may drive the cancer phenotype in a small minority of samples.

2.7.2 Mapping the network of an abnormal luminal A tumor

A major goal in cancer systems biology is to infer a specific network for each patients tumor and, as data become available, each subclone identified within the tumor. Accurate network models could be used to explore a large space of potential targets to kill the tumor in silico. Therefore, I applied TieDIE to identify a pathway solution for every tumor sample in the TCGA breast cohort. To illustrate the results, I focus here on the non-canonical tumor sample TCGA-BH-A0BR, which had an intermediate pathway state between the classic basal and luminal A subtypes as evidenced in its CircleMap plot (Figure 2.16). Tumor heterogeneity may contribute to such "mixed" samples or may reflect a tumor evolution distinct from the classic basal and luminal pattern. The sample-specific analysis of this tumor reveals it has a hybrid set of genomic perturbations with both luminal A-like events, such as an AKT1 mutation, and several basal-like events, such as a TP53 mutation, and amplifications in insulin-like growth factor receptor (IGF1R) and PAK1.

Even though the patient sample has a luminal-associated AKT1 mutation, the surrounding network is more consistent with wild-type AKT1 activity. Namely, HIF1A is active, reflecting a basal program of hypoxic response and angiogenesis, further evidenced by increased EDN1 expression. In addition, IRS1 and PIK3CA expression are basal-like, and these network properties are maintained by a basal-like PAK1 amplification that, in this patient, may promote the activity of RAC1 and MAP2K1.

Interestingly, IGF1R is known to be involved in the control of breast cancer

cell growth. Blocking or reducing the activity of this receptor has been found to reduce growth in at least one luminal breast cancer cell line [82], and increased sensitivity in trastuzumab-resistant cells are associated with IGFR1 as well as PAK1 [173]. The TieDIE network suggests that IGF1R and PAK1 amplifications may be driving the growth signaling pathways of this tumor in the absence of HER2 amplification. Experimental validation of such hypotheses are difficult to perform within patient samples, but the concordance of TCGA sample-expression-derived and cell line-expression-derived networks (see above) suggests that experiments in cell lines could be used to address such patient-specific hypotheses in the future.

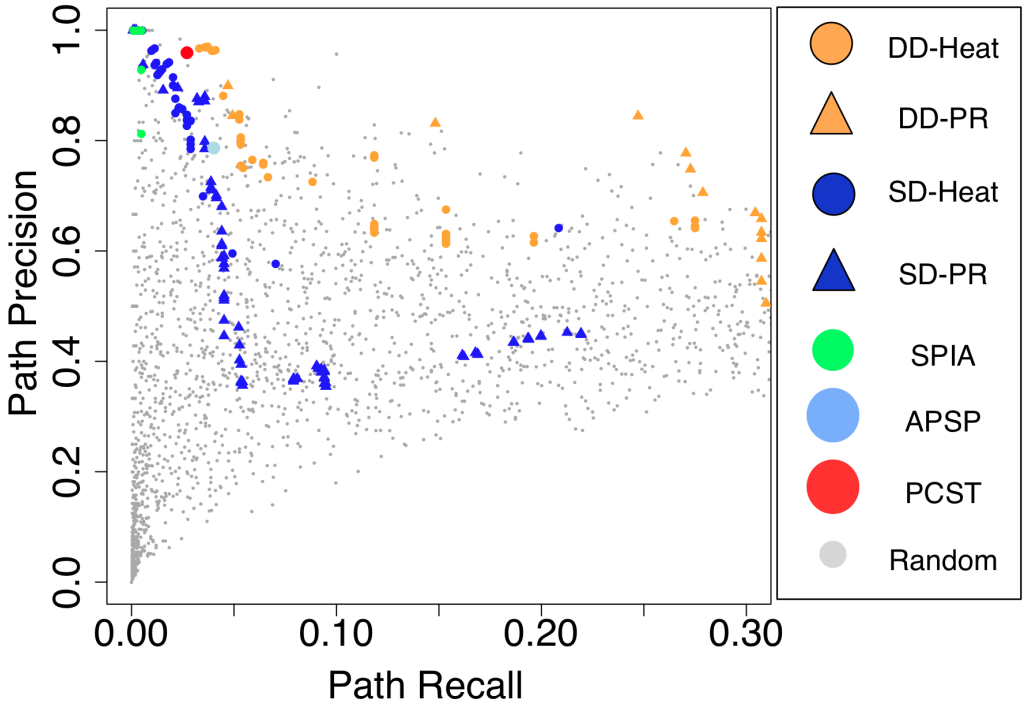

Figure 2.9: Precision of single-source (blue points) and tied-diffusion (orange points) with different relevance scores 2.2.2 for identifying pathways in a breast cancer. Any paths containing even a single randomly injected 'decoy link were considered false positives. Recall measures the number of logically consistent paths (Methods 2.2.3) out of the total possible; precision measures the number of such consistent paths in the total number returned. Relevance scores tested are heat diffusion (circles), personalized PageRank (triangles) and SPIA (green circles). For comparison, included are all-pairs shortest paths (APSP; blue circle) and prize-collecting Steiner trees (PCST; red dot). Randomly generated networks of various sizes were obtained to estimate the background distribution (gray dots). Different levels of precision and recall were obtained by varying algorithm parameters (e.g. the α parameter for single and tied diffusion; [165]

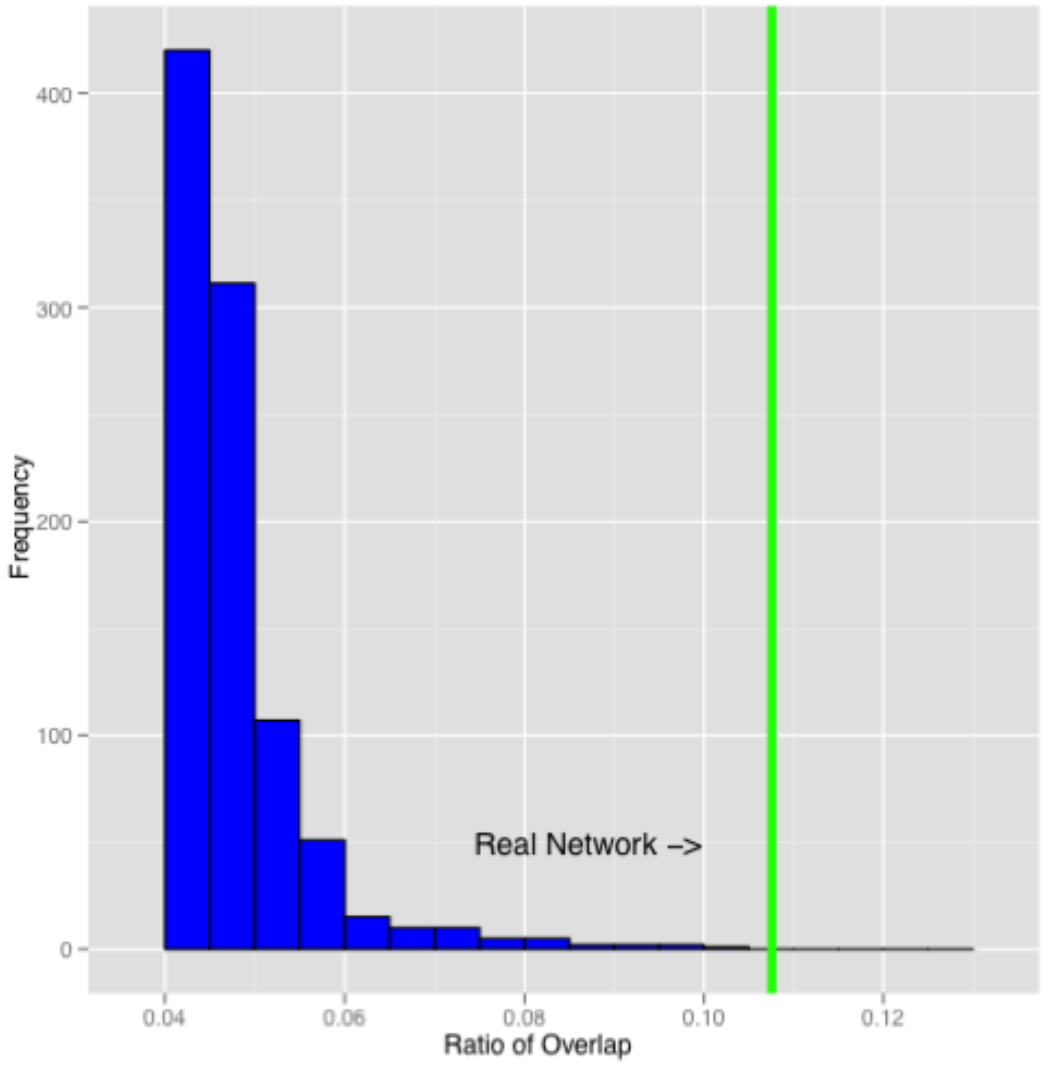

Figure 2.10: Ratio of overlap scores under the null model, for the Basal vs. Luminal network. The score for the real (non-permuted) network is shown as a green vertical line.

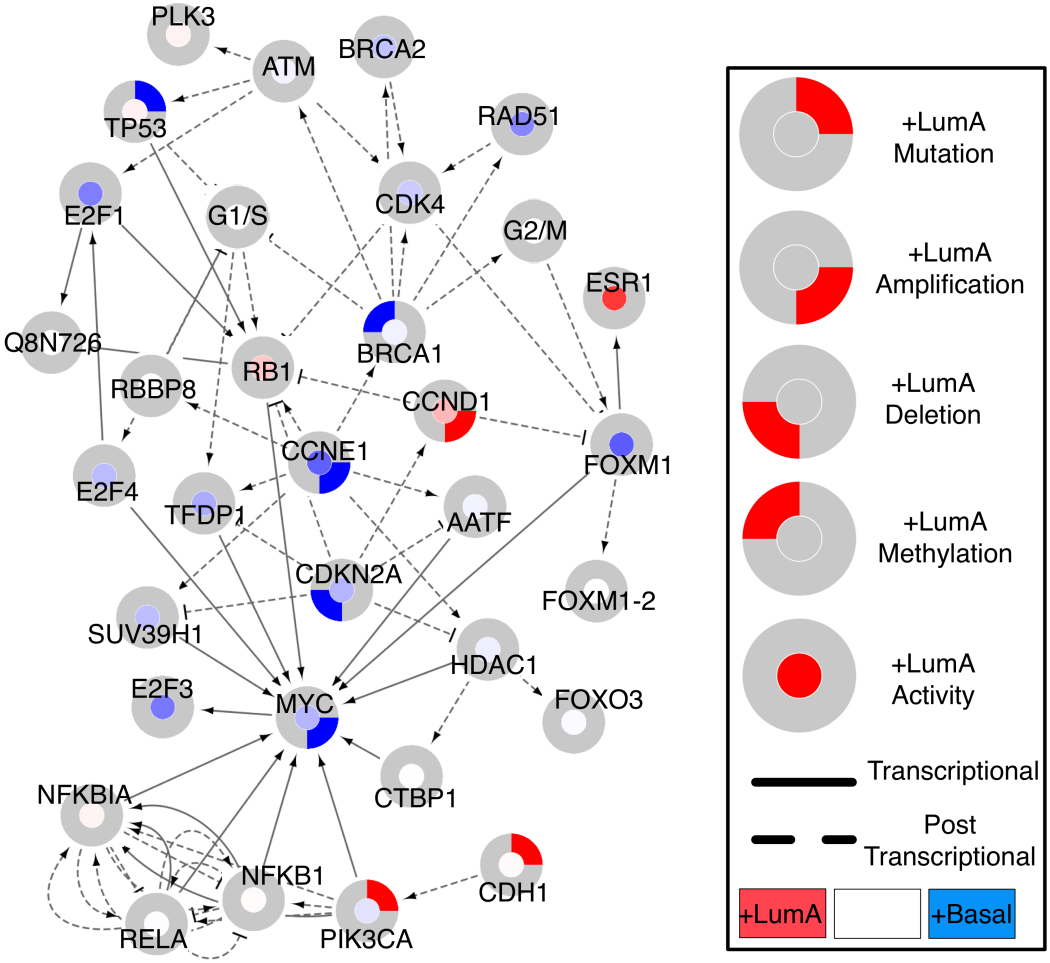

Figure 2.11: Tied-diffusion result for luminal A versus basal breast cancer subtypes. The inner coloring of the rings represents the differential expression in luminal A as compared with basal samples. The outer ring represents differential frequency of genomic perturbations in luminal samples as compared with basal samples: differential mutation (upper right), amplification (lower right), deletion (lower left) and DNA-methylated CpG islands near the promoter (upper left) [165].

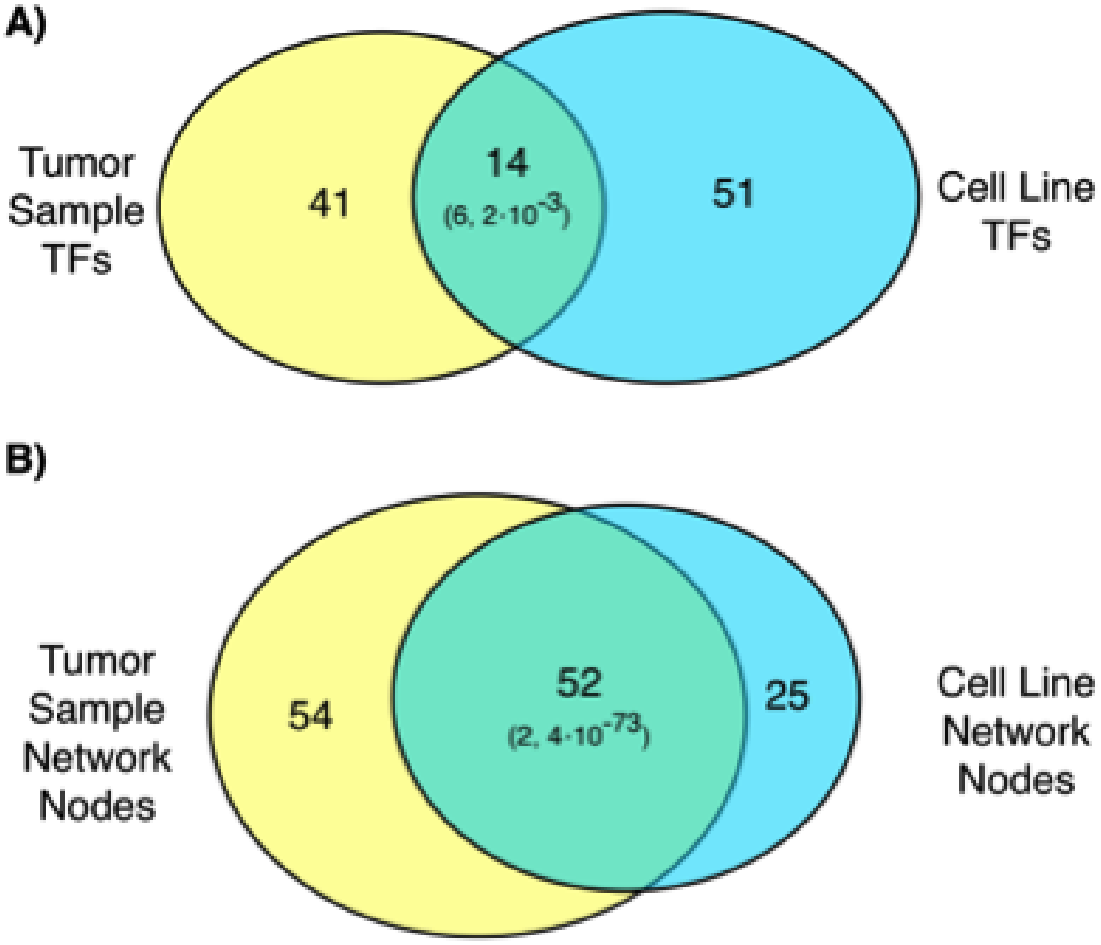

Figure 2.12: Figure S5: Overlap between TCGA sample- (yellow circle) and cell line-derived (blue circle) networks. All numbers reflect the number of network proteins. A. Comparison of transcription factors differentially active in tumors (yellow) compared to cell lines (blue). B. Comparison of linker genes uncovered by TieDIE for tumor- (yellow) versus cell line-derived (blue) solutions

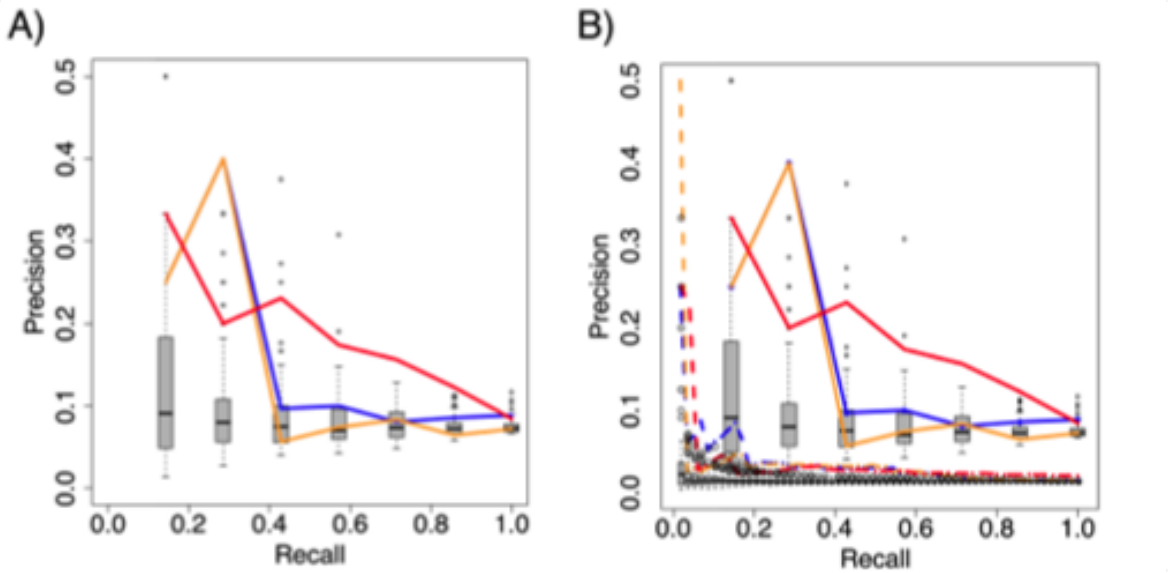

Figure 2.13: Precision/Recall for various "master regulator" algorithms, in the cell-line specific breast cancer network. A) Precision of two master regulator node-importance ranking algorithms (SPIA, PageRank) as well as node out-degree. A node is considered a true positive if it was found to cause a significant growth defect in the siRNA inhibition dataset (Turner et al., 2008). Grey boxplots show multiple random orderings of genes in the network. B) Precision of master regulators from the TieDIE network plotted against rankings generated from the entire SuperPathway.

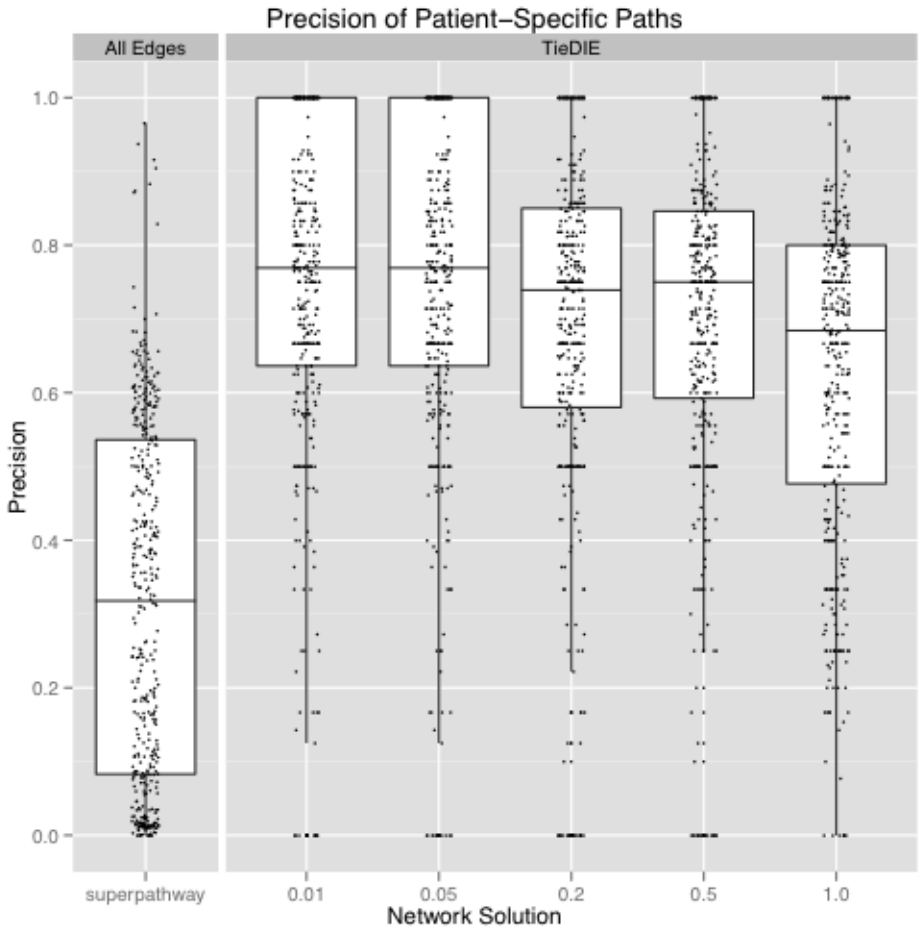

Figure 2.14: Precision of patient-specific paths over the SuperPathway and TieDIE networks of varying size. The plot of precision over the entire SuperPathway network is shown on the left; on the right, precision for 5 TieDIE networks of increasing size is shown. The number of linker genes in each TieDIE solution, relative to the number of genes in the input sets, is shown in the x-axis and ranges non-linearly from 0.01 to 1.0.

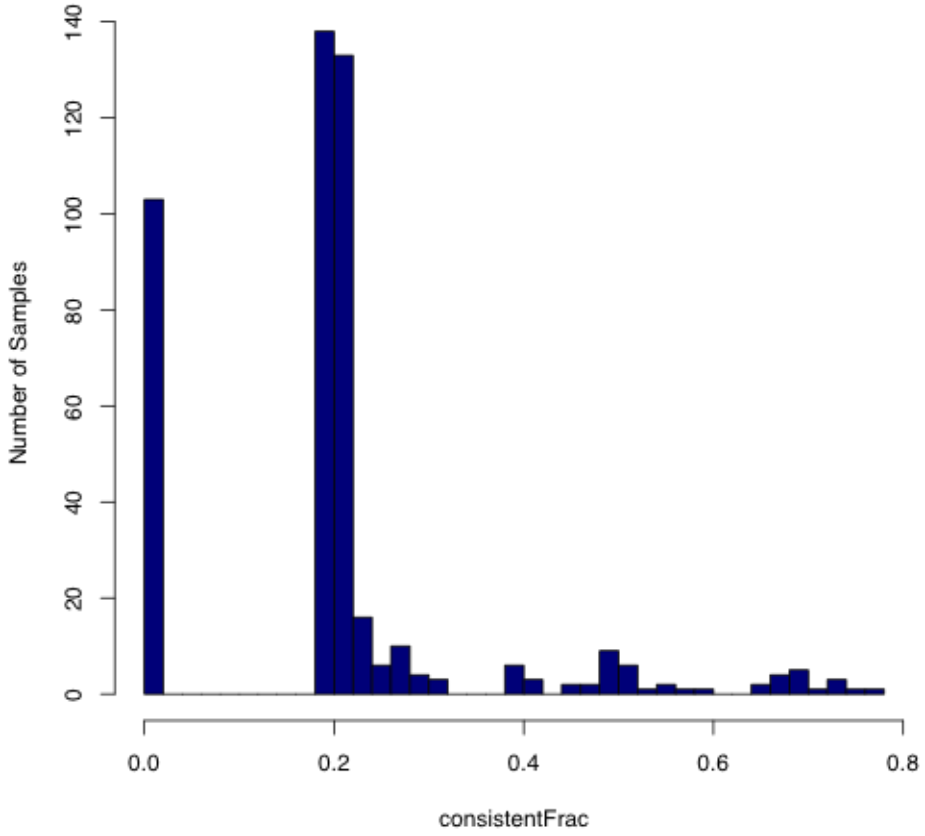

Figure 2.15: Histogram of the fraction of differentially expressed genes (downstream of our initial set of transcription factors) that are explained by genomic perturbations in that same sample. For most samples, we found logically consistent paths explaining 20 - 80% of the differentially expressed genes.

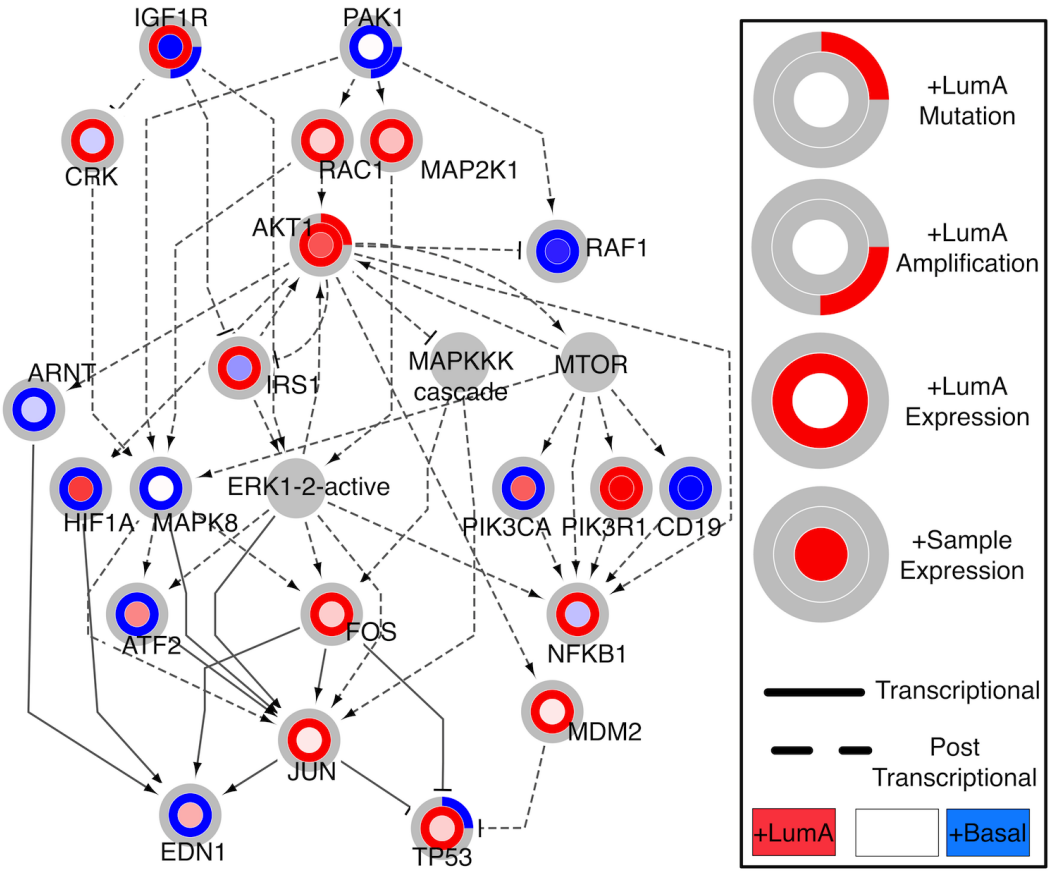

Figure 2.16: Luminal A sample TCGA-BH-A0BR specific network reveals basal-like molecular behavior. The network connects genomic perturbations in the sample, red or blue rings around nodes, to transcriptional changes in the same sample, inner node coloring. Red and blue colors indicate higher and lower cohort mutation rates in luminal A samples as compared with basal, outer ring; overall cohort differential expression of luminal A compared with basal samples, second ring; individual sample expression, inner circle. Transcriptional interactions, solid lines; post-transcriptional interactions, dashed. Activating interactions, arrow at the target node; inactivating, flat bars.

Network Analysis and Applications of the TieDIE Algorithm

3.1 TieDIE finds effects linking histone modification to transcriptional changes related to growth

3.1.1 Introduction

As with normal biology, cancer cells are characterized by a specifc pattern of epigenomic state, that are believed to contribute to the cancer phenotype [26]. While histone modification accounts for cell type specific gene expression patterns [88] in normal cells, specific patterns of modification, and their effect on transcription, have appeared for different cancers [26, 73]. For example, abberant DNA methylation can induce cancer by silencing tumor supressor genes by either by CpG island promoter hypermethylation [68], global hypomethylation [68, 90], or loss of acetylation at histone H4 [71]. In addition, the SWI/SNF (SWItch/Sucrose Non-Fermentable) complex, a conserved chromatin-remodelling complex that mobilizes nucleosomes and implicated as

having tumor-supressor activity (or required for the activity of other tumor-supressor gnes) [176], is known to have frequently mutated constituent proteins in renal carcinoma [216].

Two large-scale TCGA studies in renal-cell and bladder carcinoma, that I have been heavily involved in, also found frequent alterations in SWI/SNF complex genes, including PBRM1, ARID1A, SMARCA4 and others that are thought to contribute to the tumor phenotype [150, 151]. Through the use of prior knowledge, pathway-based methods have the potential to find at least a fraction of the (likely wide-ranging) effects of the alterations seen in these chromatin remodelling genes. These priors include protein-protein interactions that can result via complex formation, and regulatory or functional interactions between genes.

Genes that are found to interact with mutated chromatin-remodelling genes may represent the targets of mutation-driven epigenetic changes, and pathway-based methods are needed to prioritize genes using this additional prior evidence. In particular, pathway interactions with transcriptional "master regulators" that already have good evidence for altered activity in SWI/SNF mutants, based on the aggregate expression of their downstream targets, represent additional evidence that a given transcription factor is implicated in disease.

For each of these studies, I used the TieDIE algorithm to connect these mutated histone-modifying genes to transcriptional regulators that modulate a large fraction of the gene-expression signature seen in tumor samples, testing the hypothesis that the mutations in question have a causal impact on the expression signature. In each case,

pathway data, in combination with the TieDIE background model (Methods 2.2.5) supported this hypothesis, and TieDIE was able to select specific histone-modifying proteins, transcription factors and pathway interactions that underly these respective results. In addition, key "linking" genes that may mitigate the signal between mutated proteins and transcription factors were found in each case, expanding the list of potential therapeutic targets available for followup study.

3.1.2 Kidney Cancer

In a study of over 446 TCGA samples of clear cell renal carincoma, the TieDIE algorithm was used to integrate mutation, expression and pathway data. The "source" set chosen by selecting any MutSig gene with a q-value of 0.05 or smaller, which resulted in 19 genes (out of the original 39 significant at the 0.10 level found by MutSig using the pre-validation MAF).

To select the "target" set of transcription factors (TFs) , the PARADIGM algorithm was used to find transcription factors with high predicted activity in tumor samples, as compared with normal controls. Briefly, this procedure infers integrated pathway levels (IPLs) for genes, complexes, and processes using pathway interactions and genomic and functional genomic data from a single patient sample using a Bayesian statistical model [150].

All transcription factors with an average unsigned PARADIGM IPL greater than unity that also had at least one transcriptional target with an IPL above unity were selected as active hubs. An IPL above unity corresponds to having an activity of

at least one standard deviation more extreme than levels seen in normal controls. These selection criteria allowed active transcription factors with relatively few targets to be included while controlling for well connected "hub" transcription factors found to be inactive given the pathway context. This resulted in the selection of 115 transcription factors. The TieDIE solution connecting these transcription factors to the 19 MutSig genes was found to be highly significant (Figure 3.19). The resulting network contained 529 genes connected by 10,707 interactions (3396 HPRD-PPI, 4052 regulatory, 3259 component; $p \leq 0.017$). In this network, 14 (74%) of the sources were connected by some path to 115 (100%) of the targets involving 400 interconnecting linking genes.

Pathway enrichment analysis revealed that 5 genes (PBRM1, SETD2, BAP1, KDM5C, and ARID1A) participating in chromatin remodeling were overrepresented beyond chance expectation in the list of 19 significantly mutated genes based on MutSig analysis. To elucidate the pathways that may be disrupted due to mutations in these chromatin genes I extracted a chromatin-related sub-network of the TieDIE solution. A graph traversal was used to search for paths linking mutated chromatin genes to those genes with gene expression levels significantly correlated or anti-correlated with mutations in any of the chromatin genes. Correlation was determined by performing a t-test in which samples having a mutation in one of the 5 chromatin gene were grouped into one set and those without a mutation in any of the 5 were grouped into a second set. A two-sided t-test was then calculated for each gene using either the genes expression levels or its IPLs from PARADIGM and links connecting genes with t-statistics with and FDR ≤ 0.10 were retained. A depth-first search was then used to find all

paths connecting the chromatin-related genes to IPL-correlated "signaling layer" genes, through up to one "linking" gene.

Similarly, I connected the IPL-correlated signaling genes to expression-correlated "output" genes through active transcriptional hubs. The final sub network was defined as the union of all complete paths connecting the chromatin-related genes to "output" genes through the linker, signaling and transcriptional-hub layers.

The TieDIE solution gives clues into the various pathways affected by modulation in the chromatin related genes (Figures 3.17, 3.18). Of particular interest is the finding that the chromatin complex made up of PBRM1, ARID1A, and several SMARCA proteins was found to interact with NFKB1. In addition, central to the network are genes involved in TGF-beta and Wnt-related signaling. For example, betacatenin (CTNNB1) has higher activity in non-chromatin mutants. This suggests that the more advanced disease stages are driven by pathways involving beta-catenin activation, which in turn would then activate such targets as MYC leading to well-known de-differentiation programs seen in many aggressive cancers.

Several interlinking genes in the TieDIE solution are of particular interest because they may be overlooked when the data is analyzed without consultation of the known and/or predicted pathway interaction logic. For example, JUN, FOS, and SP1 are major transcriptional regulators that are inferred by PARADIGM analysis to be active in many of the samples (see outer rings of these genes in Figures 3.17,3.18). However, neither the inferred activities nor the expression of these genes is associated significantly with chromatin-specific genomic perturbations. However, these transcription factors

together interconnect several genes that are associated with chromatin mutations from the signaling layer, such as COBRA1, to genes in the transcriptional output layer, such as estrogen receptor (ESR1), TGF-beta, and IL6 differential expression. The existence of such connections suggest that mutations in chromatin modifiers enable particular transcription factor linkers such as JUN and FOS to express sets of growth factor receptors, cyclins (e.g. CCNB1) and interleukins leading to a global turn-over in the signaling circuitry of tumor cells.

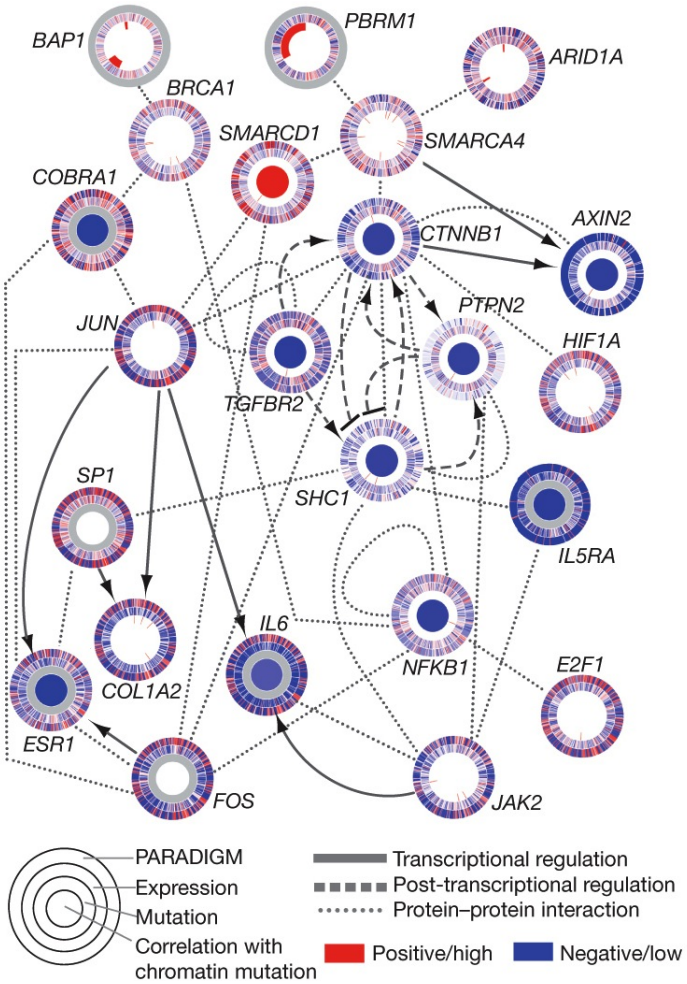

Figure 3.17: sub network Describing the Impact of Chromatin-Related Mutations in Kidney Cancer. TieDIE was used to assess the impact of mutations in genes known to participate in chromatin-remodelling processes (PBRM1, ARID1A, BAP1, SETD2, KDM5C), and identified as significant by MutSig, in a TCGA study of clear cell renal carcinoma [150]. TieDIE identified a significant sub network connecting 3 of these genes (PBRM1, ARID1A, BAP1) to active transcriptional hubs as identified by the PARADIGM method. Each gene is shown as a multi-ring circle with multiple levels of data, so that each 'spoke' in the ring represents a single patient sample. PARADIGM ring, bioinformatically inferred levels of gene activity (red, higher activity); Expression, mRNA levels relative to normal (red, high); Mutation, somatic event; centre, correlation of gene expression or activity to mutation events in chromatin-related genes (red, positive).

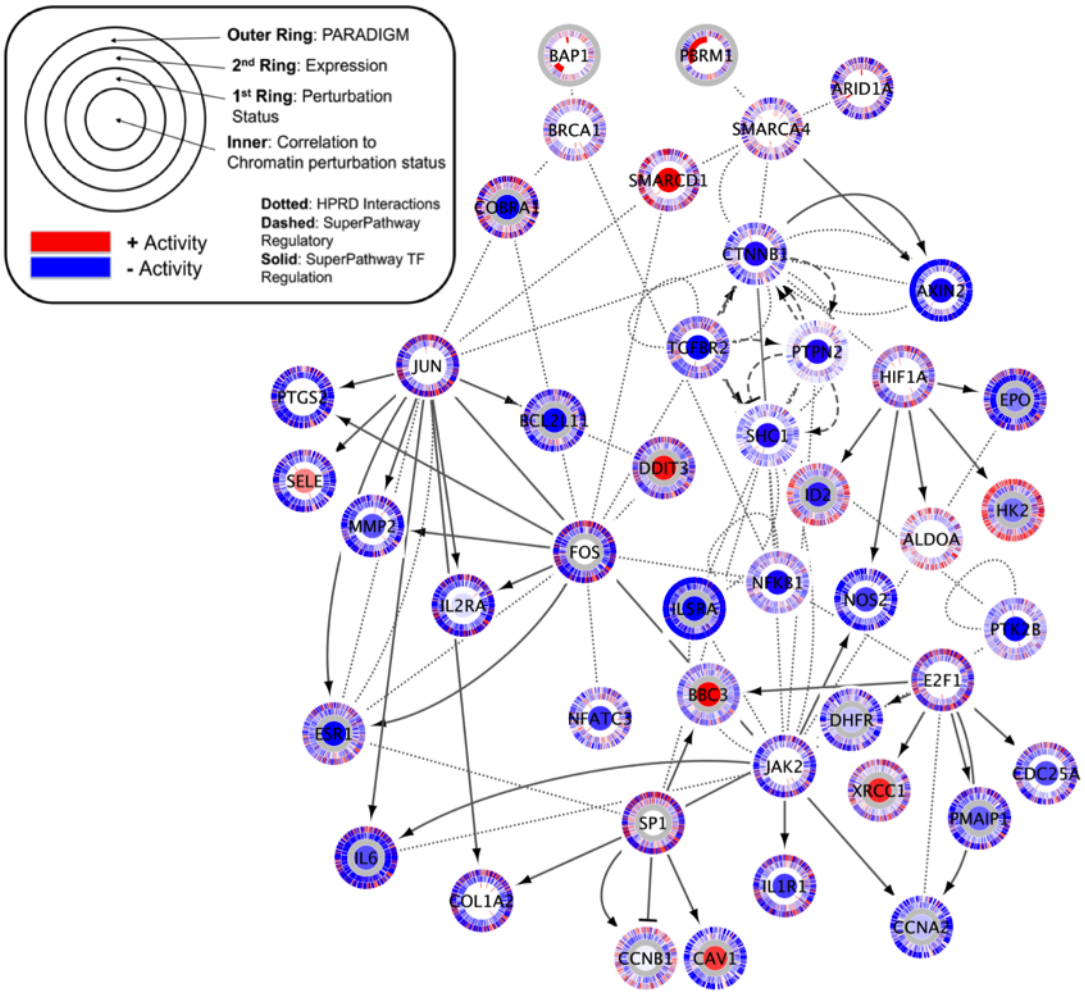

Figure 3.18: Chromatin-related TieDIE solution, alternate view showing the original, non-discriminant data and inference levels. Same solution as in figure 3.17 with an expanded view by searching all paths from mutated genes to transcriptional targets up to depth 4 (vs. 3), and original pathway inferences and gene expression levels are shown instead of the differential levels.

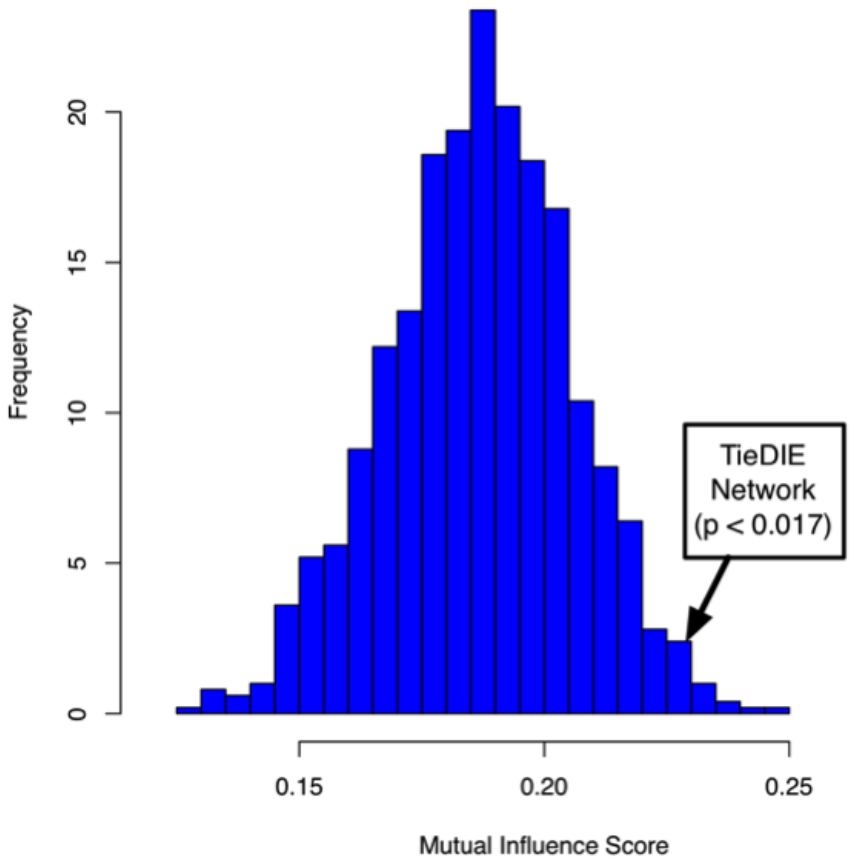

Figure 3.19: Genomic perturbations in kidney cancers are significantly associated with downstream transcriptional changes through known and novel pathway circuitry. The TieDIE algorithm was used to identify a network connecting the top 19 significantly mutated genes to transcriptional hubs identified from the identification of transcription factors with targets significantly up- or down-regulated in tumors relative to normal controls.

3.1.3 Bladder Cancer

This TCGA study of urothelial bladder carcinoma collected data from 19 tissue source sites, consisting of 131 chemotherapynaive, muscle-invasive, high-grade urothelial tumours (T2-T4a, Nx, Mx), as well as peripheral blood (n5118) and/or tumour-adjacent, histologically normal-appearing bladder tissue [151]. Genomic and transcriptional data was collected, as in the kidney cancer study (above), and MutSig analysis was performed to assess the driver potential of observed SNPs in tumor samples, again through the TCGA/Broad Institute Firehose pipeline. PARADIGM was again used to infer IPL levels for all genes.

An initial "global" TieDIE analysis was performed to assess whether the genes with predicted driver mutations were likely to influence transcriptional "master regulators" that modulate the gene-expression signature observed in tumor samples. Therefore, the TieDIE source set was defined as the 29 genes identified as significant by MutSig analysis. For the targets, a gene-expression signature between tumor and normal controls was defined using edgeR [180], using RNA-Seq data. Gene Set Enrichment Analysis (GSEA) [199] was then used to identify transcription factors (TFs) having targets with a non-random distribution of significance scores, resulting in the selection of 26 transcription factors including JUN, FOS and MYC/Max. Not surprisingly since RB1 is one of the genes in the SMG, several retinoblastoma-pathway TFs were selected including HDAC1, E2F1, E2F4, and TFDP1.

The TieDIE solution was found to be highly significant ($p < 0.008$; Figure

3.20, part A), and the resulting network (Figure 3.21) contained 103 genes connected by 1,233 interactions (523 HPRD-PPI, 409 regulatory, 301 component). 24 (83%) of the sources were connected by some path in this network to all 26 of the targets, with 56 interconnecting linking genes [151].

The resulting linking set of genes in this "global" solution contained over a dozen "linker" genes with high betweeness centrality measures (Figure 3.21). While many of the highly central genes by this measure correspond to the starting input set (TP53, RB1, HDAC1, E2F1), several genes were found as linkers that were not part of either the source or target sets used as input to TieDIE that increase the overall connectivity of the solution. These genes are depicted as large white nodes in Figure 3.21. Among these linkers were several cycle-cycle related genes such as CCND1, CDC25A, CDK1 and CDK4. Also among the list was the DNA repair gene, BRCA1 [151].

To further investigate the specific effects of mutations in histone-modifying genes, I selected 23 genes with known histone-modifying activity and at least 1 non-synonymous mutation, and weighted these genes by their mutation frequency to define the source set. For the targets, I divided the samples into 2 groups: one containing at least 1 non-synonymous mutation in the list of histone- modifying genes (100 samples) and those without (28 samples), and used edgeR to [180] to rank the genes by differential expression between those groups. Gene Set Enrichment Analysis (GSEA) [199] was then used to identify transcription factors having targets with a non-random distribution of edgeR [180] significance scores, which resulted in the selection of 35 transcription factors, weighted by GSEA significance. TieDIE was run on this source and target set, producing

a highly significant network (Figure 3.20, part B) with 107 nodes and 2463 edges (603 HPRD-PPI, 322 regulatory, and 1538 component; p <0.001) [151].

To further focus this network solution on paths relating genomic alterations and transcriptional changes, only through relevant "linking" proteins, PARADIGM was run on the samples to produce Inferred Pathway Levels (IPLs) [151] for all genes, including those not in the "source" or "target" sets. In this case, PARADIGM inferences were used in place of direct measurement of direct, phosphoyrlated protein quantification via RPPA or Mass-Spectrometry. A two-sided t-test was calculated for each gene using the IPLs, and 12 genes were found to correlate with histone status (p <0.05, uncorrected).

I generalized the TieDIE algorithm to use the original source and target sets along with this additional IPL-correlated set of genes, with the algorithm set to find linker genes with high heat in either the source and IPL-correlated sets, or the target and IPL-correlated sets. This corresponds to a generalized relevance function $z = f(r(a;A);r(b;A);r(c;A)) = max(\ min(r(a;A),r(b;A)),min(r(b;A),r(c;A))\ )$, where a, b, c represent the source, IPL, and target sets, respectively (Methods 2.2.2). Intuitively, this sets the model constraints to find two independent regions of the network connecting source genes to the IPL-correlated "signaling" genes, and the "signaling" genes to the target gene set, respectively. This constrains the results to include only paths from source to target sets that flow through the "signaling" set. As expected, the resulting network was notably smaller (49 genes, 600 edges), and I ran an additional filtering step by performing a graph traversal from the 3 most mutated histone-modifying genes (EP300, CREBBP, MLL) to the 7 most differentially active transcriptional hubs

(TP53, MYC, MAX, MYB, HES1, FOXA2, HSP90AA1), selected by prior knowledge and TCGA working group consensus. The resulting network (Figure 3.22 contained 24 genes, including 4 of the IPL-correlated input nodes (SP1, HNF4A, FOXA2, CD19), and 66 edges (55 HPRD-PPI, 10 transcriptionally activating, 1 post-transcriptionally activating).

The linker gene with the highest centrality in the "global" network solution, CREBBP, was also found in this chromatin-remodeling sub-network. This gene is a transcriptional coactivitator of several transcription factors that couples chromatin remodeling to transcription factor recognition involving growth, homeostasis, and development.

Interestingly, though 17 samples in the BLCA cohort have non-silent mutations in CREBBP, the level of significance did not reach the cutoff for CREBBP to merit inclusion into the SMG by Mutsig analysis. Therefore, the TieDIE analysis provides an important orthogonal perspective on the role of CREBBP, further implicating it as a potential driver of bladder carcinogenesis based on it's surrounding pathway context.

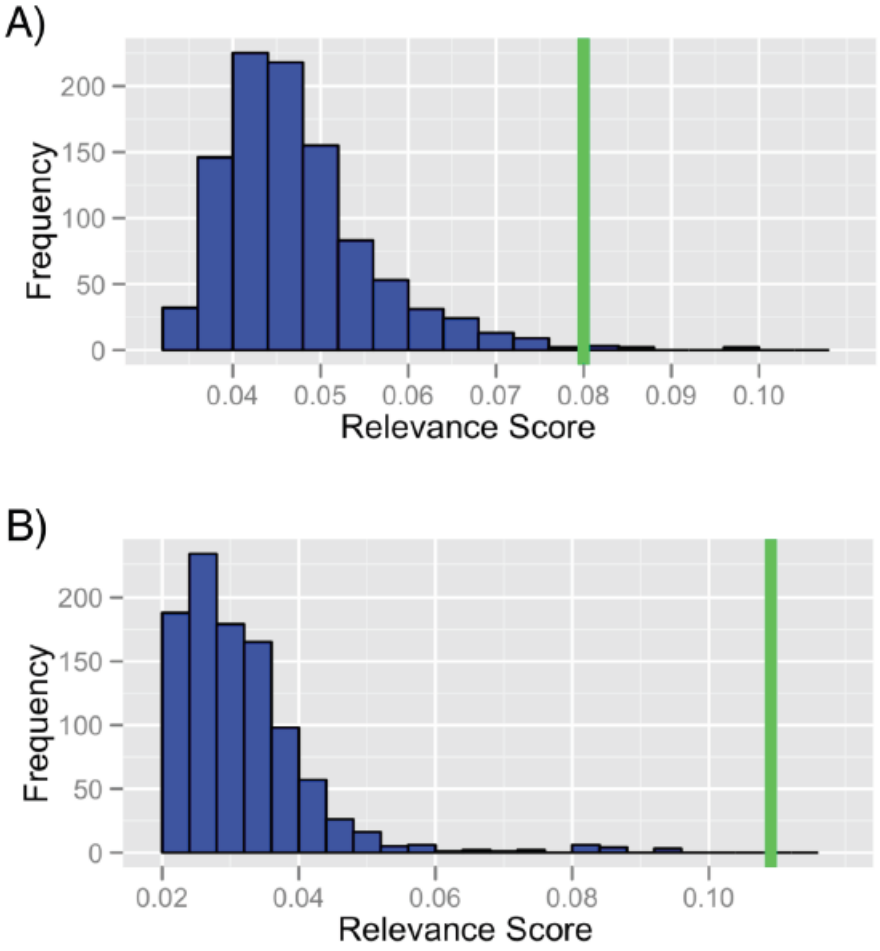

Figure 3.20: Genomic perturbations in bladder cancers are significantly associated with downstream transcriptional changes through known and novel pathway circuitry. A) The TieDIE algorithm was used to identify a network connecting the top 29 significantly mutated genes to transcriptional hubs identified from the identification of transcription factors with targets significantly up- or down-regulated in tumors relative to normal controls. These inputs sets are significantly close in pathway space, under 1000 random permutations of the input sets; blue bars are the scores of the permutations, the green line represents the score of the real network. B) The TieDIE algorithm was applied to connect 23 mutated histone-modifying genes, weighted by mutation frequency, to transcription factors with differential activity in histone-gene mutated and non-mutated samples.

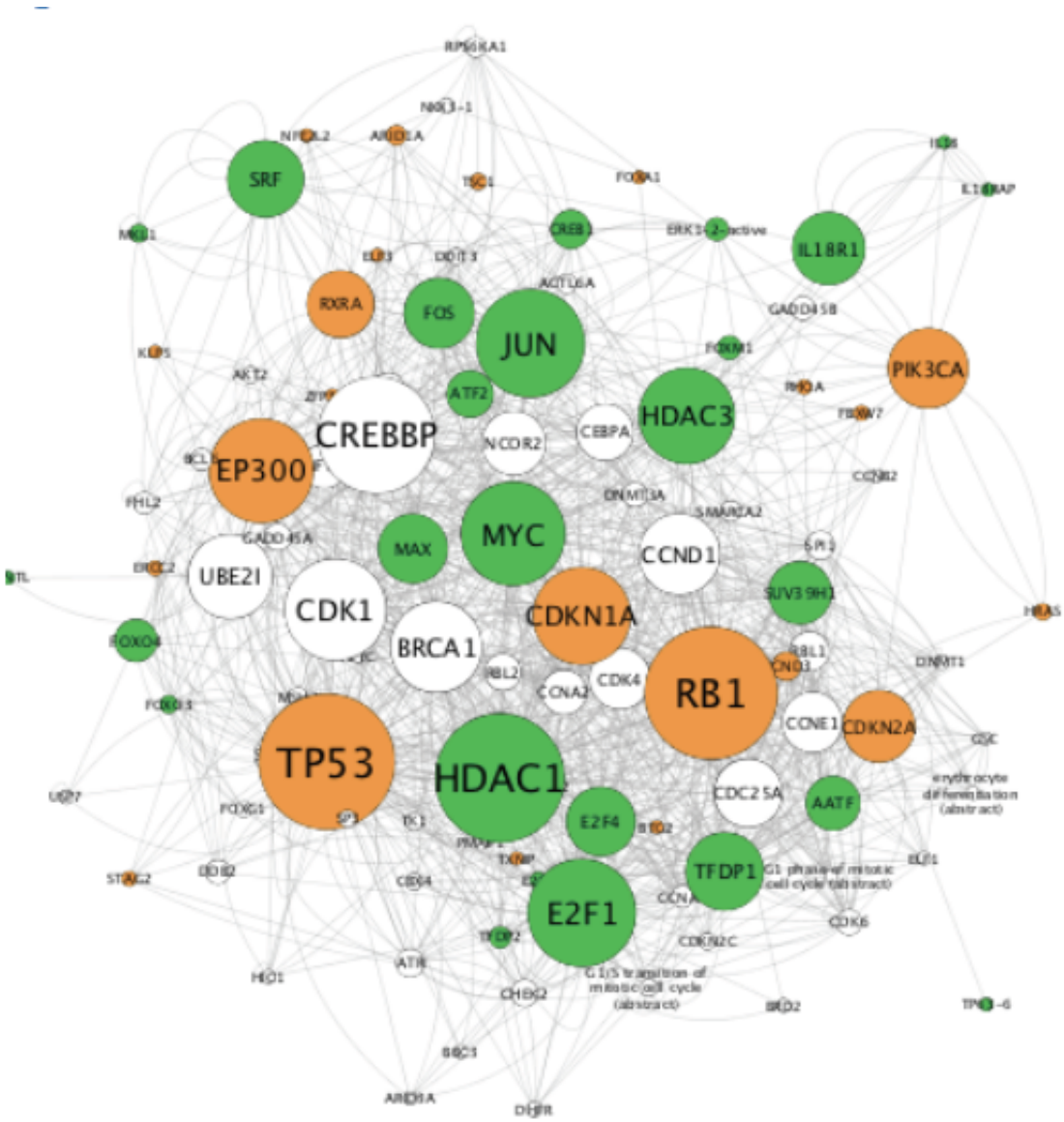

Figure 3.21: TieDIE network connecting Significantly Mutated Genes (SMGs) to transcription factors with altered activity in tumor samples. SMGs (orange) are shown as part of a network that connects these mutated genes to transcription factors (green) with altered activity. Solid lines indicate transcriptional regulation and dashed lines indicate protein regulation, and dotted lines indicate HPRD-PPI interactions or component associations. Size of the node reflects the betweeness centrality measure of the genes position in the network with larger nodes as more central to the network solution.

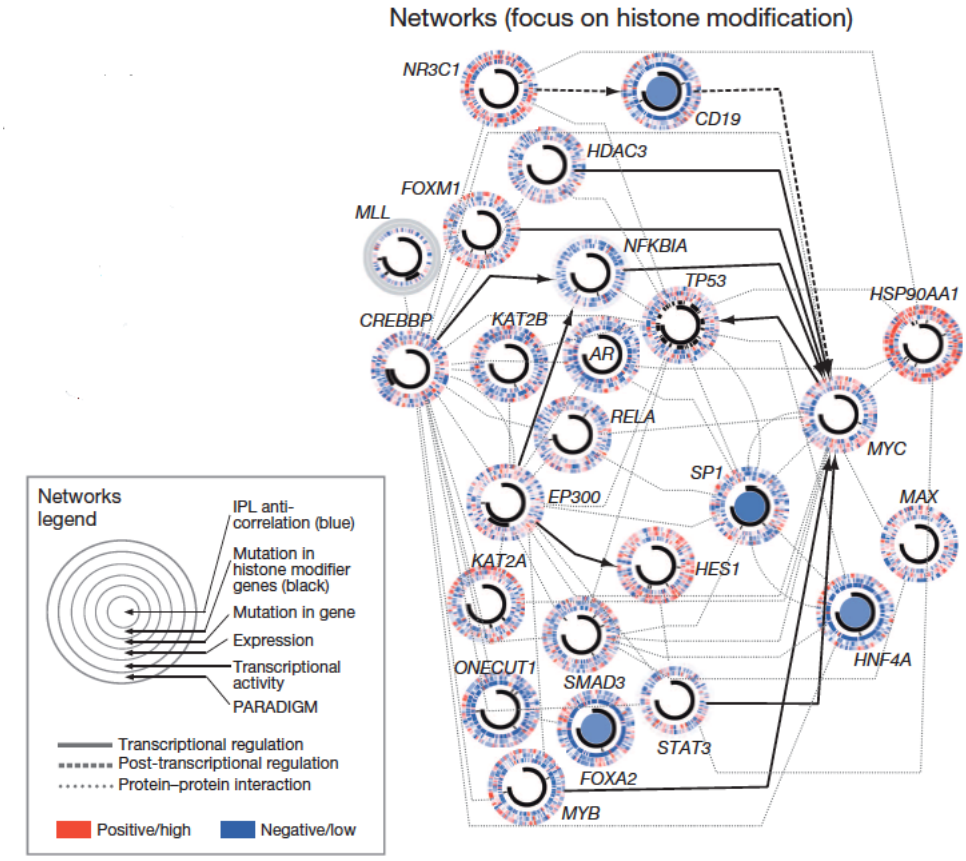

Figure 3.22: The network connecting mutated histone-modifying genes to transcription factors with differential activity. Each gene is depicted as a multi-ring circle with various levels of data, plotted such that each 'spoke in the ring represents a single patient sample (same sample ordering for all genes). 'PARADIGM ring, bioinformatically inferred levels of gene activity (red, higher activity); 'Transcriptional activity, mean mRNA levels of all of the targets of each transcription factor; 'expression, mRNA levels relative to normal (red, high); Mutation in gene, somatic mutation; 'Mutation in histone modifier genes, somatic mutation in at least one such gene; 'IPL anticorrelation, genes with PARADIGMintegrated pathway levels (IPLs) inversely correlated with histone-gene mutation status. Genegene relationships are inferred using public resources [151]

3.1.4 Discussion

The results here support the hypothesis that coding mutations in several chromatin-remodelling genes lead to a proliferative transcriptional program that underlies the tumor phenotype. The algorithm also predicts several "master regulators" that are known to modulate this transcriptional program and are likely to require the chromatin alterations that exist in the specifically mutated samples. In addition, key "linking" genes, that are not frequently mutated or modulate the expression a large number of genes, are implicated by TieDIE, including beta-catenin (CTNNB1), CREBBP, SMARCD1, SHC1 and others. In each case, the integration of multiple datasets with prior pathway knowledge is necessary to identify these genes, that lack strong cis-evidence.

Methodologically, both the standard TieDIE model (used with two input sets) and a generalized implementation that allowed for inclusion of a third set of inferred protein activities, generated with the PARADIGM algorithm [**?**, 217], were used. The latter analysis highlights the flexibility of diffusion models like TieDIE, that can be easily adapted to test more specific hypotheses as the available data is expanded.

One weakness of this anlysis is that, while validation of the findings has been done through literature validation by myself and the TCGA working groups, they currently lack any direct biological validation. One way to approach followup analysis on the set of prioritized genes would be focused on validation of differential epigenetic state in functionally mutated kidney and bladder cancer samples, and the effects of

intervention on these specific genes.

Another complication with the preceeding analysis is that the pathway interactions used as prior knowlege may be specific to specific cellular type and state, so it is difficult to conclude that each resulting network represents a coherent biological process rather than a set of unrelated biological interactions that each occur in different cellular states. This issue may be partially resolved through prediction of context-specific networks, which has beeen addressed through DREAM challenges. Specifically, I participated in the DREAM8 HPN-DREAM breast cancer network inference challenge, and as a member of the winning team for subchallenge 1A, helped to design methodology to predict signaling networks specific to each of a set of breast cancer cell lines. Other methods such as ARACNe [137] and CLR [133] have been developed to predict transcriptional specific to a given cell type, and more work is needed in this area to support methods such as TieDIE.

3.2 TieDIE identifies key kinase and scaffold proteins in BRAF and RAS mutated thyroid carcinoma samples

Papillary thyroid carcinoma (PTC) is the most common type of thyroid cancer, and is characterized by frequent activating somatic alterations of genes encoding effectors of the MAPK pathway [152], including mutations in either BRAF or RAS genes [111] , as well as fusions involving the RET [79] or NTRK1 [170] tyrosine kinases [152]. These genomic events are almost always mutually exclusive [195], suggesting similar or redundant downstream effects [152].

I participated in a TCGA study of 496 PTCs, as a pathway analysis workgroup leader, in an effort to characterize the genomic landscape and corresponding oncogenic pathway activation of these tumors. Previous analysis has observed that PTC tumors driven by BRAF-V600E mutations do not respond to negative feedback from ERK to RAF, resulting in high MAPK-signaling [171], while, conversely, tumors driven by RAS and RTK fusions signal via RAF dimers that respond to ERK feedback result in lower MAPK signaling [152]. Observing that this differential signaling results in profound differences in phenotype, the group was been able to develop a gene-expression signature to quantify the separation between BRAF and RAS mutatant samples, defining an expanded set of BRAF-like (BVL) and RAS-like (RL) samples using this signature (Supplement [152]). Ultimately, the group was able to show that these expanded BVL and RL subtypes show clear mechanistic differences in MAPK activation and corresponding downstream transcriptional activation, including mTOR activation in RL tumors asso-

ciated with ERK substrate p90RSK phosphorylation, and BCL2 overexpression [152].

TieDIE was then used to assess differential pathway activation between BVL-PTC and RL-PTC. I identified 221 samples with a non-synonymous mutation in either BRAF or one of (NRAS/HRAS/KRAS/EIF1AX) but not both, that had RPPA proteomic data available. The goal here was to find sub networks characterizing the alternate signaling activities and transcriptional effects between BRAF and RAS (NRAS, HRAS, KRAS, EIF1AX) mutants. To define the source set I used the 5 mutated BRAF/RAS gene set defined above, weighted by mutation frequency in the sample set.

For the transcriptional targets, Gene Set Enrichment Analysis (GSEA [199]) was used to identify transcription factors having targets with a non-random distribution of edgeR [180] scores, generated by a differential analysis of RNA-Seq data between the BRAF and RAS mutants. This resulted in the selection of 36 transcription factors, including JUN, FOS and TP53. A t-test [210] was performed to identify 82 genes with significantly different protemic activity between BRAF and RAS mutants at a conservative q-value cutoff of 0.05, which served as the third "signaling" set of gene inputs, weighted by the test statistic. TieDIE was run with parameters that gave equal weight to each of the three input sets, and the solution was found to be highly significant in two independent tests comparing the "source" set with the "signaling" set ($p = 0.01$) and the "signaling" and "target" sets ($p < 0.01$; $p < 0.00005$ combined), over a background network that combined NCIPID, BioCarta, Reactome, and KEGG databases (see supplement of [152] for details).

The resulting network contained 82 genes connected by 343 interactions (83 HPRD-PPI, 160 regulatory, 100 complex membership) and a depth-first search was used to connect BRAF and NRAS genes to genes in the target set with paths up to length 4, through post-transcriptional (directed) and HPRD (undirected) interactions (Figure 3.23); this filtered network contained 31 genes with 71 interactions (48 HPRD, 23 regulatory), describing alternate signaling mechanisms of BRAF and RAS mutants through key signaling genes RAF, Ras homolog enriched in brain (RHEB), MAP-Kinase proteins as well as MEK and ERK pathways.

The small GTPase RHEB, a known regulator of mTOR activity [80], was identified as the most influential factor driving differences between tumor types, through ranking of the final TieDIE "linker" heat scores, confirming the association of mTOR activating with the RL samples identified in the gene-signature analysis described above.

Notably, BRAF mutants were enriched for activity in the ERK-1-2-active protein complex, as inferred by PARADIGM, along with a corresponding decrease in TSC2 activity, an inhibition target of ERK via p90RSK signaling. This integrative analysis sheds new light on the role of p90RSK as a crucial crossroad for MAPK, mTOR, and BCL2 signaling in RAS-driven tumors, and also supported by the higher levels of p90 phosphorylation in RL tumors [152].

Additionally, KSR1 (Kinase Suppressor of RAS) activity was found to be higher in RAS mutants as determined by both RPPA A.1.5 assays and PARADIGM inferences (see Figure 3.23). KSR1 is a scaffold protein that modulates Ras/Raf-mitogen-activated protein kinase [114]; knockout of this factor has been shown to selectively

blunt oncogenic signaling in at least one Ras-driven cancer [129, 177], supporting the idea that this protein deserves further consideration as a potential therapeutic target. TieDIE was able to re-discover the role of this gene via this completely data and prior-knowledge driven analysis, thanks in part to the regulatory interactions between both RAS/BRAF/RAF1 and MAPK/MEK/ERK–supported by the genomic and transcriptomic data, respectively–as well as the RPPA dataset. This result highlights the strength of multi-modal data integration when combined with prior-knowledge, as each of these datasets and pathway links was necessary in re-discovery of this essential scaffold protein.

These findings confirm many of the known signaling changes induced by BRAF V600E and RAS mutations in PTC, provide a framework for explaining MAPK downstream activity in tumors with other driver alterations.

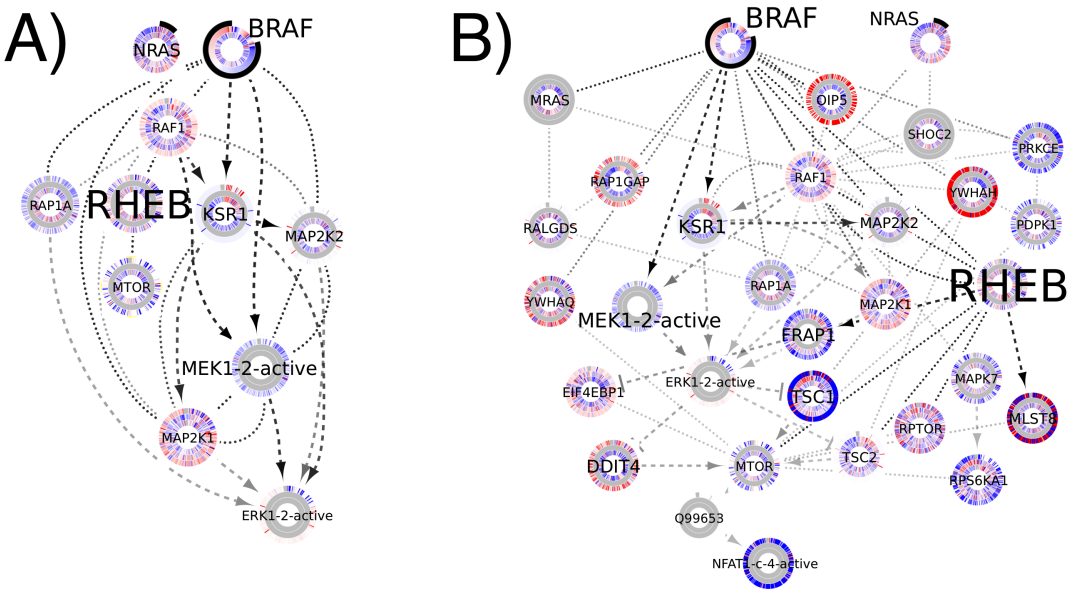

Figure 3.23: TieDIE networks connecting BRAF/RAS mutated genes to transcription factors and signaling proteins with altered activity in tumor samples. A) The "core" network of genes connecting mutant genes NRAS and BRAF to inputs generated with RNA-Seq (ERK1-2-active complex) with paths of length up to 3. Nodes correspond to proteins or complexes in the TieDIE solution; the size of the label text indicates the influence attributed to each node, in relation to the three input data sets. Solid lines indicate transcriptional regulation and dashed lines indicate protein regulation, and dotted lines indicate HPRD-PPI interactions or component associations. Each tick-mark in each circle represents a single patient sample. Circles represent genomic perturbations, PARADIGM inferred activities, RPPA and expression for proteins and complexes. Outer ring represents mutation status for BRAF and NRAS genes; for all other genes it represents the PARADIGM inferred activity levels. Second most outer ring represents RPPA activity level, and the inner ring represents the gene expression. All rings are sorted in the same order and according to the presence of a mutation in BRAF or RAS-related genes (NRAS/KRAS/HRAS/EIF1AX), and secondarily according to the activity of BRAF within mutant samples, as measured by RPPA. B) The larger TieDIE network generated by allowing all paths between BRAF/NRAS to RNA-Seq generated inputs with paths of length up to 4.

Table 3.1: Mutation frequency across pancan clusters.

mRNA Cluster	Mutation Cluster	P value	Proportion
1	12	5.5E-11	28/8
2	13	7.9E-11	52/10
2	8	5.9E-13	58/11
3	4	3.3E-56	203/32
3	9	5.8E-18	224/56
4	5	1.9E-22	76/51
5	10	2.1E-136	198/23
5	9	2.4E-12	153/56
6	1	1.9E-12	32/7
6	4	2.1E-12	72/32
11	5	4.6E-71	133/51
13	14	5.3E-14	19/6

3.3 TCGA PanCancer Dataset

To assess the hypothesis that multiple mutational and copy-number profiles cause similar transcriptional effects in a large number of tumor samples, I devised a pathway-based approach that could incorporate both prior network knowledge along with transcriptional and genomics data. To this end, I performed a Hypergeometric test with Bonferroni correction between the 16 mutation-based clusters and the 16 expression-driven clusters. I found 12 clusters with highly significant overlaps ($p < 1 * e^{-10}$), as shown in table 3.1 and, in particular, found that mutation clusters 8 and 13 each overlap highly with mRNA cluster 2. Roughly half of the samples in each of these mutation clusters intersect with the mRNA cluster 2 that consists of mostly squamous cell HNSC and LUSC tumor samples along with some BLCA and LUAD samples.

I asked whether the genomic perturbations in each of the overlapping mutation subgroups were significantly associated with the transcriptional signal associated

with the squamous mRNA cluster. To accomplish this, I used TieDIE to search for significant interconnections between genomic perturbations and downstream transcriptional changes. For this analysis, I used genes with mutations in at least 2% of the intersecting set of samples (in both mutation cluster 8 and mRNA cluster 2) as the source set, and used the combined frequency of mutations and copy-number changes to define the weights of these 75 resulting genes. For the targets, I used Gene Set Enrichment Analysis (GSEA) [199] to identify identify 42 transcription factors having targets with a non-random distribution of edgeR [180] significance scores, based on a tumor vs. normal comparison with all the samples in the second mRNA cluster. TieDIE was run on these input sets, and produced a significant network ($p < 0.043$) according to the permutation-based background model, used to determine whether the inputs sets are significantly close in pathway space. The resulting network contained 215 genes connected by 2,531 interactions (774 HPRD-PPI, 644 regulatory, 1113 component). Similarly, a source input set consisting of 80 genes was generated for the intersection of mutation cluster 13 and mRNA cluster 2 and TieDIE was run as before, this time producing a highly significant network consisting of 219 genes, connected by 2733 interactions (784 HPRD-PPI, 593 regulatory, 1356 component; p <0.006, Figure 3.24).

To visualize the result in Figure 3.26, a tag cloud (word cloud) representation was used with free online software from http://www.wordle.net/.

Pathway Input. As input to both the TieDIE and GSEA analyses, pathways were obtained in BioPax Level 3 format, and included the NCIPID and BioCarta databases from http://pid.nci.nih.gov, the Reactome database from http://reactome.org,

and the set of signaling and metabolic pathways in the last public release of the KEGG database. Gene identifiers were unified by UniProt ID then converted to Human Genome Nomenclature Committees HUGO symbol using mappings provided by HGNC (http://www.genenames.org/). Interactions from all of these sources were then combined into a merged Superimposed Pathway (SuperPathway).

Genes, complexes, and abstract processes (e.g. "cell cycle" and "apoptosis") were retained and referred to collectively as pathway features. Before merging gene features, all gene identifiers were translated into HUGO standard identifiers. A breadth-first traversal starting from the feature with the highest number of interactions was performed to build one single component. The resulting pathway structure contained a total of 20,713 concepts, representing 7706 proteins, 8998 complexes, 1700 families, 55 RNAs, 15 miRNAs and 582 processes. In addition, 39,240 protein-protein interactions from the HPRD database were acquired from http://www.pathwaycommons.org/ and included in the input network used for TieDIE.

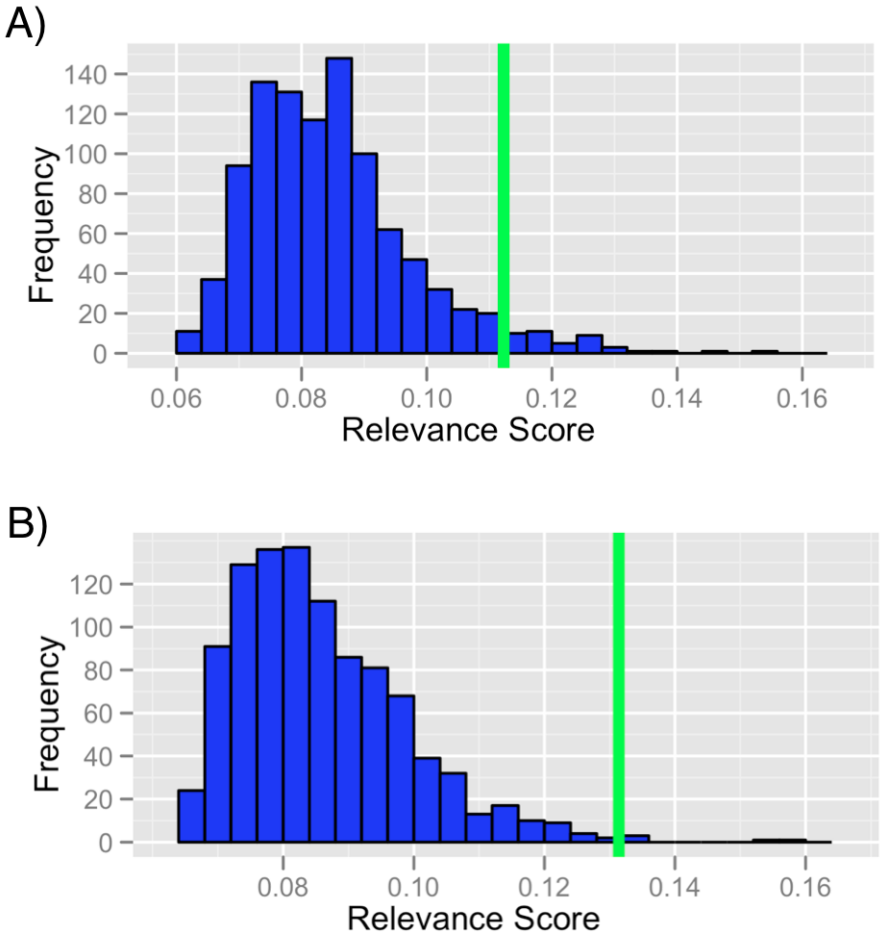

Figure 3.24: Genomic perturbations in mutation clusters 8 and 13 are significantly associated with downstream transcriptional changes in the squamous expression subtype, through known and novel pathway circuitry. A) The TieDIE algorithm was used to identify a network connecting 75 mutated genes within mutation cluster 8 to transcriptional hubs identified from the identification of transcription factors with targets significantly up- or down-regulated in tumors relative to normal controls, within the squamous mRNA cluster 2. These inputs sets are significantly close in pathway space, under 1000 random permutations of the input sets; blue bars are the scores of the permutations, the green line represents the score of the real network. B) The TieDIE algorithm was applied to connect 80 altered genes in the mutation cluster 13 to the same set of transcriptional hubs.

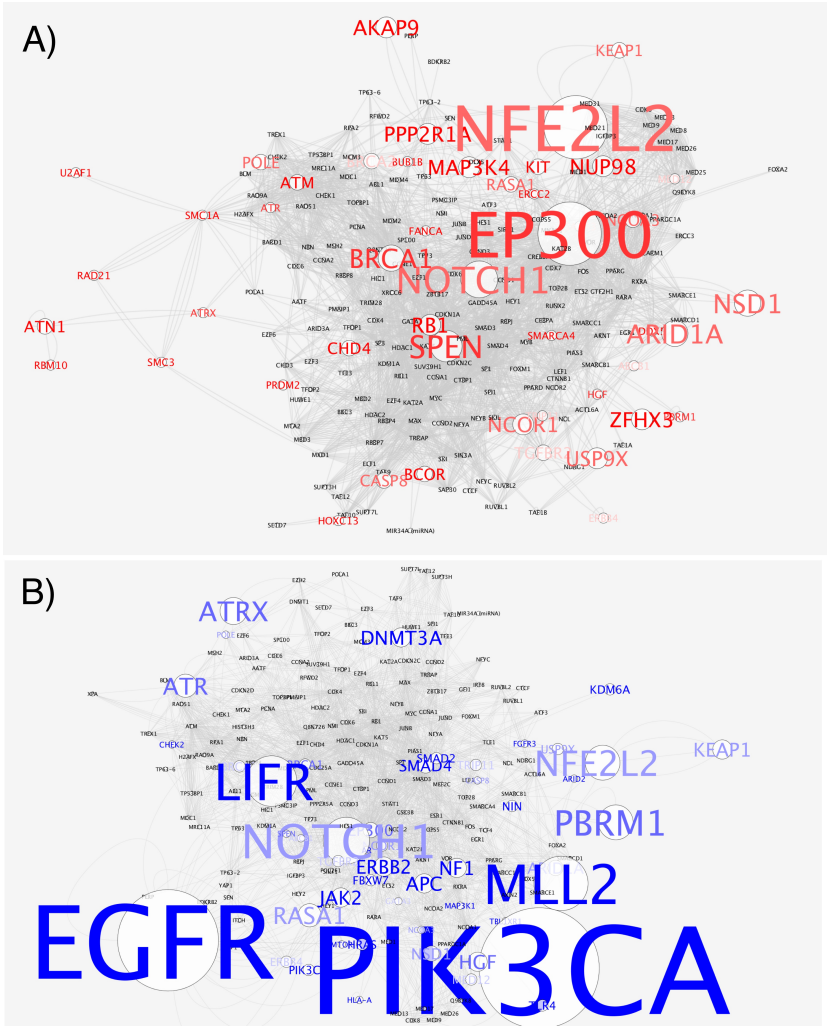

Figure 3.25: TieDIE networks connect genomic perturbations in mutation clusters 8 and 13 connect to downstream transcriptional changes in the "squamous' expression subtype. A) The TieDIE algorithm was used to identify a connecting network between 75 perturbed genes within mutation cluster 8 to transcriptional hubs identified from the identification of transcription factors with targets significantly up- or down-regulated in tumors relative to normal controls, within the "squamous" mRNA cluster 2. Node and label font size represents the genomic alteration frequency within the sample group; brighter shades of red represent genomic alterations more exclusive to the mutation cluster 8, as compared with mutation cluster 13. Genes commonly mutated in both networks are removed (TP53, CDKN2A) for contrast. B) The TieDIE algorithm was applied to connect 80 altered genes in the mutation cluster 13 to the same set of transcriptional hubs. Brighter shades of blue indicate genomic alterations more exclusive to mutation cluster 13, as compared with mutation cluster 8.

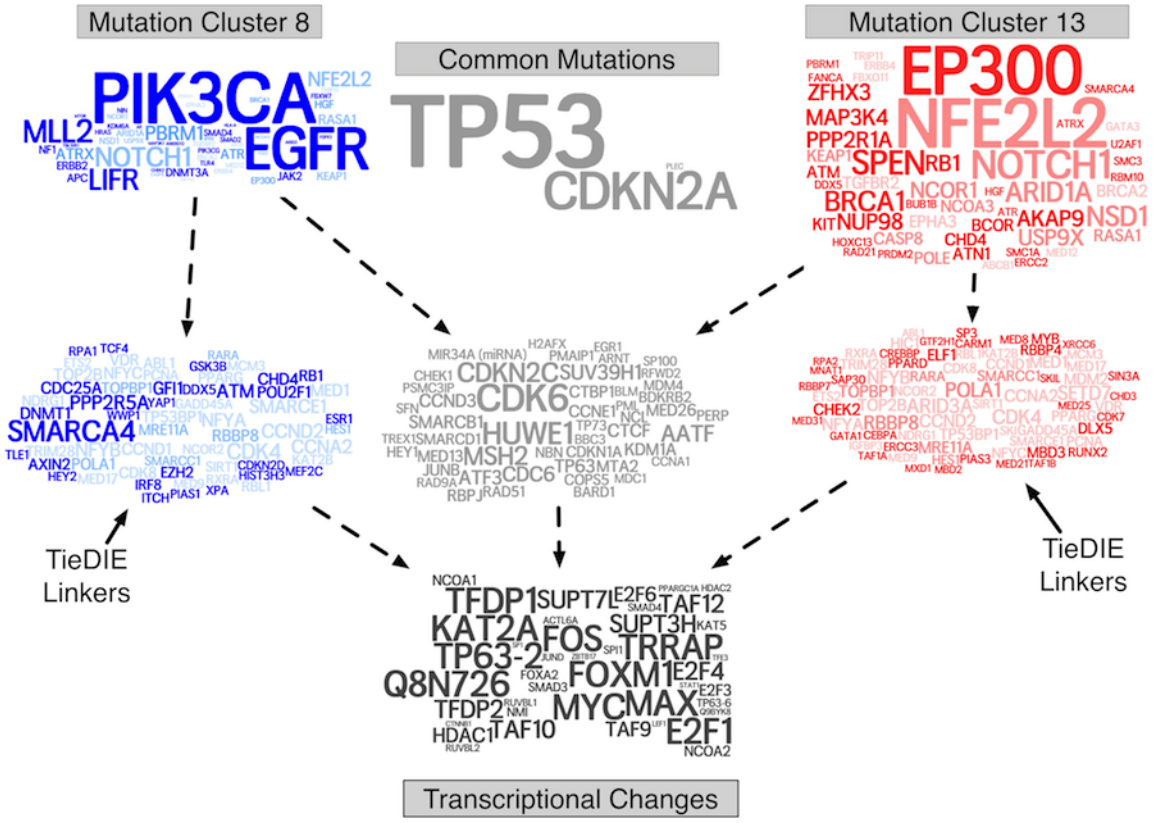

Figure 3.26: To visualize the result in Figure 3.26, a tag cloud (word cloud) representation was used with free online software from http://www.wordle.net/. Font sizes were scaled linearly either by frequency of genomic alteration (top left, top right clouds), linker gene importance from the TieDIE algorithm (middle left, middle right clouds), and by the significance of GSEA scores for each transcriptional hub (bottom cloud). Font color gradients represent differential frequency of genomic perturbation (top left, right clouds) and differential linker gene importance (middle left, right), with bright red representing a gene fully exclusive to mutation cluster 13, or a linker gene in the corresponding TieDIE network that is not found in the mutation cluster 8 TieDIE network. Similarly, bright blue words represent genes fully exclusive to mutation cluster 8 and its corresponding TieDIE network

.

3.4 TieDIE Cytoscape Implementation

In the summer of 2014 I served as a Google Summer of Code (GSoC) mentor, as part of the National Resource for Network Biology (NRNB; *http://nrnb.org/gsoc.html*) academy. I had the pleasure of working with Srikanth Bezawada, a student at the Birla Institute of Technology and Sciences, who implemented the TieDIE algorithm as a Cytoscape [188] application in the Java programming language, under my guidance. The application is currently available from the Cytoscape app store (*http://apps.cytoscape.org/*) as a Beta release.

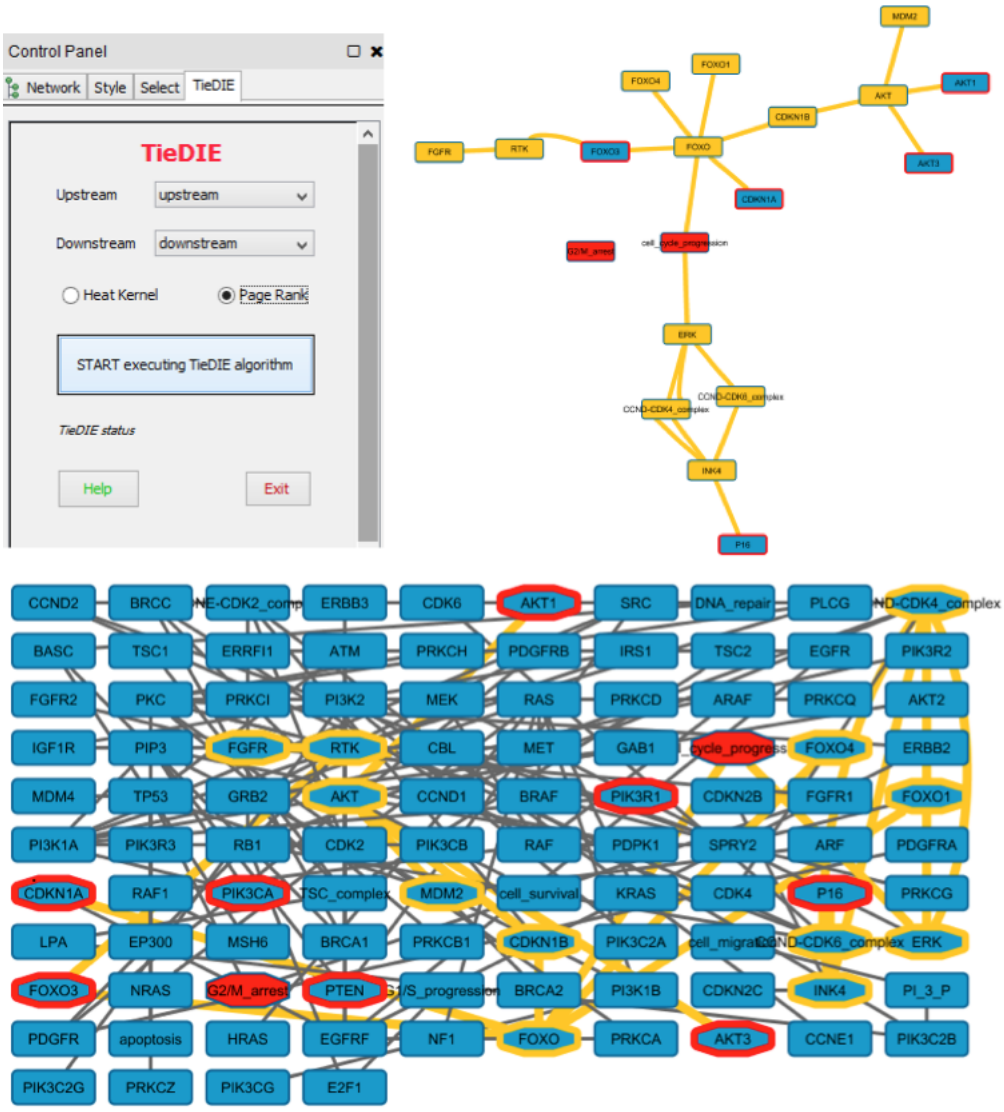

Figure 3.27: TieDIE is available as a plugin application to the Cytoscape software [188]; developed in collaboration with Java software engineer Srikanth Bezawada and the 2014 Google Summer of Code (GSoC). Top left: the TieDIE control panel allows users to select input data from tables loaded into the Cytoscape software; radio buttons allow the user to select either the heat-diffusion kernel or personalized pagerank methods for information diffusion over their selected network. The rectangular button below starts the TieDIE algorithm as an a-synchronous background process, producing a sub network (top right) on completion, as well as highlighting the source (red outline), linker (yellow outline) and target (red center) nodes, as shown in the bottom figure.

Patient-Specific Network Analysis and Applications

4.1 Introduction

A major goal in cancer systems biology is to infer a specific network for each patient's tumor, and, as data becomes available, each subclone identified within the tumor. Accurate network models could be used to explore a large space of potential targets to kill the tumor *in silico*, while new experimental technologies such as Patient Derived Xenografts (PDX) [204] will make testing a subset of these computational predictions possible.

In contrast to the cohort-wide analyses that find a summary of common biological changes across a cohort of tumors with similar tissue of origin, biological markers or subtype, patient-specific analyses attempt to find specific biological changes that drive the disease of a single patient. In (rough) statistical terms, cohort-level analyses capture a "mean" view of relevant biology of tumors, often related to genomic or transcriptional events that occur in a meaningful fraction of patients, while patient-level analysis is

necessary to summarize the variance (or heterogenaity) between patients. This heterogeneity is widely believed to be present in most cancer types [85], and the need for patient-specific analyses is particularly necessary for generating clinincal utility as most the therapeutic strategies target specific proteins or molecular processes [85].

Finding these actionable genes in each patient, however, provides a challenge due to the lack of statistical power when doing an $N = 1$ analysis, making it difficult to find robust predictions for each patient. Any reasonable approach to $N = 1$ analysis, therefore, will attempt to use data on additional cancer samples to inform predictions made for a specific patient. One of the simplest examples of this strategy is the use of the COSMIC [3] database, which can help determine if a single nucleotide variant (SNV) observed in a given patient is recurrent across other cancer samples, and therefore more likely to be an evolutionary driver of the tumor in question. Similarly, the driver potential of some mutational variants observed in new patient samples can now be interpreted thanks to large-scale studies by the TCGA in cancers of various tissues of origin [150, 151, 153].

Proteins and signaling components that are not directly altered, but essential to the growth and survival of a given tumor, can pose an ever greater challenge in $N = 1$ analysis than assessment of genomic drivers. For instance, as proposed earlier in this document 3.1, alterations in certain histone-modifying genes can lead to the corruption of SWI/SNF complex machinery and wide-ranging epigenomic and transcriptional changes in genomically unmodified genes. In addition, multiple genomic alterations upstream of the mitogen-activated protein kinase (MAPK) [56] and AKT [104] pahtways

can lead to the constitutive activation of these growth-signaling pathways, which are themselves genomically stable. Pathway analysis is a powerful tool in predicting activation of this tumor-essential biology as it provides a framework to compose various types of high-throughput data (genomic, transcriptomic, proteomic and others) into predicted pathway activation states, generating large-scale datasets–analygous to the COSMIC and TCGA examples for SNVs–that can be used for interpretation of a single patient's pathway-transformed data.

4.2 Tissue-Specific network inference in cell-line model systems

4.2.1 Inference of gene regulatory networks with phosphoproteomic data

A major goal of systems biology is to infer the genetic "circuitry" that governs how cells respond to environmental stimuli, developmental cues, and therapeutic interventions. The idea is to find a genetic pathway model that can accurately predict the consequences of perturbations not seen during model construction [55]. Much work has focused around the reverse engineering of gene regulatory networks (GRNs) from high-throughput datasets of various manifestations. The DREAM series of challenges was originally launched to spur the development of methods to infer GRNs and has found that an ensemble or "wisdom of crowds" approach that combines several strategies often performs better than any stand-alone approach [55]. Consequently, the accuracy

of the top-performing ensembles reveal that there still exists a considerable room for improvement in the ability of individual methods to reverse-engineer GRNs. Thus, new methodology is needed, one that may draw inspiration from approaches developed in different fields of research and that could complement the current set of methods [55].

In collaboration with a small team of graduate students at UC Santa Cruz, I participated in the DREAM 8 breast cancer network inference challenge, in an effort to predict tumor-specific protein signaling pathways. A panel of breast cancer cell lines were used as a model for prediction in this challenge, in the hope that methods learned here would generalize to patient samples. Our winning solution to the HPN-DREAM8 1A problem combines the principle of correlation suggesting causation with prior biological knowledge to produce a robust prediction of causation from time series phosphoproteomics data.

This novel method, Prophetic Granger Causality (PGC), for inferring gene regulatory networks (GRNs) infers networks from time series data measuring protein levels as a function of time, stimuli, and other perturbations or conditions [55]. The method uses an L1-penalized regression framework to adapt a Granger Causality approach for use in GRN construction. When combined with a data-independent network prior, the method outperformed all other methods submitted to the HPN-DREAM 8 network inference challenge 1A sub-challenge. Our investigations reveal that PGC on its own is a weak predictor but provides complementary information to other approaches, boosting ensemble learners. Thus, PGC serves as a valuable new tool in the bioinformatics toolkit for analyzing temporal datasets [55].

Derivation of a network prior

Our hypothesis was that a network prior would aid in the interpretation of a dataset like the HPN phosphoproteomics data, if enough of the protein-protein interactions are relevant to the condition(s) represented in a study. For this challenge we used established pathway collections, since their protein interactions tend to fit most closely with a notion of causal influence. Because there are biological feedback mechanisms that are not accounted for, perturbations in a child species also often lead to perturbations in parents even in the absence of a documented mechanism, making it difficult to assess the directionality of causality, as was required in this competition. Therefore, we decided to ignore the directionality of pathway diagrams, and merely seek some metric of "closeness in the graph that represents the pathway, choosing heat diffusion as the metric, the merits of which have been extensively discussed in this thesis [55].

The biological prior was derived from the Pathway Commons database version 3 [33]. Of these, 495 pathways describe previously established canonical biological processes with HGNC symbols for genes. As a first guess at causality, without even considering the phosphoproteome data, we used the pathways to describe how "close the proteins were in biochemical processes, regardless of the types of interactions that linked them. We did this by treating each pathway that contained two or more of the proteins in the chip (263 pathways) in the following manner: first, each pathway was reduced to a simple, undirected graph where each node was a gene and each edge was an interaction. For proteins that act in a complex, each of the constituents of the com-

plex was assigned an edge to the targets of the complex. On this graph, I calculated a heat diffusion kernel for the pathway. If L_p is the graph Laplacian, then the heat kernel corresponding to a 0.1 time unit diffusion, as suggested by [172] was calculated as: $H_p = exp(-L_p * 0.1)$, where $exp()$ is the matrix exponential, and identically to the computation used with the TieDIE algorithm. This step resulted in a symmetric matrix of pairwise edges between some of the nodes in the training data. The final prior network was then obtained by aggregating all of these symmetric matrices across all of the Pathway Commons pathways considered [55, 197].

The heat kernel prior uses no information about the training data in its calculation; however, this matrix alone achieved second place in the DREAM challenge 1A (see figure 4.29. The performance of this entry by itself in the DREAM challenge can be seen as a confirmation that the contest recapitulates known biology, and is an important reminder that prior biological knowledge should always be taken into account in practical applications of machine learning to biology [55, 197].

Regression framework and results

The DREAM8-1A sub-challenge provided contestants with a time series of phosphoproteomics data on four different breast cancer cell lines, subjected to eight different stimuli, in the presence of four different inhibitor conditions. The goal of the HPN sub-challenge was to infer the directed protein-protein signaling network governing the response of 45 phosphoproteins for each of the cell lines. The correctness of the inferred networks was assessed by follow-up wet lab experiments in which the stimuli

were presented to the cells in the presence of a new inhibitor, and the resulting responses of each protein in the network were recorded; this allowed verification of phosphoproteins predicted to be differentially regulated under the new inhibitor and, therefore, downstream of the inhibitors target [55].

A regression-based analysis was performed, extending L1-penalized Granger causality [20] to consider future time points in addition to past time points (see upcoming DREAM8 project paper [197]). This strategy provided the regression model with additional training data to assess the correlation structure of the data, while still allowing resolution of the directionality. Treating each (cellline, stimulus) context independently, each value of the phosphoprotein data for a particular inhibitor, time point, and phosphoprotein was approximated as a linear combination of all other (time point, phosphoprotein) pairs for the same inhibitor, including future time points. An L1-penalty was applied to the regression and tuned such that all regression coefficients associated with past and future instances of the regressed node (i.e. autoregression terms) were set to zero. Any non-zero coefficients among the exogenous terms were then assumed to contain causal information. Regressor nodes from time points prior to, or concurrent with, the regressed node contributed evidence in favor of the regressor node having a causal influence on the regressed node. Regressor nodes from time points after the regressed node contributed evidence in favor of the regressed node having a causal influence on the regressor node. Finally, the inferred network matrices were normalized to a [0,1] range and averaged with the prior network to produce a final causal matrix [55, 197]; an overview of our approach is shown in figure 4.28.

As can be seen in figure 4.29, the causal model combined with the prior network was the best overall performer in the challenge, slightly (but signficantly) improving on the prior network alone [197].

Discussion

The DREAM competitions are an invaluable way to assess the relative utility of a large number of methodologies and datasets, and provide rapid feedback to the scientific community. In the DREAM8 network inference challenge, we found that the use of prior knowledge is essential for the determination of tumor-specific protein networks. As with the many TieDIE algorithm results presented in this thesis, we demonstrated the ability of a heat-diffusion model to effectively integrate prior data with tumor-specific datasets.

Further analysis (presented in detail in our methods paper in *Nature Scientific Reports*) using our regression framework found evidence that cell-type, and the corresponding genomic background, is a greater determinant of cell-specific network wiring than stimulus-dependent effects [55]. This, along with evidence from other recent studies [66, 94] supports the idea that tissue-specific networks are needed to perform more complex computational analyses. I test this approach with a multi-omic dataset derived from metastatic prostate tumors later in this chapter.

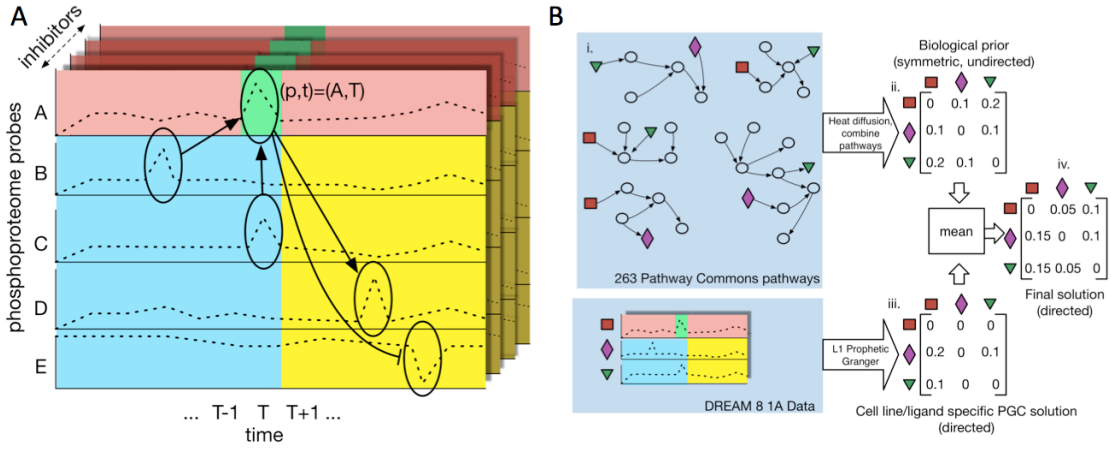

Figure 4.28: a). Prophetic Granger Causality method description taken from the team paper [55]. The method is given a set of probes (rows; y-axis) measuring the level of a particular phospho-protein state at particular time points (columns; x-axis). Each probe value p at each time point t (green) is considered in turn as a linear regression of all other feature times and probes. Probe A is being considered at time T. The penalty parameter L1 is chosen such that autoregression contributions (red) are set to zero. Any remaining non-zero regression coefficients for other probes suggest causality; past or concurrent time point probes (blue) are considered causal of the target; future time point probes (yellow) are considered to be caused by the target. The different inhibitor conditions are treated as different examples in the regression task. This process was repeated for each time and probe, with each regression task contributing to the final connectivity matrix [55]. b) Overview of the overall PGC plus network prior approach for the HPN DREAM8 submission [55]. Prediction for a single (cell line, ligand) pair task. (i.) 263 Pathway Commons pathways having at least two proteins in the DREAM dataset (colored shapes). (ii.) Heat diffusion kernel used to measure closeness between the proteins in each individual pathway. Pathways were combined into a single weighted biological prior adjacency matrix. (iii.) The Prophetic Granger solution, obtained as shown in part A. (iv.) The final solution for the (cell line, ligand) pair produced by averaging the heat diffusion kernel with the absolute value of the Prophetic Granger solution [55].

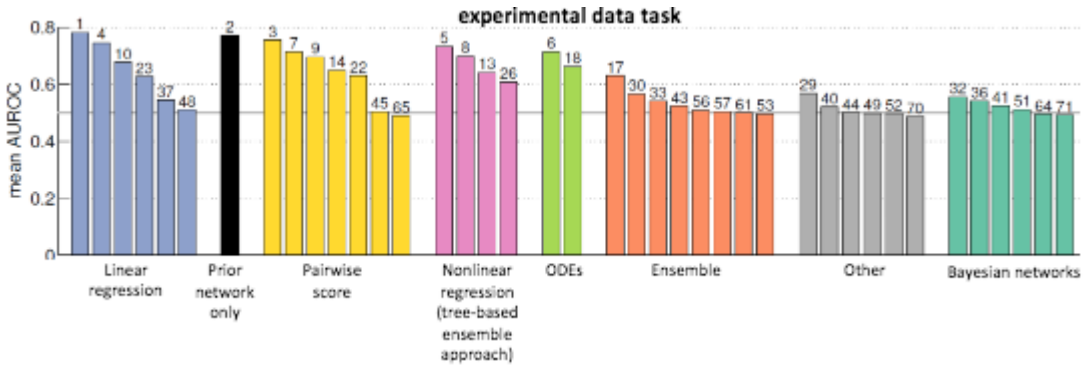

Figure 4.29: Summary of subchallenge 1A result for all models submitted to the HPN-DREAM8 network inference challenge. The prior network (2) scored higher than all other submitted models; combining with the Prophetic Granger Causality (PGC) regression predictions further increased the performance (1, left) [197].

4.2.2 Gene-Essentiality Prediction in cell-line model systems

As described in the introduction to this thesis, cancer often relies heavily on unaltered components of cellular biology to sustain agressive growth; this results in tumor-specific vulnerabilities that can be exploited with target drug therapies [39]. Prediction of these cancer-essential genes is an important dual to the problem of sub-network prediction, as reliance on specific gene interactions (as explored in the the previous section) can inform gene-essentiality prediction, and vice versa.

In the summer, 2014, I lead a team at UC Santa Cruz to compete in the Broad-DREAM Gene Essentiality Prediction Challenge. This challenge represented a community effort to assess the potential of a large-scale genomic screening effort (project Achilles) to link gene dependencies to the molecular characteristics of each cancer, in order to identify molecular targets and guide therapeutic development [39] (www.synapse.org). The Achilles project, developed by the Broad Institute, uses genome wide genetic perturbation reagents (shRNA) to selectively silence genes and assess the effect on survival of each (www.broadinstitute.org/achilles). Our goal was to predict the specific dependence of each gene in a panel of cell lines from the Cancer Cell Line Encyclopedia (CCLE), producing a ranking of cell lines from most to least sensitive. Our leading hypothesis was that prior information in the form of pathway databases and literature review could improve our ability to predict specific essentiality of many key genes; by the end of the competition, we had developed methodology to employ this prior knowledge to our prediction task, and were able to win one of the key subchallenges

in the DREAM competition.

4.2.2.1 Methodology

Data Filtering Strategy Given that we were asked to predict a relatively small number of samples (150) with an extremely large number of features (over 40,000), feature selection was a large factor in the competition, both for the accuracy and interpretability of the results. Because cancer-related genes are more likely to be driving the underlying biology of tumor cells, I hypothesized that prior-knowledge of these genes will improve our ability to pick robust biomarkers; in particular, data and analyses performed on the TCGA Pan-cancer dataset can filter the number of genomic features to those related to "driver" events. In addition, the TCGA Pan-cancer datasets includes thousands of unique samples, meaning the application of these data vastly improves the amount of information available to select a robust set of features. I implemented this concept by filtering CCLE copy-number training and test data include only the 102 genes found to have significant copy-number alterations in the TCGA Pan-cancer data set [25], and found that it provided a significant boost in performance on throughout all (3) stages of the competition. Therefore, we included only this subset of copy-number features in all of our final submissions.

In addition, a "baseline" filter was constructed by selecting the 25% of expression probes with the highest variance in the combined training and test datasets, leaving 4740 probes. After this we filtered out any probes with significantly different distributions in the training and test data (two-tailed t-test, $p < 0.05$), as these probes

were less likely to generalize between the two data partitions; 4561 probes were left after this step.

Prediction Framework To introduce prior knowledge into our prediction task, we used MSigDB [200] c2 (curated gene sets) and c5(GO gene sets) to construct a list of pathways to be used with a Multiple Kernel Learning (MKL) framework. Starting with 6176 gene groups, we first filtered using our baseline filter (above), and then through a heuristic to remove redundant gene sets: gene sets were ranked according to size, and starting with the largest, every subsequent gene set was only included if it did not overlap more than 65% with any previously included gene sets. Following this second filtering step, we ended up with 2804 pathways that were used to construct expression kernels.

Multiple Kernel Learning is a kernel-based machine learning method which looks for an optimal linear combination of (given) kernels that gives the best learner for a partiular task [196]. These kernels can represent heterogeneous data sources or different biological perspective of the same data: in our case, each kernel corresponded to one of the 2804 expression-based and 2338 copy-number based gene sets that passed the pathway filtering step previously described. Our hope was that, by using these kernels to learn activation or de-activation states of clusters of related genes, we would obtain more robust (and more interpretable) features than a single-gene predictor could provide. Vladislav Uzunangelov, a Ph.D. student colleage on our team used Tomioka and Suzuki's regularized MKL elastic net formulation [207] to implement this part of

our entry.

$$\min \sum_{i=1}^{N} (y_i - \sum_{m=1}^{M} K_m(x; x_i)\alpha_m - \beta)^2 + C_1 \sum_{m=1}^{M} |\alpha_m| + C_2/2 \sum_{m=1}^{M} \alpha_m^2$$

where N is the number of samples, M is the number of kernels with non-zero weights, and (α) are the individual kernel weights. We solved this with Tomioka and Suzuki's MATLAB implementation.

We constructed 2804 expression-based and 2338 copy -number based Gaussian kernels by limiting the feature space to only the expression/CN features that belonged to the given gene set. Before computing the kernels, we standardized the distribution of each feature across the training set to a mean of 0 and variance of 1. Furthermore, once the kernels were calculated, they were all re-scaled to have 1's along the main diagonal, which is equivalent to ensuring that each vectors lies on the unit sphere in the (transformed) feature space.

I implemented a Random Forest predictor to predict the essentiality of each gene i, the features for all cell lines were combined into a single data matrix that was then used as input to the randomForest package. The regression problem for each target gene i was then reduced to minimizing the squared error loss over all cell lines l and features j:

$$\sum_{j,l} (\delta_l^i - f_i(x_{(j,l)}))^2)$$

where (δ_l^i) is the gene essentiality for gene i and cell line l, and ($f_i(y)$) is the function of the feature data to be optimized by our regression model.

Regression trees approximately solve this problem by recursively splitting the learning sample with binary tests, each based on one input variable, trying to reduce the variance of the output variable/essentiality score across the cell lines [123]. At each node/split, only a subset of the input features are sampled and considered as a candidate variable for this split; we set this parameter to the base-2 logarithm of the number of the total number of features, plus one, as it was determined to provide the best performance in a related DREAM prediction task [50]. Votes from an ensemble of 1000 trees, each built from a bootstrapped sample of the data, is then averaged to produce the final predictions.

4.2.2.2 Conclusion

Our DREAM9 team demonstrated the effectiveness of using the additional data of a large sequencing initiative (TCGA Pancancer) as well as pathway data, to improve the prediction of gene essentiality in cell line models. The methods we developed use this additional data in combination with the competition-provided expression and copy-number data to select a more robust set of predictive features for each task (gene). Specifically, an analysis of a Pan-cancer dataset consisting of thousands of diverse tumor samples was able to identify a small set of "driver" genomic features [94] (102 genes) that consistently improved predictive accuracy on the phase 1 and phase 2 leaderboards, when compared with the entire (19,000) gene set.

In addition to the feature-selection methods, we leveraged multiple non-linear (random forest, my contribution; multiple kernel learning, implemented by Ph.D. student Vladislaw Uzunangelov) predictors, allowing us to create ensemble predictions that increased the performance on the first subchallenge.

4.3 Patient-Specific Networks in Metastatic Prostate Cancer

Incorporation of phosphoproteomic data sets has not reached the standardized level to what is routine with whole genome or RNA sequencing data. Typically, these approaches have defined algorithms that analyze the data in several ways to depict key transcriptional targets, cell surface molecules, or pathways [13,19,78,148,179,202]). This has led to the current state of personalized medicine using genomics and transcriptomics to identify targetable mutations or pathways that can differentiate tumors from a similar tissue of origin for selective therapies [13]. However, one step that is missing from genomic or transcriptomic analyses is further confirmation or validation of the activated pathways that have been found with these data; Examples include confirmation of known driver mutations such as EGFR in lung adenocarcinoma and c-KIT in gastrointestinal stromal tumors (GIST) or the identification of signaling pathways from tumors lacking consistent driver mutations such as prostate cancer, where the functional understanding of how these pathways drive tumor growth is essential [13].

Protein phosphorylation is an essential, rate limiting step for the regulation

of signaling pathways over numerous biological events; determining both the level of phosphorylation and what residues are phosphorylated on a given protein may provide information about the activity of kinases and phosphatases as well as uncover new functional information [13]. Cellular signaling can also be controlled through the recruitment of protein domains (such as SH2 and SH3) to specific phosphorylation sites on the cytosolic tail of kinases [166]. Protein phosphorylation leads to a cascade of downstream signaling events important for cell maintenance and survival, and dysregulation of this process has been implicated in many diseases including cancer [100]. It follows that the implementation of phosphoproteomics, coupled with traditional mRNA-based approaches, may provide greater clues to these signaling events than either alone [13].

To discover treatment-relevant patterns of phosphoproteomic signaling in CRPC, a team lead by Justin Drake and Owen Witte (UCLA) first developed a complete and extensive dataset of the phosphoproteome in metastatic CRPC by performing an extensive analysis of tyrosine, phosphoserine and phosphothreonine peptides [63]. In collaboration with them, I adapted the TieDIE and Master Regulator Analysis (MRA; see appendix A.3) presented in earlier chapters of this thesis, to develop a computational pipeline that could integrate this new protein signaling information with existing "omics" datasets. This approach was motivated by recent attempts at multi-omic data integration of proteomic and genomic data, that have revealed (in other cancer types) pathways that can be targeted [21, 28, 227].

With TieDIE, we developed comprehensive pathway networks that are both enriched and activated in CRPC, integrating independent datasets of the transcriptome,

mutational data, and the phosphoproteome from a partially overlapping set of tumor samples obtained at rapid autopsy [182]. For six patients (seven samples) that had both mRNA and phosphoproteome data, I computed patient-specific pathways that mark key signaling events within the pathway for interrogation, by extending the methods for single-patient analysis developed for TieDIE and demonstrated on TCGA breast cancer data, earlier in this thesis.

4.3.1 Results

Generation of a comprehensive phosphoproemic dataset

Metastatic CRPC was analyzed at the phosphoproteome level by the team at UCLA using strong cation exchange (SCX) chromatography (phosphotyrosine (pY) peptides, [63] and phosphoserine (pS) and phosphothreonine (pT) peptides [13]); in addition to the tyrosine peptides, CRPC tumors were evaluated for pST peptides using quantitative label free mass spectrometry (Figure). This analysis was able to identify 8,051 total pST peptides from 72 total runs corresponding to 36 samples of interest, resulting in over a 25-fold increase of pST to pY identifications due to the stoichiometry of serine/threonine phosphorylation (98%) relative to tyrosine phosphorylation (2%) [160].

Similar hierarchical clustering patterns were apparent when compared to previously published pY peptide data [63] as cell lines were distinct from primary tissues and treatment nave localized prostate cancer clustered independently from metastatic CRPC tissues (Figure 4.30).

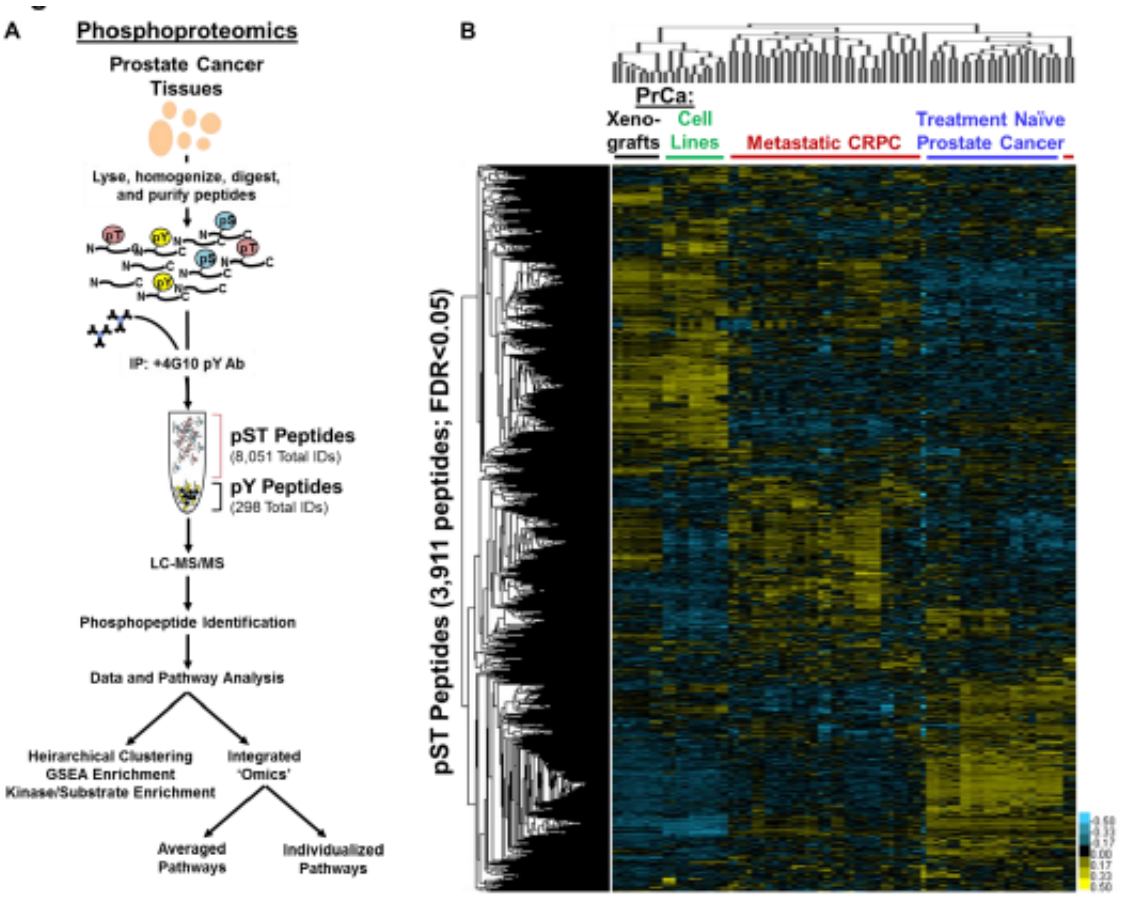

Figure 4.30: Experimental setup for prostate cancer phosphoproteomic data collection, designed and collected by Justin Drake and Owen Witte **??**. (A) Diagram depicting the workflow for phosphopeptide enrichment and quantitative mass spectrometry as described in [63]. (B) Unsupervised hierarchical clustering heatmap of phosphoserine and phosphothreonine peptides identified from prostate cancer cell lines and tissues. Over 3,900 unique phosphopeptides were significantly identified from over 36 samples [13].

TieDIE pathway analysis of clinical prostate cancer samples

I first applied the MARINa algorithm [17] to find transcriptional "master regulators" as well as kinase regulators (based on kinase/substrate relationship data) with differential activity in metastatic CRPC samples as compared with treatment nave prostate cancers (Table 4.2). In addition, kinases directly identified by the mass spectrometer in our phosphoproteomic dataset were merged with the kinase regulators, before being input to TieDIE (phosphorylated kinases). From this differential analysis, we were able to input 74 transcriptional master regulators, 14 inferred kinases, and 24 phosphorylated kinases (Fig. 4.31 (b)). The inclusion of both inferred and phosphorylated kinases allowed us to increase the number of kinases that will be interrogated in the network (Fig. 4.31 (c)). As a third input to TieDIE, the background of somatic mutation and copy-number aberration was estimated using a large number of prostate cancer samples from multiple datasets (see methods 4.3.2).

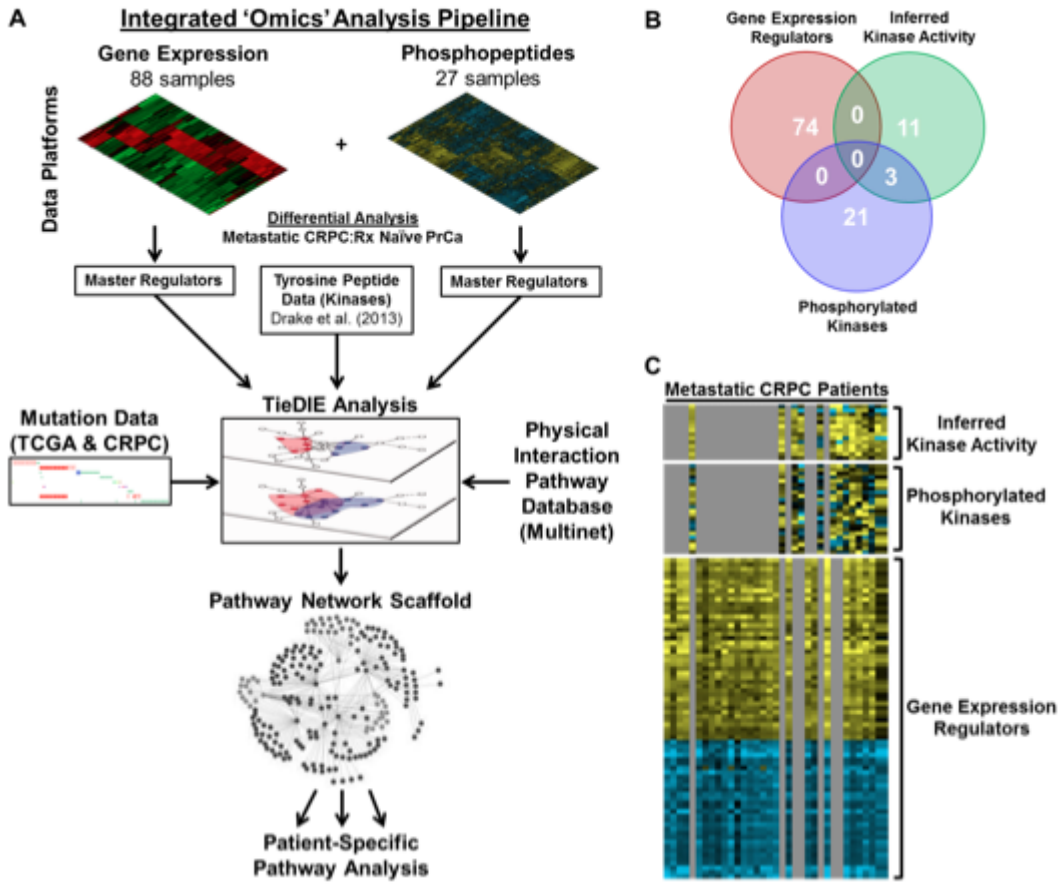

Figure 4.31: Pipeline for omic dataset integration. (A) Flow diagram depicting the integration pipeline. Gene expression and phosphoproteomic datasets were integrated with mutational data and combined using TieDIE to generate the resulting integrated network. (B) Overlay of input gene expression and kinase master regulators and phosphorylated kinases and the identification of each of these genes via a heatmap (C). The 6 patient samples used for individual pathway networks are displayed as a group on the right side of the heatmap. Yellow = hyperphosphorylation, Blue = hypophosphorylation [13].

This approach found that the identified kinases were significantly interrelated to the expression-driven transcription factor and genomic-derived input sets ($p < 0.012$), according to a conservative background model that performed 1,000 permutations of the data. The integrated network consisted of 338 nodes (40 kinases, 53 transcription factors, 86 genomically altered genes and 163 linkers) and 1,889 edges. To simplify this network, we kept only the nodes that were directly identified in the phosphoproteomic analysis (i.e. each node has a phosphoresidue identification). This resulted in a "scaffold network" for metastatic CRPC containing 122 nodes and 256 edges. Interestingly, the 61 "linker" genes identified through network topology were found to have protein phosphoresidues that were significantly more phosphorylated in metastatic samples (Figure 4.32 (b); $p < 4.5 \times 10^{-6}$). Using a set of established cancer "hallmark" gene sets from GSEA/MSigDB (http://www.broadinstitute.org/gsea/msigdb), we performed enrichment analysis to identify activities related to the disease state in the scaffold network solution. Six predominant sub-networks involved in AKT/mTOR/MAPK signaling, nuclear receptor signaling (Estrogen Response and Androgen Response), the cell cycle (G2M checkpoint, mitotic spindle, TP53 Pathway, E2F Targets, and Apoptosis), DNA repair (DNA repair, UV response), stemness (TGF-beta signaling, WNT-beta-catenin signaling, notch signaling, MYC targets, and hedgehog signaling), and migration and invasion (Figure 4.32 c-h) (Apical junction and epithelial and mesenchymal transition) were identified representing metastatic CRPC biology.

To determine how many of these proteins within the sub-networks were dependent on the inclusion of the phosphoproteomic dataset, I re-ran the same TieDIE

analysis with or without the phosphoproteomics information (Figure 4.34 a,b). I found that the resulting, comparable sized networks showed significantly different hallmark outputs as we observed more enrichment in proteins involved in cell cycle, DNA repair, AKT/mTOR/MAPK and nuclear receptor pathways when the phosphoproteomic data was included (Figure 4.33; cell cycle: 20.5 vs. 14.7, -log10 hypergeometric p-value; DNA repair: 6.6 vs. 5.1; AKT/mTOR/MAPK signaling: 4.6 vs. 1.6; nuclear receptor signaling: 8.0 vs 5.8). Inspecting each sub-network through our phosphoproteomic data revealed several enzymatically active phosphoresidues enriched in metastatic CRPC. This includes MAPK signaling targets (RPS6KA4 S343/S347, S682/T687), cell cycle targets (MCM2 S40/41, S27), and the DNA repair kinase PRKDC T2609, S2612 (Figure 4.33 b-i). Several other kinases within these sub-networks were hyperphosphorylated at residues with unknown function in metastatic CRPC including PRKAA2 S337, MAPK14 S2, STK39 S385, NIPBL S318, and SNW1 S14. These results provide evidence of actionable phosphorylation events in metastatic CRPC, several of which have previously been implicated in this disease [76,164,224] while others represent novel drug targets.

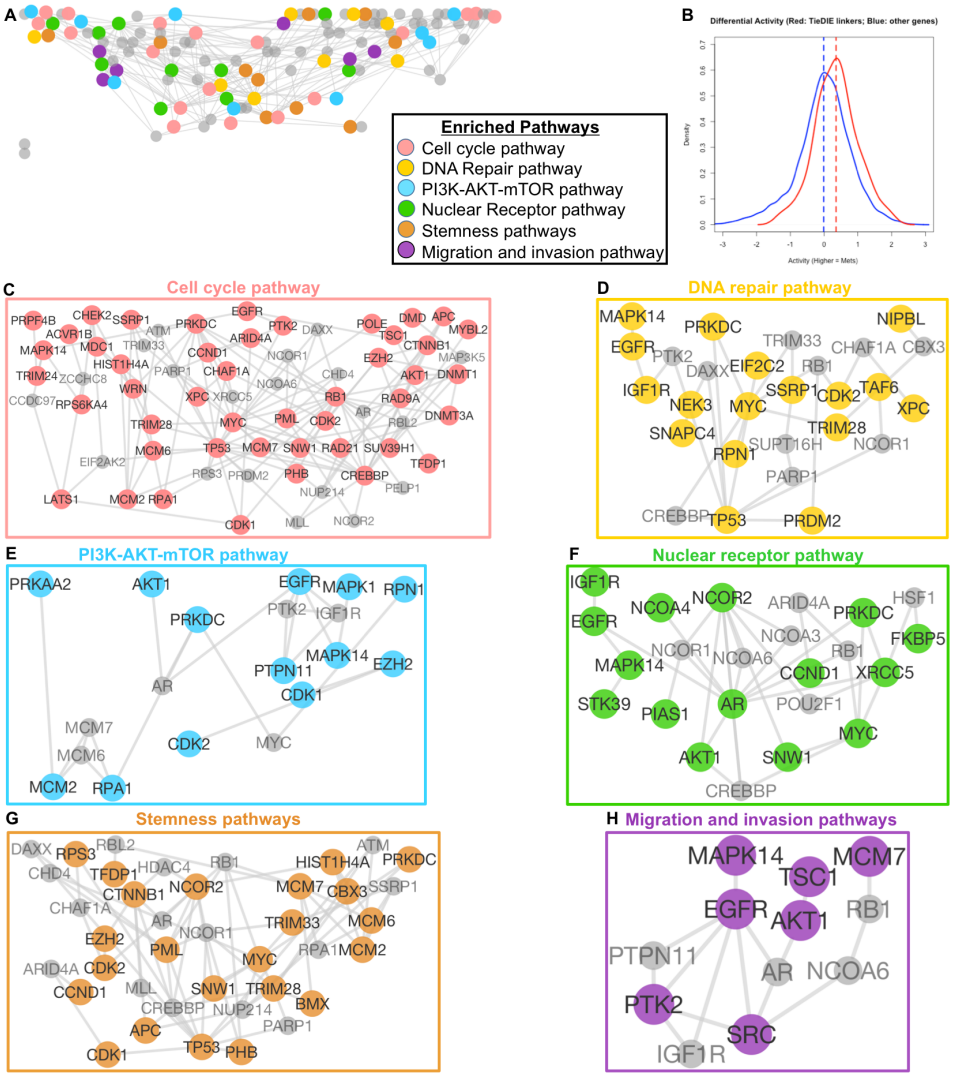

Figure 4.32: (A) TieDIE "Scaffold" network components centered on each of 6 cancer hallmark categories: for each, the set of genes in both the hallmark and the scaffold network are shown (colored) along with all adjoining edges, as well as all scaffold network genes that connect two or more of these hallmark genes (grey). (B) For each peptide in the dataset, a t-test was run between values in metastatic CRPC samples and primary controls. TieDIE "linker genes were defined as those proteins in the scaffold network that were not included in any of the 3 (genomic; kinase; transcription factor) input sets. I found the overall phosphoryation of linker genes (red) to be significantly higher ($p = 4.5 \times 10^{-6}$; two-sample Kolmogorov-Smirnov test) in metastatic CRPC samples, as compared with the distribution of differential phosphorylation in all other genes (blue), which is centered at zero. (C-H) Genes and subnetworks related to each hallmark category are shown. [13].

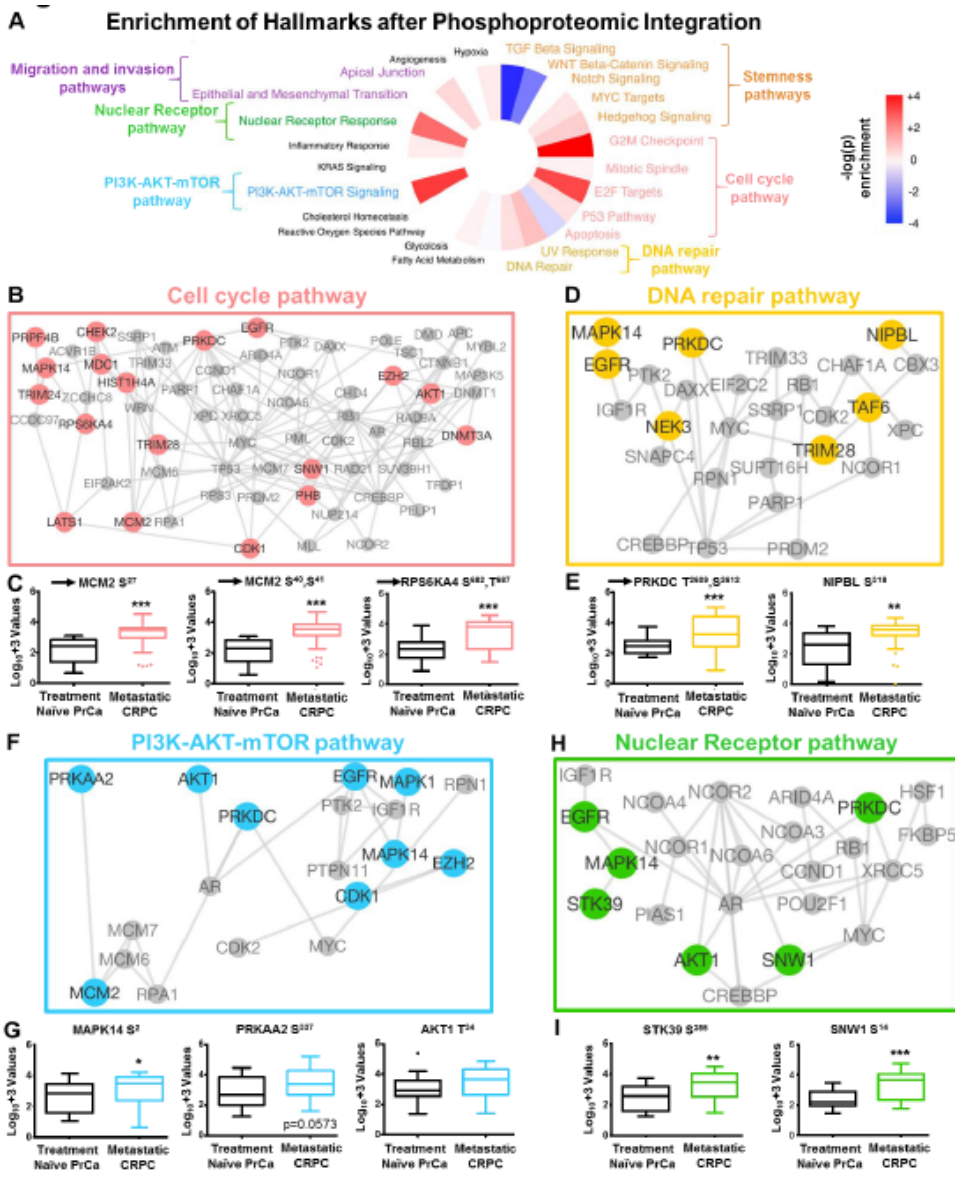

Figure 4.33: Pathway analysis of metastatic CRPC. Enriched cancer hallmarks generated by dataset integration using TieDIE after inclusion of the phosphoproteomic data (A). Several cancer hallmarks were enriched after inclusion of the phosphoproteomic data including the cell cycle pathway (B, red nodes), DNA repair pathway (D, yellow nodes), AKT/mTOR/MAPK pathway (F, blue nodes), and the nuclear receptor pathway (H, green nodes). Inspection of a select number of kinases and phosphoproteins from each network confirmed their elevated phosphorylation state (C, E, G, I) including some with direct phosphorylation on their enzymatic active residue (C, E). Black arrow represents phosphoresidues that result in enzymatic activity of the given protein [13], further supporting the activation state of the networks.

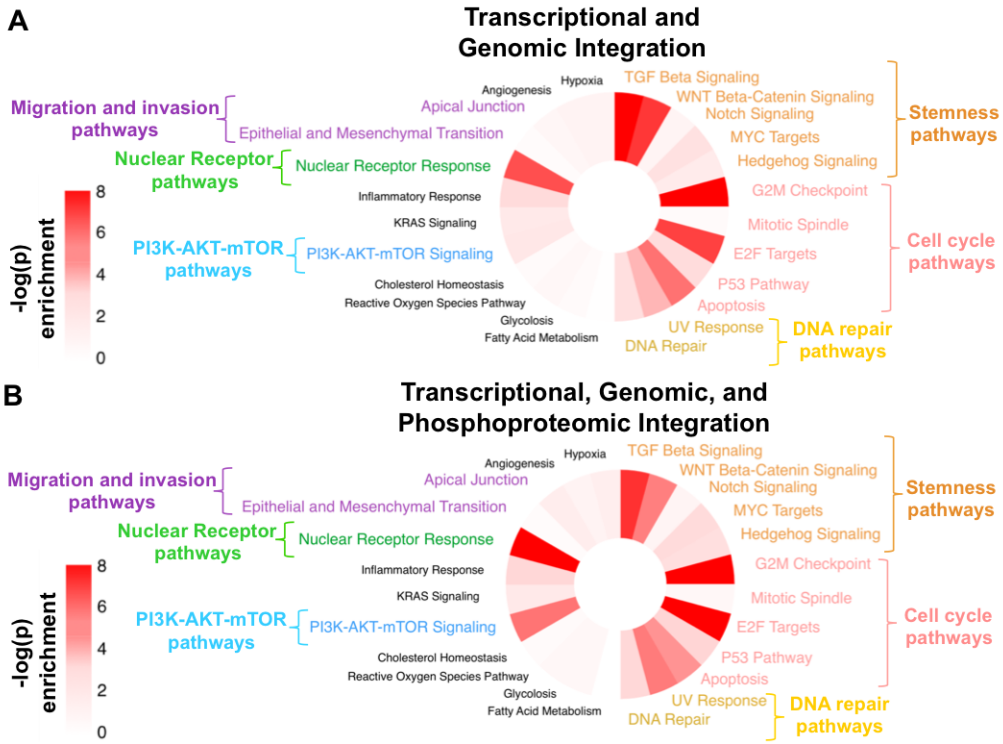

Figure 4.34: Hallmark wheels of genomic/transcriptomic and full integrated datasets. (A) Hallmark wheel showing enrichment of hallmark pathways when the transcriptional and genomic information is included. (B) Further enrichment of these hallmark pathways are observed when the phosphoproteomic data was included as a 3rd dataset.

Development of patient-specific networks using VIPER

The overall network consisting of the transcriptomic and phosphoproteomic assessment of metastatic CRPC tissues described in the previous section represents the comprehensive signaling network averaged across all of the analyzed metastatic CRPC tissues. To de-convolute the signaling networks generated by TieDIE for the purpose of identifying patient-specific signaling routes we proceeded to analyze 6 metastatic CRPC patients, of which one patient (RA55) presented with two different metastatic lesions [13]. These were the only 6 patients in our cohort that had both transcriptomic and phosphoproteomic data available, thus confining our analysis to a select few. We ran the VIPER algorithm [17] twice, to summarize the high-dimensional transcriptomic and phosphoproteomic data vectors of each patient into reliable inferences of a relatively small number of transcriptional and kinase "master regulators", respectively (Figure 4.35a). It is interesting to note that the transcriptional master regulators are highly similar across this patient cohort but the inferred and phosphorylated kinases are more variable between each of the individual patients (Figure 4.35 b). This suggests that the dominant signaling networks driving the biology of each patient may be derived from the phosphoproteomic data rather than the relatively convergent gene expression data. Importantly, we found that phosphoproteomic-driven VIPER inferences for patient RA55 (with two samples from multiple metastatic lesions; liver and dura) were very highly correlated (Spearman Rho approx. 0.87) between the metastatic sites (Figure 4.37 (a)). However, the phosphorylation abundance of specific residues on these same proteins displayed a lower correlation (Spearman Rho approx 0.1) (Figure 4.37

(b)). This additionally held true for a second patient, RA43, where we found that the overall cross-comparison correlations were higher on average for VIPER scores than with phosphorylation abundance for all three metastatic lesions (Figure 4.37 c-h). While not surprising, this finding bolstered our confidence that VIPER analysis is able to detect patient-specific protein signaling activity that is consistent across different metastatic sites.

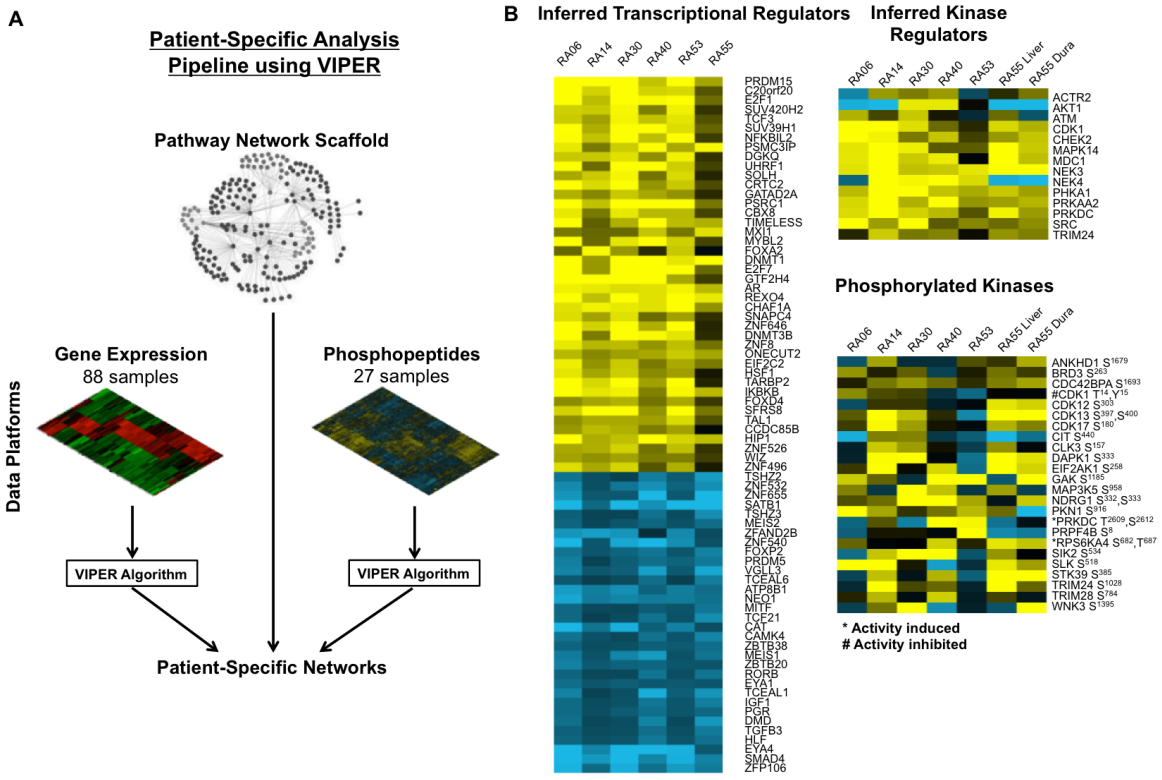

Figure 4.35: (A) Flow diagram depicting the integration of gene expression and phosphoproteomic datasets for VIPER analysis. (B) Heatmap of the gene expression and kinase master regulators and phosphorylated kinases for all 6 patients. This data was used as the input for patient-specific network analysis.

Analogous to the integrative analysis presented above, I intersected sample-specific VIPER inferences and the phosphorylation abundance of select phosphoresidues

(with published biological activity: see Patient-specific network generation methods in next section) with the integrated TieDIE "scaffold network" solution to generate separate network models. This approach allowed us to prioritize proteins not only by their activity, but also by their ability to regulate (or be regulated by) other genes implicated in that patients dataset. Importantly, we can now make use of all available cohort-level data when predicting a single patients signaling profile. Full utilization of this methodology revealed the variation in both gene-level representations of detectable targets as well the enrichment of cancer-hallmarks (Figure 4.36).

Assessment of clinically actionable pathways for personalized medicine Using this information, we have now integrated a beta version of what we call personalized cancer hallmarks using an integrated phospho-signature, or hallmark (Figure 4.36 a). hallmark will allow us to visually inspect and prioritize the signaling and regulatory pathways specific to each individual patient. These patient-specific patterns will be useful for evaluating differences in pathway events across patients but also for personalized treatment approaches.

Dissecting the hallmarks of patient RA40 revealed 4 enriched sub networks (cell cycle, nuclear receptor, PI3K-AKT-mTOR, and stemness pathways) including a large active network related to cell cycle processes (Figure 4.38). Interestingly, this was the only patient of the 6 analyzed with a missense mutation and deletion in the tumor-suppressor gene APC. While frequently observed in colorectal cancers, APC mutations can occur in other cancers [106] where its inactivation leads to increased -catenin activity [146]. Indeed, we observed strong phosphorylation of the enzymatic active site

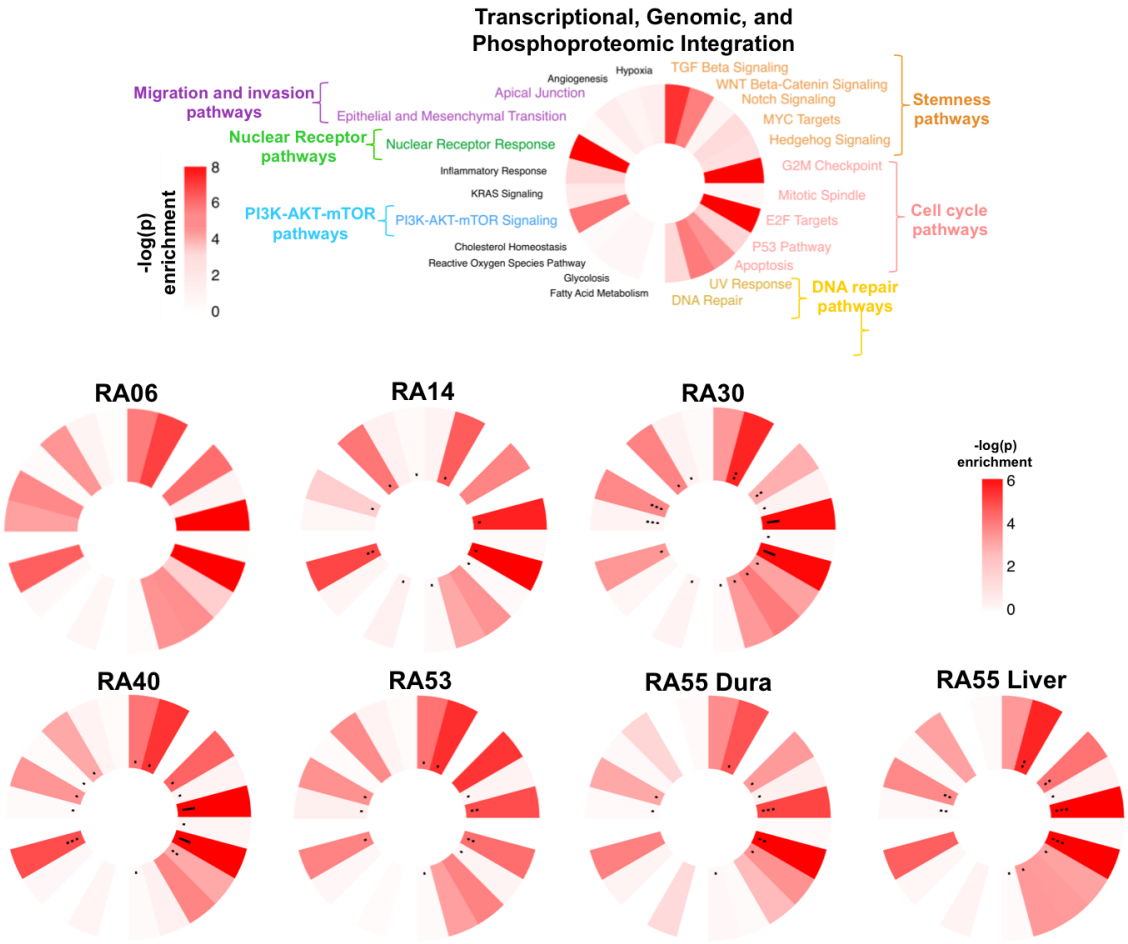

Figure 4.36: Hallmark wheels for each patient. Top: the integrated TieDIE network solution shows strong enrichment for genes in Migration and Invasion Pathways, Stemness Pathways, Nuclear Receptor Response, PI3K-AKT-mTOR Signaling, Cell Cycle and DNA Repair related categories. Bottom: hallmark wheels for 6 individual patients show varying, but strong enrichments for categories. Color indicates the log(p) value (uncorrected) for the hypergeometric overlap test between genes in each patient-specific network and the corresponding hallmark category.

of -catenin (S675) with very moderate phosphorylation of APC residues, though the function of these residues is currently unknown. The putative activation of EZH2 is also linked to -catenin activation in several cancers including hepatocellular carcinoma and breast cancer [35, 38, 191] . EZH2 activation in this sample is supported by both

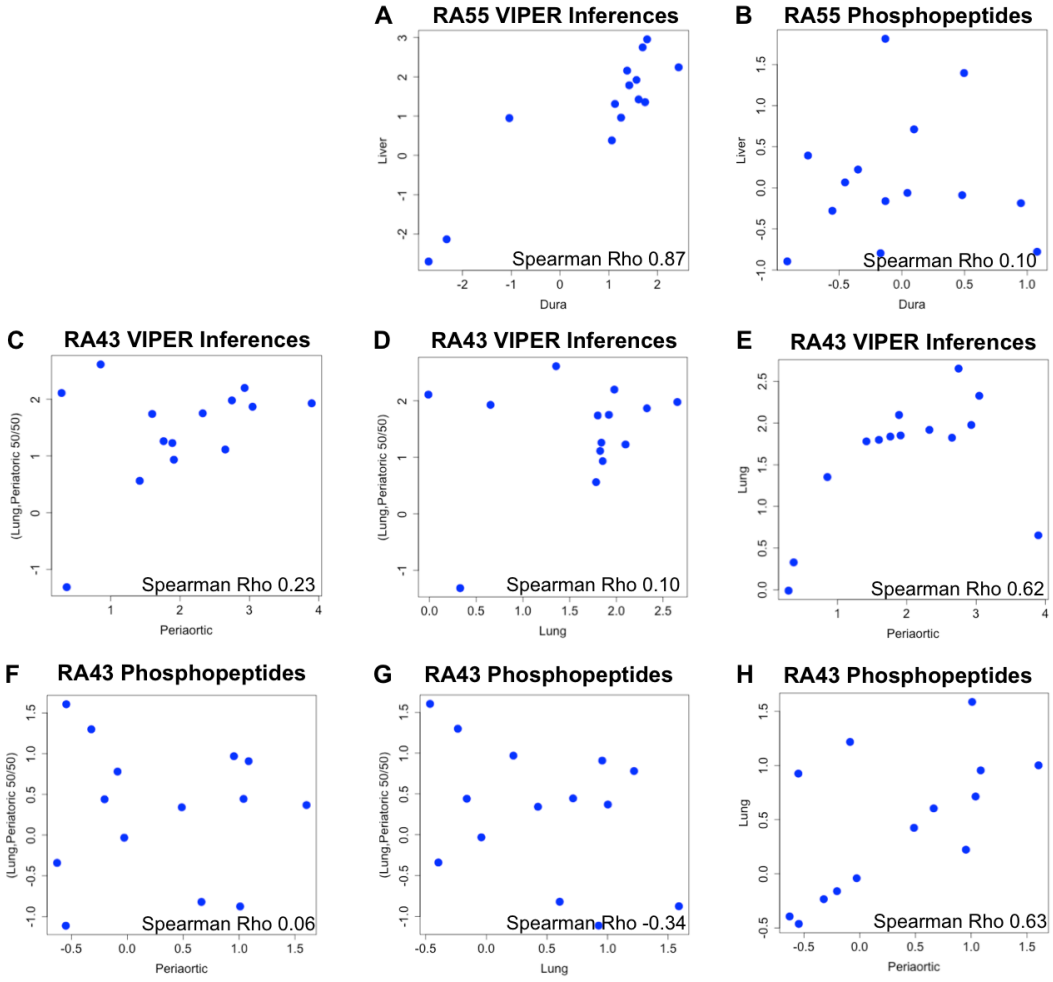

Figure 4.37: Scatter plots of correlation between tumors found in multiple metastatic sites in patients WA55 and WA43. (A) Comparison between VIPER inferences in liver and dura metastases of patient WA55 show very high correlation (0.87 Spearman Rho), while in (B) we see low (0.1 Spearman Rho) correlation between individual phosphopeptides on the same proteins. (C) Comparison between VIPER inferences in samples taken from periaortic, lung and mixed metastatic sites show low to moderately high correlation (left to right: 0.23, 0.1, 0.62 Spearman Rho, respectively) in patient WA43. (D) Comparison between phosphopeptides abundance in the same proteins shows negative to moderately high correlation (left to right: periaortic vs. mixed, 0.06; lung vs. mixed, -0.34; periaortic vs. lung, 0.63 Spearman Rho) in patient WA43.

heterozygous amplification and de-phosphorylation of residue T487 (a marker for ubiquitination of EZH2; phosite.org) as well as amplification of DNA methyltransferase 3

Table 4.2: VIPER scores for kinase regulators enriched in metastatic CRPC samples.

VIPER Scores for Kinase Regulators				
Kinase	Regulon Size	NES	p.value	FDR
PRKDC	759	4.01	6.09E-05	0.00237
MAPK14	327	3.83	0.000126	0.00237
MDC1	236	3.37	0.000756	0.00423
CDK1	792	3.33	0.000854	0.00423
NEK4	87	3.29	0.00102	0.00423
AKT1	403	3.16	0.00157	0.0044
NEK3	89	3.12	0.00182	0.00442
PHKA1	209	2.65	0.008	0.0143
PRKAA2	307	2.33	0.02	0.0293
SRC	25	2.27	0.023	0.0314
ACTR2	79	2.24	0.025	0.032
ATM	136	2	0.046	0.0555
CHEK2	244	2	0.046	0.0555
TRIM24	26	1.9	0.058	0.0679
PRKCZ	44	1.51	0.132	0.135

(DNMT3) and predicted activity of DNTM1 (inferred by the gene-expression activity of its downstream transcriptional targets) [159] (Figure 4.39 (b)). Further, the amplification and predicted (transcriptional) activity of SUV39H1 correlates with EZH2 expression in tumor development [163], consistent with our observations. Mechanistically, EZH2 activity is sufficient for activation of AKT1 [75], which we observed through both hyperphosphorylation of the enzymatic active site T308 as well as high predicted kinase activity inferred through patient-specific VIPER analysis (Table 4.2). Together, this information implicates the involvement of AKT1 in contributing to altered cell cycle regulation and growth, and makes a strong case for targeted inhibition in this patient. Similar mechanisms relating to other signaling pathways for other patients can also be described.

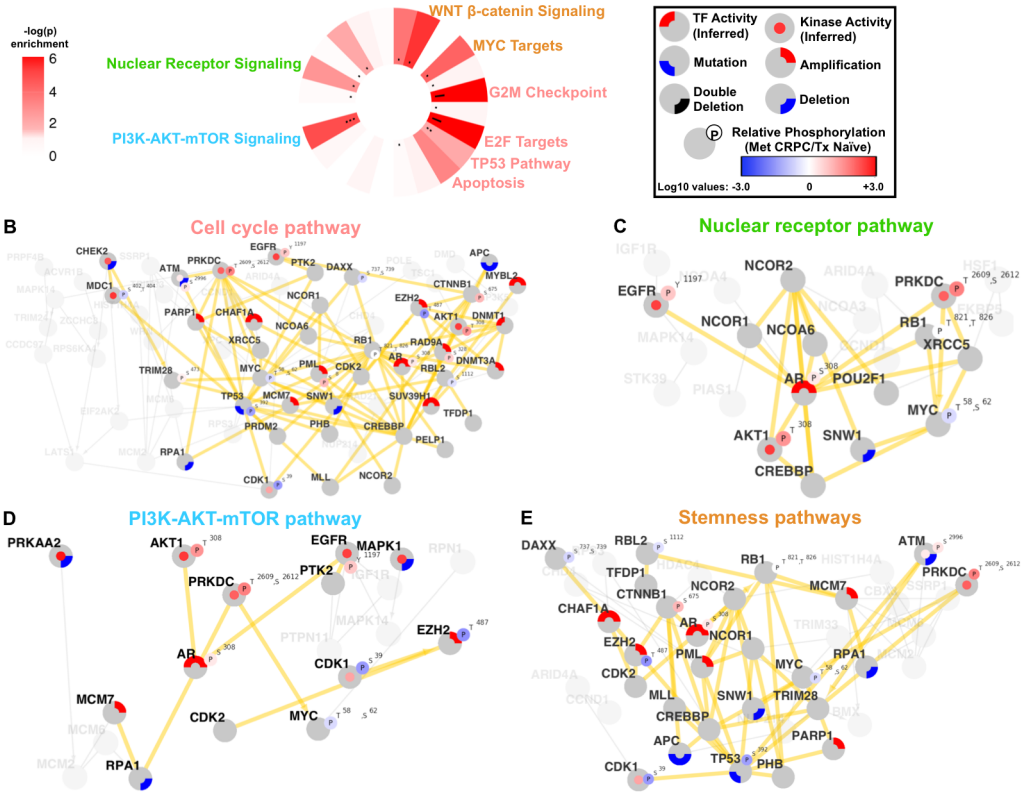

Figure 4.38: Integrated pathway network of patient RA40. (A) Hallmark analysis for patient RA40 revealed strong enrichment of cell cycle and PI3K-AKT-mTOR pathway networks. The hallmark wheel summarizes enrichment between genes in each patient-specific network and the corresponding category: labels indicate categories with significant enrichment after multi-hypothesis correction (q <0.1). Dots indicate SNV and copy-number genomic events in this patient. Patient-specific network nodes and edges related to cell cycle pathway (hallmark categories G2M checkpoint, mitotic spindle, E2F targets, TP53 pathway, and apoptosis) nuclear receptor pathway, PI3K-AKT-mTOR pathway, and stemness pathways (hallmark categories, TGF-beta signaling, WNT-beta-catenin signaling, notch signaling, MYC targets, and hedgehog signaling). Edges belonging to both the patient-specific network model and the cell cycle related scaffold network are shown as thick yellow edges, while corresponding genes are shaded in dark grey. Yellow arrows indicate that the upstream kinase directly phosphorylates the downstream substrate. "Circleplot" quadrants for each gene summarize genomic, transcriptomic and phosphoproteomic activity relevant to metastatic CRPC phenotype (upper right: amplification; lower right: deletion; lower left: mutation; upper left: transcriptional regulatory activity; center: kinase regulatory activity). Node "ears" peripherally attached to circleplots represent relative phosphorylation of specific, functionally annotated peptides sites on each protein. Genes and edges in the scaffold network are only shown in light grey.

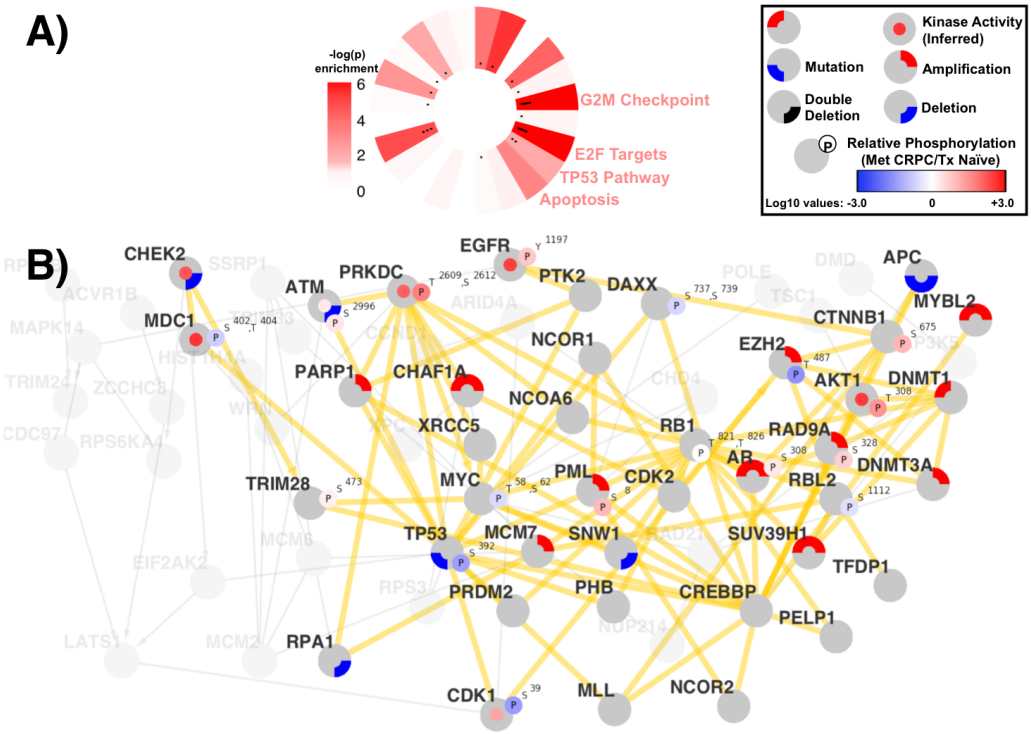

Figure 4.39: Integrated pathway network of patient RA40, focusing on cell cycle related pathways. A) Hallmark categories related to cell cycle: G2M checkpoint, mitotic spindle, E2F targets, TP53 pathway, and apoptosis show significant enrichment after multi-hypothesis correction in Patient RA40. B) Network diagram corresponding to cell cycle related pathways, as described in Figure 4.38.

Under the assumption that we seek to reverse as many altered gene activities found in a patient, we considered the idea of using a minimum combination of targets that influence the largest area in a patients network. Inhibiting genes at the top of the network may offer more powerful targets than those at the bottom. Understanding the nesting of gene regulatory signals provides information about how to select genes for this purpose [13]. Therefore, I developed a hierarchy of therapeutic kinase targets through topological evaluation of the cancer hallmarks and pathways of each individual patient (Figure 4.40). The hierarchy reveals the top kinase targets for every individual patient that we analyzed and the corresponding therapeutic intervention including PRKDC, CDK1, CDK2, AKT1, SRC, and EFGR. Given this structure, targeting of a single kinase such as PRKDC may be sufficient to blunt the activity of other kinases that phosphorylate many of the same targets (NEK3, PRKCZ) and those that are, additionally, predicted to be phosphorylated by PRKDC (ATM, IKBKB) [13].

The hierarchy allows for the construction of feasible strategies for patients with high activity in multiple kinases, such as WA30 (or WA53) where the joint inhibition of SRC and AKT1 may reduce activity of PRKDC and CDK1 as well as multiple downstream kinases (Figure 4.40). Both WA30 and WA53 may also need EGFR inhibitors in addition to these. As another example of how to use the regulator hierarchy, different combinations could be selected for those patients with low SRC or AKT1 levels such as WA14. For example, PRKDC stands out for WA14 since the patients CDK1 levels are also low. In other cases, such as WA55, inhibition of either SRC or CDK1 may be worthwhile, with SRC still relevant as well for the subclone(s) represented in the dura

sample taken from this patient [13].

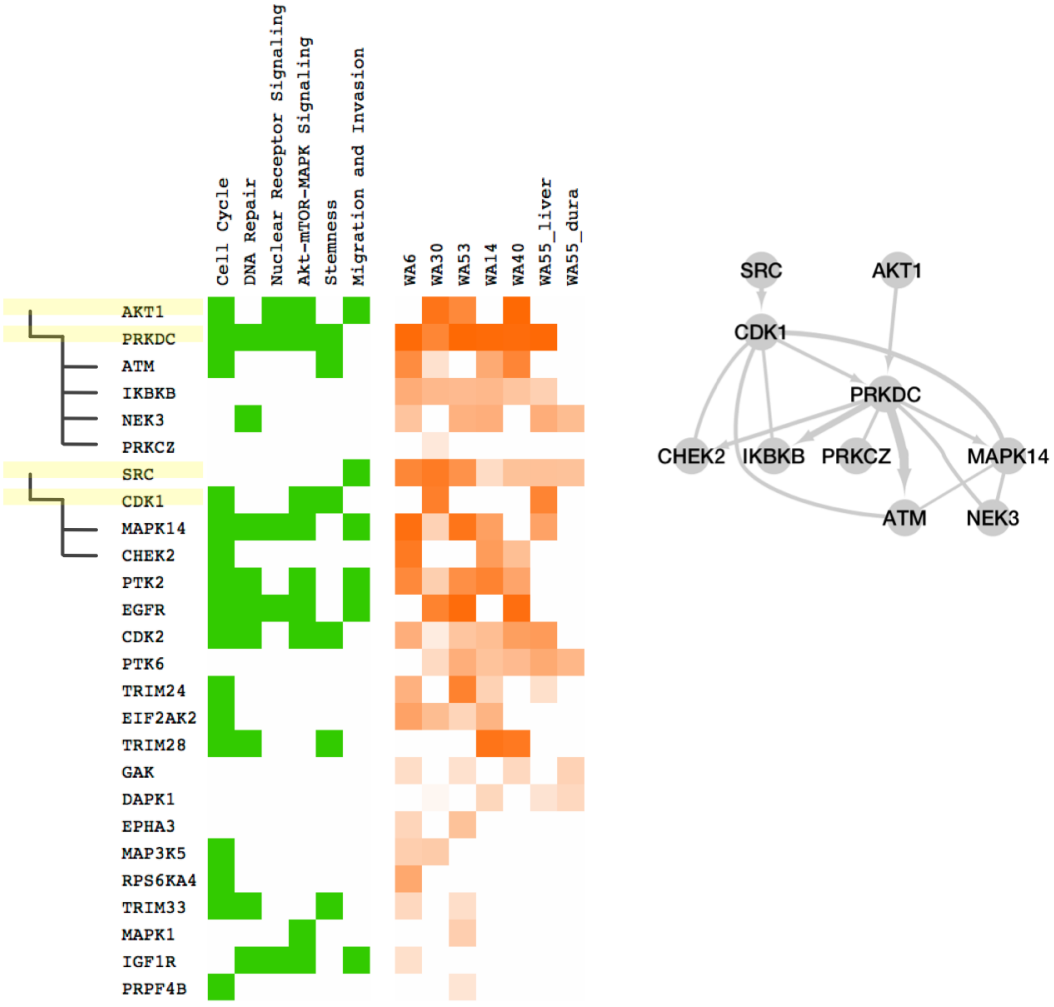

Figure 4.40: Summary of kinase targets in patient-specific networks. Left: green boxes indicate kinases (rows) that are members of each of the six major hallmark subnetworks (columns) shown in (Figure 4.32). Right: Orange boxes indicate the predicted importance of kinase targets based on the combined evidence from VIPER-inferred kinase activity, phosphorylation status of functionally annotated peptides, and connectivity, for each patient specific network (columns). Far Right: Network diagram of hierarchy of kinase interactions. Edge thickness represents the degree of overlap of protein targets; directed arrows are drawn when a given source kinase is predicted to phosphorylate a site on the target.

4.3.2 Methods

Development of a metastatic CRPC network using TiEDIE

To prioritize kinases that are likely to regulate the observed gene-expression profile of metastatic samples, and be related to genomic aberrations observed in prostate cancer, I applied the TieDIE algorithm as part of a novel computational pipeline. This approach allowed for the integration of complementary transcriptomic and genomic datasets collected from partially overlapping metastatic CRPC tissues or patients, as well as prior knowledge in the form of pathway databases, to find proteins implicated by multiple forms of biological evidence (Fig. 2a).

Replicates were averaged before running a t-test between 11 treatment naive and benign prostate cancer samples and 16 metastatic CRPC samples corresponding to 13 distinct patients [210]. 1,045 differentially phosphorylated peptides with higher phosphorylation in metastatic samples above a FDR of 0.1 were retained, 24 of which mapped to known kinases [135]. In addition, we performed "Master Regulator Analysis (MRA; see appendix A.3) to find kinases implicated by the combined phosphorylation activity of their putative targets [17]. To facilitate this analysis, Justin Drake and Nicholas Graham [63] combined multiple databases of predicted kinase-substrate predicted interactions to produce a comprehensive 'regulome of candidate regulator kinases that are predicted to phosphorylate at least 25 proteins on at least one site [63, 64, 115]. We ran the MARINa algorithm [17], to find 'kinase regulators with significantly higher activity—as inferred from the peptides they are predicted to phosphorylate—in CRPC

samples compared to the control primary and benign tumor samples. This analysis found 14 kinases with higher inferred activity in CRPC samples (FDR <0.1), 3 of which overlapped with the set of differentially active kinases, resulting in a final list of 35 putative kinase regulators in the CRPC samples.

Separately, I ran master regulator analysis (see appendix A.3) on microarray expression data consisting of 49 primary tumor and 27 metastatic samples. This analysis finds transcription factors (TFs) that are likely drivers of a large fraction of the observed expression differences between these CRPC and control samples. Because this analysis requires a computationally derived 'interactome between TFs and putative targets that must be tissue specific, we used a pre-determined interactome of transcription factor-to-target regulatory edges, inferred using a diverse sample of normal, primary and metastatic prostate cancer samples as well as cell lines [19]. While this dataset does not include CRPC samples, it contains samples with a high degree of genomic and clinical diversity, leading to a highly comprehensive interactome [19]. This step found transcriptional regulators with significantly higher or lower activity in CRPC samples than the control primary tumor samples, resulting in 74 TF genes (q <0.1), weighted by the absolute value of the MARINa test statistic. We again used the VIPER package to infer the sample-specific activity of each of these regulators in each metastatic sample, using the 27 primary samples to compute the reference distribution of activity scores.

We combined genomic data from multiple sources to estimate the genomic background of metastatic prostate cancer. Briefly, data and analyses from The Cancer Genome Atlas (TCGA) were downloaded; a MutSig 2.0 [119] analysis on 261 TCGA

primary tumor samples revealed 14 predicted driver genes that were found to be significant at a q-value cutoff of 0.1. This, combined with a partially overlapping list of 9 predicted driver genes from 49 metastatic samples [78] produced a list of 21 genes with putative driver mutations in prostate cancer samples. Similarly, an analysis of the same 49 metastatic samples revealed significant amplification events encompassing several key genes (PIK3CA, HOXA3, AR) as well as significant deletion events (NT5E, CHD1, PTEN, RB1, TP53). An analysis of 492 primary TCGA tumors revealed 28 significant focal amplifications, and 37 significant focal deletion events implicated as drivers according to the GISTIC algorithm [140]. Because a large number of genes were involved in these events we further filtered these genes by intersecting with the most recent COSMIC census, representing 572 cancer-related genes according to our prior knowledge. When combined with the genes found in the CNV analysis of metastatic samples, this resulted in 96 candidate genes with copy number alterations and 112 genes with either mutation or copy number events with driver potential. Of these, 108 had at least one copy-number or mutation event in the 49 metastatic samples.

We used the TieDIE algorithm [165] to connect 35 kinases and "kinase regulators", 108 putative cancer driver genes with genomic perturbations in CRPC and 74 transcription factors, using the "Multinet" [110] pathway database consisting of a diverse set of literature-based gene-gene interactions (43,722 protein-protein interactions; 27,900 direct phosphorylation; 27,914 transcriptional/regulatory; 9,714 metabolic; genetic interactions excluded). Each of these three inputs were treated as a separate, equally weighted, input set for the algorithm, while the gene members of each input

set were weighted by the total evidence for each protein: kinases by combined SAM d-statistic and MARINa inferred activity level, transcription factors by MARINa inferred activity level, and genomic events by the number of mutations and copy-number alterations observed in the 49 metastatic prostate cancer samples. The kinase, genomic event and TF gene sets were found to be significantly close in pathway space ($p < 0.012$), according to a conservative background model run with 1,000 permutations of the input data.

The resulting network consisted of 338 nodes and 1,889 edges (597 direct phosphorylation; 1,184 protein-protein interaction; 102 transcriptional/regulatory; 6 metabolic). This network was filtered further by restricting to gene-gene edges with at least one pair of constituent phosphopeptides with at least modest correlation (Spearman rank correlation, Rho ≥ 0.3), resulting in a final scaffold network of 122 nodes and 256 edges (190 protein-protein interaction; 131 phosphorylation).

VIPER analysis of individual prostate cancer patients

I used the VIPER package [17] to infer sample-specific activity (pseudo z-scores) of each of each of the 14 kinase regulators that were found to have significantly higher average activity in CRPC samples, using the 11 treatment naive or benign samples to compute the reference distribution of activity scores. Similarly, I used VIPER to generate inferences for each of the 74 TFs, for each sample, using the same microarray data, regulon and reference samples used in the MARINa analysis on the Grasso dataset.

Patient-specific network generation

To generate sample-specific networks, I applied thresholds to each samples data, generating binary calls for each of the 35 kinase regulators and 74 TFs, respectively. Scores for the 24 peptides with significant differential phosphorylation activity were z-normalized by gene and thresholded at a z-score of 1.0 or above, while VIPER pseudo z-scores were thresholded at the level corresponding to a 0.1 FDR cutoff in each corresponding Network Enrichment Score (NES) for MARINa analysis, 1.89 and 2.83 for the kinase and transcriptional regulators, respectively. Kinases were called as "active" if either the VIPER inference or cis-phosphorylation status passed the respective thresholds. For each sample, we searched all paths connecting any active kinase to any active TF over edges contained in the scaffold network, using the NetworkX python package [83] and up to an edge-depth of 4; directionality was used when available, while protein-protein interactions were converted to bi-directional edges. Networks were visualized using the Cytoscape 3.2.1 software package [188].

For all proteins in each patient-specific network, I performed three independent rankings based on the phosphorylation activity of functionally annotated peptides, VIPER inferred activity scores, and the network connectivity. For the first metric, I used log10 normalized phosphorylation values to rank all phosphopeptides with an available functional annotation (phosphosite.org) that map to a protein/gene in the corresponding patient-specific network. VIPER scores inferred from the phosphoproteomic data were ranked separately, and the network connectivity ranking was generated by computing the shortest-path between centrality for all genes, with the NetworkX python pack-

age [83]. For the first two rankings, genes without a corresponding functional peptide or VIPER score were given a rank of N+1, where N is the number of ranked genes with each respective method. These three independent rankings were then equally weighted and summed by gene, generating a final combined ranking for each patient. Figure 4.40 illustrates the combined rankings of kinases for each patient; hallmark membership of each kinase is determined by inclusion in the corresponding TieDIE scaffold sub network shown in figure 4.32.

Further computational validation

To assess the ability of the patient-specific network predictions to remain robust to noise in the phosphoproteomic dataset, I designed a procedure to select random subsets of the mass-spec data, and performed ten repetitions. For each, 20% of the peptides were randomly sampled for each of the 7 samples with patient-specific networks, and the corresponding mass-spec data was removed. I then repeated the procedure to generate a TieDIE scaffold network with this diminished dataset, using the same parameters to generate a network of comparable size; similarly, I repeated the VIPER analysis of each patient and the patient-specific network generation procedures as before. Each finalized patient-specific network was then evaluated by counting the number of functionally annotated peptides that had been absent from the subsampled data for that specific patient, had a corresponding gene in the patient-specific network prediction, and high phosphorylation (z-score <1.0) in that particular sample. I measured the number of these "recovered" peptides and divided by the expected number (<1, given the size

of the networks) to get a ratio statistic for each patient, in each subsampled dataset, finding that most networks had a significantly high enrichment of these peptides (Figure 4.41), confirmed with a hypergeometric overlap test.

The contribution of the TieDIE scaffold network to the observed robustness measures was assessed using the same data-subsamples as described above, to generate patient-specific networks in a scaffold-independent manner. For each patient, VIPER scores were thresholded as before and then used as direct inputs to the TieDIE algorithm, replacing the kinase, and TF gene sets outlined in the TieDIE methods above. Genomic alterations were used as a third input set, included if reported in both that patients data and in the list of 108 altered genes established (see TieDIE pathway analysis of clinical prostate cancer samples methods). TieDIE was run with the same size parameter established in the scaffold generation step, to ensure final networks of a similar size. The ratio statistic of "recovered" peptides was measured as done with the scaffold-enabled networks; because this metric counts only peptides measured in the phosphoproteomic data in the denominator, this measure controlled for network nodes/genes without any measured peptides. I found a higher average ratio of hyper-phosphorylated, functionally annotated peptides in 6 of 7 sample-specific networks (Figure [**?**]), and similar levels in the seventh.

4.3.3 Discussion

Current clinical inhibitors targeting AR in late stage prostate cancer patients provide survival benefits of 3-4 months [57, 187]. This work provides clues into the

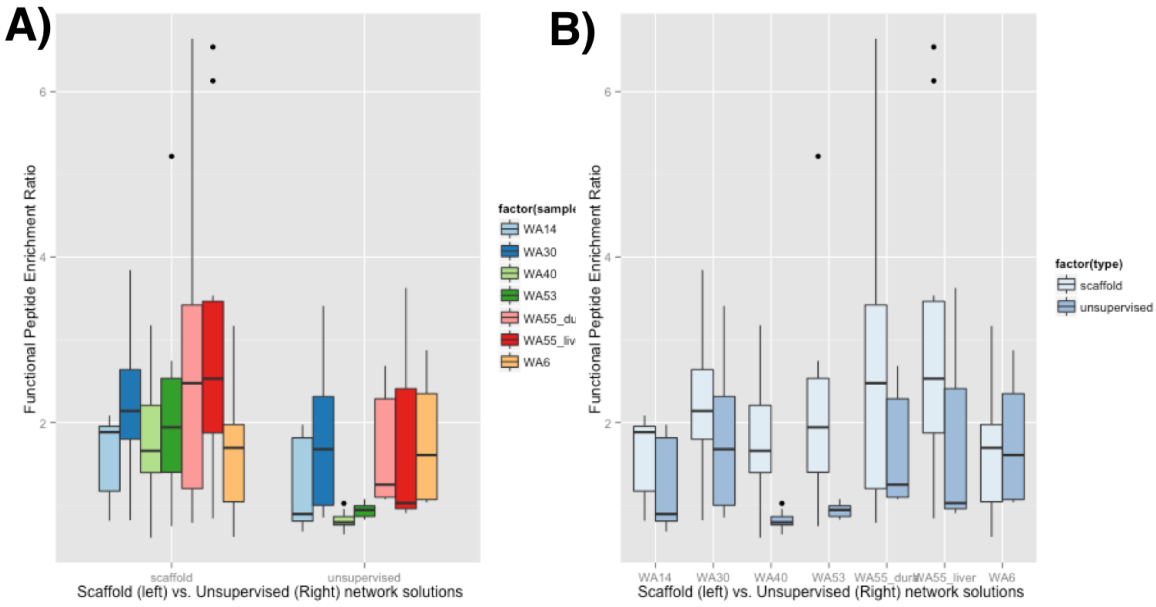

Figure 4.41: Patient-specific network evaluation. A) Patient-specific network solutions are evaluated by recall of hyperphosphorylated, functionally annotated residues (x-axis, ratio of recovered by expected) for 7 metastatic samples, using the TieDIE scaffold methodology (left) and scaffold-free solutions for each patient (right). B) The same robustness measures shown with a side-by-side comparison between patients.

signaling pathways that are activated in metastatic prostate cancer patients, and as a result of resistance to current AR targeted therapies. Although the way to target these kinases in these patients is not entirely clear (recent clinical trials using single agent kinase inhibitors do not increase overall survival in CRPC patients [**?**, 118]), co-targeting strategies for cancer therapy in other cancer types has proved to be better than administration of single agents in pre-clinical models [54, 126]. It follows that advanced prostate cancer patients could also benefit from targeted, combination therapies, and several clinical trials are currently underway to address this possibility including inhibition of AKT, MET/VEGFR2, or SRC in combination with AR blockade (NCT01485861,

NCT01995058, NCT01685125).

In collaboration with Justin Drake, Owen Witte and other experimental scientists at UCLA, the computational pipeline I developed identified many interesting protein targets in CRPC patients. The identification of mTOR and MAPK signaling pathways through this computational approach are not new in prostate cancer, but several signaling molecules such as PRKDC, AMPK, PTK6, RPS6KA4, and even CDK family members within these pathways provide strong support to therapeutically target these proteins as well [13]. Additionally, the patient-specific analysis presented here sheds some light into the diversity of the activated signaling pathways in metastatic CRPC patients, and highlights the strength of systems biology methodology that can integrate multiple biological perspectives in a single model [13]. In particular, the model developed here may be generally useful for analyzing datasets that have only partial sample overlap between datatypes (i.e. gene expression and mass spec data). By combining the signatures constructed on a larger sample group, I was able to create a tissue-specific "scaffold" network that was enriched for proteins with significantly higher phosphorylation in metastatic CRPC than when compared with primary and benign controls. Incorporating this network in the patient-specific analysis greatly improved the recall of functionally annotated and hyperphosphorylated peptides in patient-network models, when tested with a held-out data cross validation setup.

As with previous studies of tyrosine kinases [63], we found a large degree of hetereogeneity between samples, while also observing stronger similarity between different metastatic sites in the same patients. In addition, the enrichment-based visualization de-

veloped in collaboration with Justin Drake provides a systematic way to rapidly identify cancer-related processes in each patient while simultaneously summarizing the complex relationships between genes and proteins, and the evidence provided by multiple assays of biological data. Taken as a whole, the system to analyze metastatic CRPC presented here represents a prototype of a personalized treatment strategy. However, further interrogation of these patient-specific networks in appropriate pre-clinical models are necessary, and more experimental and computational work is needed before personalized medicine can propose sound treatment recommendations for lethal prostate cancer patients such as these.

Future Directions

This work points toward the need for developing personalized network models of disease that combine all concepts presented in the previous chapters. These include data integration over a biological network prior, the generation of tissue-specific networks as an intermediate step in personalized network construction, identification of cancer-essential genes and new techniques to explain how pathway and process-level aspects of a tumor are explained by a patient's genomic and kinase-signaling alterations. Each of these concepts represents part of an overall strategy to integrate data produced by multiple high-throughput assays, each capturing a unique perspective of cancer biology, into a single personalized model of disease.

In the first chapter, I presented a new methodology (TieDIE) to select genes and interactions (sub networks) from prior literature that are dependent on the measurable genomic, transcriptional and (if available) phosphoproteomic signaling context of cancer cohorts. This method provides a way to integrate multiple, disparate types of biological data with prior knowledge, and has been rigorously tested on a large breast cancer cohort where it was better able to extract relevant genes and interactions than previous methods.

The second chapter details multiple applications of the TieDIE method, where it was applied to three additional TCGA tissue-specific cohorts with genomic, transcriptomic and proteomics A.1.5, as well as one "pan-cancer" dataset. In kidney and bladder cancer, this analysis revealed key literature-backed interactions that connect mutated genes with histone-related activity to a wide range of transcriptional processes relevant to tumor maintenance and growth. In the thyroid cancer cohort, the inclusion of RPPA [**?**] proteomics data allowed the TieDIE algorithm to find key areas of the signaling pathways that are downstream of activating BRAF V600E and RAS mutations. In each case, the TieDIE algorithm provided new evidence that genomic events already presumed to drive tumorogensis do influence the observed pattern of gene expression, mediated by transcriptional "master regulators." The algorithm also revealed specific mechanisms of that influence. In broader terms, these results help to address the question of how tissue and disease-specific processes work to change the way genes and proteins interact, with consequences for the progression and potential treatment of disease.

I continued to explore this question with my participation in the DREAM8 (network inference) and DREAM9 (gene essentiality) challenges: in each case I was part of a competition winning team that explored two sub-questions related to the central question of finding patient-specific pathway models. In the first, our team used time-course proteomics data to predict protein interaction networks specific to cell line/stimulus pairs. As a model system, the cell line and stimulus conditions provide an approximation to the genomic background of patient-tumors, and corresponding

extra-cellular environments. We found that these networks are reasonably predictable, and that the use of prior knowledge is fundamental in making good predictions. This is an encouraging result, as knowledge of the specific rewiring of protein interactions under a given genomic background may lead to key observations that can guide treatment options in the future. At the same time, this result is a strong datapoint supporting the hypothesis that pathway knowledge derived from normal biology is highly relevant to the study of cancer tissue.

The second (DREAM9) challenge required prediction of the relative gene essentially—potential for reduced growth after knockout—in a panel of cell lines, a proxy for the therapeutic potential of targeted drugs in patients. The use of prior knowledge again proved critical for this challenge, for us and other participants (a DREAM9 marker paper is in progress), and we found that the essentiality of many genes can be predicted with high accuracy. However, the genomic state of these cell lines proved to be of little use in predicting gene essentiality (with our and all other teams in the competition), meaning far more work still needs to be performed to find biologically interpretable patterns of gene expression and genomic modification that can find specific vulnerabilities in patient tumors.

In my next research phase, encouraged by the results of previous analysis, as well as recent studies [66, 94] supporting the idea that the interaction structure of networks is dependent on both genomic context as well as tissue-specific effects, I applied the same approaches to a study of lethal prostate tumors obtained by rapid autopsy. Borrowing from the ideas developed by running TieDIE on large TCGA datasets, I

used the same tools to integrate genomic, transcriptomic and phosphoproteomic data into a tissue specific network specific to metastatic prostate cancer. I then mapped patient-specific data and activity inferences onto this network, extending on the original methodology developed for the TieDIE algorithm 2.2.2.4 [165], and used it to predict patterns of oncogenic signaling activity in a sample of patients with phosphorproteomic, genomic and expression data. This methodology provides a way to integrate multiple datatypes with only partially overlapping samples, and my results show that this allows for better recovery of held-out, true-positive data in at least one dataset. In addition, a new visualization paradigm (the cancer "dashboard") provides a more intuitive way than many complex network visualizations for researchers and clinicians to see how pathway and process-level aspects of a tumor are changed by a patient's genomic and signaling alterations.

As of now, phosphoproteomic data is expensive to collect and may not be practical for general clinical assessment. However, the ability to relate patterns of kinase and signaling pathway activation to more accessible measurements of genomic and transcriptional state should begin to be accomplished with only moderately larger sample sizes than presented here, where machine-learning methods can be applied to create such a mapping. However, at present it is also easy to imagine a future where genomic and transcriptional data will continue to be insufficient to find clinically actionable, tumor-specific vulnerabilities in all cases, meaning that direct measurement of kinase activity may be extremely valuable in choosing the right drugs for a given patient. In either case, more complex models will be needed to predict combinatorial therapies through

joint analysis of these data and prior-knowledge.

Crucially, through rigorous computational validation and a detailed analysis of cohort and patient-specific network models, this work demonstrates that computational models that simultaneously analyze data produced by multiple high-throughput assays can reach conclusions that far exceed any single-data analysis. Extrapolating from this conclusion, we can assume that future technologies will not only improve our understanding of cancer biology but also accelerate the pace of discovery, pending computational models that can fully exploit this data. Model complexity necessarily increases with adding additional data modalities, so standards and technologies that aim to make data more computable will become increasingly important to the field of computational systems biology.

Bibliography

[1] A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[2] An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13550–13555, 2005.

[3] The catalogue of somatic mutations in cancer (cosmic). *Current protocols in human genetics editorial board Jonathan L Haines et al*, Chapter 10(March):Unit 10.11, 2008.

[4] Overexpression of separase induces aneuploidy and mammary tumorigenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 105(35):13033–13038, 2008.

[5] Monitoring proteins and protein networks using reverse phase protein arrays. *Disease markers*, 28(4):225–232, 2010.

[6] Chasm and snvbox: toolkit for detecting biologically important single nucleotide mutations in cancer. *Bioinformatics*, 27(15):2147–2148, 2011.

[7] De novo discovery of mutated driver pathways in cancer. *Genome research*, 22(2):375–85, 2011.

[8] A decade's perspective on dna sequencing technology. *Nature*, 470(7333):198–203, 2011.

[9] The gene ontology 2011, 2011.

[10] Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[11] Wikipathways: building research communities on biological pathways. *Nucleic acids research*, 40(Database issue):D1301–7, 2011.

[12] Comprehensive molecular portraits of human breast tumors the cancer genome atlas network. *Nature*, 2012.

[13] Patient-specific ... 2016.

[14] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010.

[15] Ruedi Aebersold and Matthias Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[16] Leonidas G Alexopoulos, Julio Saez-Rodriguez, Benjamin D Cosgrove, Douglas A Lauffenburger, and Peter K Sorger. Networks inferred from biochemical data reveal profound differences in toll-like receptor and inflammatory signaling between normal and transformed hepatocytes. *Molecular cellular proteomics MCP*, 9(9):1849–1865, 2010.

[17] Mariano J. Alvarez, Federico Giorgi, and Andrea Califano. Using viper, a package for virtual inference of protein-activity by enriched regulon analysis. 2014.

[18] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[19] Alvaro Aytes, Antonina Mitrofanova, Celine Lefebvre, Mariano J Alvarez, Mireia Castillo-Martin, Tian Zheng, James A Eastham, Anuradha Gopalan, Kenneth J Pienta, Michael M Shen, et al. Cross-species regulatory network analysis identifies a synergistic interaction between foxm1 and cenpf that drives prostate cancer malignancy. *Cancer cell*, 25(5):638–651, 2014.

[20] Mohammad Taha Bahadori and Yan Liu. Granger causality analysis in irregular time series. In *SDM*, pages 660–671. SIAM, 2012.

[21] O. A. Balbin, J. R. Prensner, A. Sahu, A. Yocum, S. Shankar, R. Malik, D. Fermin, S. M. Dhanasekaran, B. Chandler, D. Thomas, D. G. Beer, X. Cao, A. I.

Nesvizhskii, and A. M. Chinnaiyan. Reconstructing targetable pathways in lung cancer by integrating diverse omics data. *Nat Commun*, 4:2617, 2013.

[22] Shantanu Banerji, Kristian Cibulskis, Claudia Rangel-Escareno, Kristin K Brown, Scott L Carter, Abbie M Frederick, Michael S Lawrence, Andrey Y Sivachenko, Carrie Sougnez, Lihua Zou, and et al. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, 486(7403):405–409, 2012.

[23] Shantanu Banerji, Kristian Cibulskis, Claudia Rangel-Escareno, Kristin K Brown, Scott L Carter, Abbie M Frederick, Michael S Lawrence, Andrey Y Sivachenko, Carrie Sougnez, Lihua Zou, and et al. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, 486(7403):405–409, 2012.

[24] Albert-László Barabási et al. Scale-free networks: a decade and beyond. *science*, 325(5939):412, 2009.

[25] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

[26] Stephen B Baylin and Peter A Jones. A decade of exploring the cancer epigenome-biological and translational implications. *Nature Reviews Cancer*, 11(10):726–734, 2011.

[27] Daniela Beisser, Gunnar W. Klau, Thomas Dandekar, Tobias Muller, and Marcus T. Dittrich. Bionet: an r-package for the functional analysis of biological networks. *Bioinformatics*, February 2010.

[28] E. S. Boja and H. Rodriguez. Proteogenomic convergence for understanding cancer pathways and networks. *Clin Proteomics*, 11(1):22, 2014.

[29] Daniel Branton, David W Deamer, Andre Marziali, Hagan Bayley, Steven A Benner, Thomas Butler, Massimiliano Di Ventra, Slaven Garaj, Andrew Hibbs, Xiaohua Huang, and et al. The potential and challenges of nanopore sequencing. *Nature Biotechnology*, 26(10):1146–1153, 2008.

[30] Cameron W Brennan, Roel GW Verhaak, Aaron McKenna, Benito Campos, Houtan Noushmehr, Sofie R Salama, Siyuan Zheng, Debyani Chakravarty, J Zachary Sanborn, Samuel H Berman, et al. The somatic genomic landscape of glioblastoma. *Cell*, 155(2):462–477, 2013.

[31] Network Cancer Genome Atlas Research. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–15, 2011.

[32] P. Casado, J. C. Rodriguez-Prados, S. C. Cosulich, S. Guichard, B. Vanhaesebroeck, S. Joel, and P. R. Cutillas. Kinase-substrate enrichment analysis provides insights into the heterogeneity of signaling pathway activation in leukemia cells. *Sci Signal*, 6(268):rs6, 2013.

[33] Ethan G Cerami, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Özgün

Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander. Pathway commons, a web resource for biological pathway data. *Nucleic acids research*, 39(suppl 1):D685–D690, 2011.

[34] Soumen Chakrabarti. Dynamic personalized pagerank in entity-relation graphs. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 571–580, New York, NY, USA, 2007. ACM.

[35] C. J. Chang, J. Y. Yang, W. Xia, C. T. Chen, X. Xie, C. H. Chao, W. A. Woodward, J. M. Hsu, G. N. Hortobagyi, and M. C. Hung. Ezh2 promotes expansion of breast tumor initiating cells through activation of raf1-beta-catenin signaling. *Cancer Cell*, 19(1):86–100, 2011.

[36] J. C. Chen, M. J. Alvarez, F. Talos, H. Dhruv, G. E. Rieckhof, A. Iyer, K. L. Diefes, K. Aldape, M. Berens, M. M. Shen, and A. Califano. Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. *Cell*, 159(2):402–14, 2014.

[37] James C Chen, Mariano J Alvarez, Flaminia Talos, Harshil Dhruv, Gabrielle E Rieckhof, Archana Iyer, Kristin L Diefes, Kenneth Aldape, Michael Berens, Michael M Shen, et al. Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks. *Cell*, 159(2):402–414, 2014.

[38] A. S. Cheng, S. S. Lau, Y. Chen, Y. Kondo, M. S. Li, H. Feng, A. K. Ching,

K. F. Cheung, H. K. Wong, J. H. Tong, H. Jin, K. W. Choy, J. Yu, K. F. To, N. Wong, T. H. Huang, and J. J. Sung. Ezh2-mediated concordant repression of wnt antagonists promotes beta-catenin-dependent hepatocarcinogenesis. *Cancer Res*, 71(11):4028–39, 2011.

[39] Hiu Wing Cheung, Glenn S Cowley, Barbara A Weir, Jesse S Boehm, Scott Rusin, Justine A Scott, Alexandra East, Levi D Ali, Patrick H Lizotte, Terence C Wong, et al. Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proceedings of the National Academy of Sciences*, 108(30):12372–12377, 2011.

[40] Lynda Chin, William C Hahn, Gad Getz, and Matthew Meyerson. Making sense of cancer genomic data. *Genes & development*, 25(6):534–555, 2011.

[41] Gilbert Chu, Balasubramanian Narasimhan, Robert Tibshirani, and Virginia Tusher. Significance analysis of microarrays (sam) software. *Nature*, 5:436–442, 2002.

[42] Fan Chung. The heat kernel as the pagerank of a graph. *Proceedings of the National Academy of Sciences*, 104(50):19735–19740, 2007.

[43] Fan Chung. A local graph partitioning algorithm using heat kernel pagerank. *Internet Mathematics*, 6(3):315–330, 2009.

[44] Giovanni Ciriello, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Mu-

tual exclusivity analysis identifies oncogenic network modules. *Genome research*, 22(2):398–406, 2012.

[45] Giovanni Ciriello, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Research*, 22(2):398–406, 2012.

[46] Giovanni Ciriello, Martin L Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. Emerging landscape of oncogenic signatures across human cancers. *Nature genetics*, 45(10):1127–1133, 2013.

[47] Melissa S Cline, Michael Smoot, Ethan Cerami, Allan Kuchinsky, Nerius Landys, Chris Workman, Rowan Christmas, Iliana Avila-Campilo, Michael Creech, Benjamin Gross, and et al. Integration of biological networks and gene expression data using cytoscape. *Nature Protocols*, 2(10):2366–2382, 2007.

[48] Thomas H. Cormen, Clifford Stein, Ronald L. Rivest, and Charles E. Leiserson. *Introduction to Algorithms*. McGraw-Hill Higher Education, 2nd edition, 2001.

[49] Michael Costanzo, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D Spear, Carolyn S Sevier, Huiming Ding, Judice L Y Koh, Kiana Toufighi, Sara Mostafavi, and et al. The genetic landscape of a cell. *Science*, 327(5964):425–31, 2010.

[50] James C Costello, Laura M Heiser, Elisabeth Georgii, Mehmet Gönen, Michael P Menden, Nicholas J Wang, Mukesh Bansal, Petteri Hintsanen, Suleiman A Khan,

John-Patrick Mpindi, et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature biotechnology*, 32(12):1202–1212, 2014.

[51] Kevin D Courtney, Ryan B Corcoran, and Jeffrey A Engelman. The pi3k pathway as drug target in human cancer. *Journal of Clinical Oncology*, 28(6):1075–1083, 2010.

[52] J. Cox and M. Mann. Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*, 26(12):1367–72, 2008.

[53] Chad Creighton and Samir Hanash. Mining gene expression databases for association rules. *Bioinformatics*, 19(1):79–86, 2003.

[54] A. S. Crystal, A. T. Shaw, L. V. Sequist, L. Friboulet, M. J. Niederst, E. L. Lockerman, R. L. Frias, J. F. Gainor, A. Amzallag, P. Greninger, D. Lee, A. Kalsy, M. Gomez-Caraballo, L. Elamine, E. Howe, W. Hur, E. Lifshits, H. E. Robinson, R. Katayama, A. C. Faber, M. M. Awad, S. Ramaswamy, M. Mino-Kenudson, A. J. Iafrate, C. H. Benes, and J. A. Engelman. Patient-derived models of acquired resistance can identify effective drug combinations for cancer. *Science*, 346(6216):1480–6, 2014.

[55] Kiley Graim Chris Wong Adrian Bivol Peter Ryabinin Kyle Ellrott Joshua M. Stuart* Artem Sokolov Daniel E. Carlin, Evan O. Paull. The prophetic granger

causality method to infer gene regulatory networks. *Nature Scientific Reports (submitted)*, 2015.

[56] MA Davies and Y Samuels. Analysis of the genome to personalize therapy for melanoma. *Oncogene*, 29(41):5545–5555, 2010.

[57] J. S. de Bono, C. J. Logothetis, A. Molina, K. Fizazi, S. North, L. Chu, K. N. Chi, R. J. Jones, Jr. Goodman, O. B., F. Saad, J. N. Staffurth, P. Mainwaring, S. Harland, T. W. Flaig, T. E. Hutson, T. Cheng, H. Patterson, J. D. Hainsworth, C. J. Ryan, C. N. Sternberg, S. L. Ellard, A. Flechon, M. Saleh, M. Scholz, E. Efstathiou, A. Zivi, D. Bianchini, Y. Loriot, N. Chieffo, T. Kheoh, C. M. Haqq, and H. I. Scher. Abiraterone and increased survival in metastatic prostate cancer. *N Engl J Med*, 364(21):1995–2005, 2011.

[58] S. J. Deeb, R. C. D'Souza, J. Cox, M. Schmidt-Supprian, and M. Mann. Super-silac allows classification of diffuse large b-cell lymphoma subtypes by their protein expression profiles. *Mol Cell Proteomics*, 11(5):77–89, 2012.

[59] Nathan D Dees, Qunyuan Zhang, Cyriac Kandoth, Michael C Wendl, William Schierding, Daniel C Koboldt, Thomas B Mooney, Matthew B Callaway, David Dooling, Elaine R Mardis, et al. Music: identifying mutational significance in cancer genomes. *Genome research*, 22(8):1589–1598, 2012.

[60] AS Dhillon, S Hagan, O Rath, and W Kolch. Map kinase signalling pathways in cancer. *Oncogene*, 26(22):3279–3290, 2007.

[61] Marcus T. Dittrich, Gunnar W. Klau, Andreas Rosenwald, Thomas Dandekar, and Tobias Muller. Identifying functional modules in ppi networks: an integrated exact approach. *Bioinformatics*, 24 ISMB, 2008.

[62] Sorin Draghici, Purvesh Khatri, Adi Laurentiu Tarca, Kashyap Amin, Arina Done, Calin Voichita, Constantin Georgescu, and Roberto Romero. A systems biology approach for pathway level analysis. *Genome Research*, 17(10):1537–1545, 2007.

[63] J. M. Drake, N. A. Graham, J. K. Lee, T. Stoyanova, C. M. Faltermeier, S. Sud, B. Titz, J. Huang, K. J. Pienta, T. G. Graeber, and O. N. Witte. Metastatic castration-resistant prostate cancer reveals intrapatient similarity and interpatient heterogeneity of therapeutic kinase targets. *Proc Natl Acad Sci U S A*, 110(49):E4762–9, 2013.

[64] J. M. Drake, N. A. Graham, T. Stoyanova, A. Sedghi, A. S. Goldstein, H. Cai, D. A. Smith, H. Zhang, E. Komisopoulou, J. Huang, T. G. Graeber, and O. N. Witte. Oncogene-specific activation of tyrosine kinase networks during prostate cancer progression. *Proc Natl Acad Sci U S A*, 109(5):1643–8, 2012.

[65] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–8, 1998.

[66] Jonathan D Ellis, Miriam Barrios-Rodiles, Recep Çolak, Manuel Irimia, TaeHyung Kim, John A Calarco, Xinchen Wang, Qun Pan, Dave O'Hanlon, Philip M Kim,

et al. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Molecular cell*, 46(6):884–892, 2012.

[67] Rebecca Elston and Gareth J Inman. Crosstalk between p53 and tgf- signalling. *Journal of signal transduction*, 2012:294097.

[68] Manel Esteller. Aberrant dna methylation as a cancer-inducing mechanism. *Annu. Rev. Pharmacol. Toxicol.*, 45:629–656, 2005.

[69] Eric R Fearon and Bert Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767, 1990.

[70] Simon A Forbes, Nidhi Bindal, Sally Bamford, Charlotte Cole, Chai Yin Kok, David Beare, Mingming Jia, Rebecca Shepherd, Kenric Leung, Andrew Menzies, et al. Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic acids research*, page gkq929, 2010.

[71] Mario F Fraga, Esteban Ballestar, Ana Villar-Garea, Manuel Boix-Chornet, Jesus Espada, Gunnar Schotta, Tiziana Bonaldi, Claire Haydon, Santiago Ropero, Kevin Petrie, et al. Loss of acetylation at lys16 and trimethylation at lys20 of histone h4 is a common hallmark of human cancer. *Nature genetics*, 37(4):391–400, 2005.

[72] Nir Friedman. *The Bayesian structural EM algorithm*, volume 98, pages 129–138. Citeseer, 1998.

[73] Olivier Gevaert, Robert Tibshirani, and Sylvia K Plevritis. Pancancer analysis of dna methylation-driven genes using methylmix. *Genome biology*, 16(1):17, 2015.

[74] D. Gioeli, J. W. Mandell, G. R. Petroni, Jr. Frierson, H. F., and M. J. Weber. Activation of mitogen-activated protein kinase associated with prostate cancer progression. *Cancer Res*, 59(2):279–84, 1999.

[75] M. E. Gonzalez, M. L. DuPrie, H. Krueger, S. D. Merajver, A. C. Ventura, K. A. Toy, and C. G. Kleer. Histone methyltransferase ezh2 induces akt-dependent genomic instability and brca1 inhibition in breast cancer. *Cancer Res*, 71(6):2360–70, 2011.

[76] J. F. Goodwin, V. Kothari, J. M. Drake, S. Zhao, E. Dylgjeri, J. L. Dean, M. J. Schiewer, C. McNair, J. K. Jones, A. Aytes, M. S. Magee, A. E. Snook, Z. Zhu, R. B. Den, R. C. Birbe, L. G. Gomella, N. A. Graham, A. A. Vashisht, J. A. Wohlschlegel, T. G. Graeber, R. J. Karnes, M. Takhar, E. Davicioni, S. A. Tomlins, C. Abate-Shen, N. Sharifi, O. N. Witte, F. Y. Feng, and K. E. Knudsen. Dna-pkcs-mediated transcriptional regulation drives prostate cancer progression and metastasis. *Cancer Cell*, 28(1):97–113, 2015.

[77] C. S. Grasso, Y. M. Wu, D. R. Robinson, X. Cao, S. M. Dhanasekaran, A. P. Khan, M. J. Quist, X. Jing, R. J. Lonigro, J. C. Brenner, I. A. Asangani, B. Ateeq, S. Y. Chun, J. Siddiqui, L. Sam, M. Anstett, R. Mehra, J. R. Prensner, N. Palanisamy, G. A. Ryslik, F. Vandin, B. J. Raphael, L. P. Kunju, D. R. Rhodes, K. J.

Pienta, A. M. Chinnaiyan, and S. A. Tomlins. The mutational landscape of lethal castration-resistant prostate cancer. *Nature*, 487(7406):239–43, 2012.

[78] Catherine S Grasso, Yi-Mi Wu, Dan R Robinson, Xuhong Cao, Saravana M Dhanasekaran, Amjad P Khan, Michael J Quist, Xiaojun Jing, Robert J Lonigro, J Chad Brenner, et al. The mutational landscape of lethal castration-resistant prostate cancer. *Nature*, 487(7406):239–243, 2012.

[79] Michele Grieco, Massimo Santoro, Maria Teresa Berlingieri, Rosa Marina Melillo, Rosangela Donghi, Italia Bongarzone, Marco A Pierotti, Giuseppe Della Ports, Alfredo Fusco, and Giancarlo Vecchiot. Ptc is a novel rearranged form of the ret proto-oncogene and is frequently detected in vivo in human thyroid papillary carcinomas. *Cell*, 60(4):557–563, 1990.

[80] Marlous J Groenewoud and Fried JT Zwartkruis. Rheb and mammalian target of rapamycin in mitochondrial homoeostasis. *Open biology*, 3(12):130185, 2013.

[81] Mads Grønborg, Troels Zakarias Kristiansen, Allan Stensballe, Jens S Andersen, Osamu Ohara, Matthias Mann, Ole Nørregaard Jensen, and Akhilesh Pandey. A mass spectrometry-based proteomic approach for identification of serine/threonine-phosphorylated proteins by enrichment with phospho-specific antibodies identification of a novel protein, frigg, as a protein kinase a substrate. *Molecular & Cellular Proteomics*, 1(7):517–527, 2002.

[82] Marina A Guvakova and Ewa Surmacz. The activated insulin-like growth factor

i receptor induces depolarization in breast epithelial cells characterized by actin filament disassembly and tyrosine dephosphorylation of fak, cas, and paxillin. *Experimental cell research*, 251(1):244–255, 1999.

[83] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Laboratory (LANL), 2008.

[84] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. *Proccedings of the 7th Python in Science Conference (SciPy 2008)*, pages 11–15, 2008.

[85] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[86] Taher Haveliwala, Sepandar Kamvar, Dan Klein, Chris Manning, and Gene Golub. Computing pagerank using power extrapolation. Technical Report 2003-45, Stanford InfoLab, 2003.

[87] Taher H Haveliwala. Topic-sensitive pagerank. *Proceedings of the eleventh international conference on World Wide Web WWW 02*, 20(650):517, 2002.

[88] Nathaniel D Heintzman, Gary C Hon, R David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F Harp, Zhen Ye, Leonard K Lee, Rhona K Stuart, Christina W Ching, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, 2009.

[89] Laura M Heiser, Anguraj Sadanandam, Wen-Lin Kuo, Stephen C Benz, Theodore C Goldstein, Sam Ng, William J Gibb, Nicholas J Wang, Safiyyah Ziyad, Frances Tong, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proceedings of the National Academy of Sciences*, 109(8):2724–2729, 2012.

[90] James G Herman and Stephen B Baylin. Gene silencing in cancer in association with promoter hypermethylation. *New England Journal of Medicine*, 349(21):2042–2054, 2003.

[91] Maureen Heymans and Ambuj K Singh. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, 19 Suppl 1(90001):i138–i146, 2003.

[92] Maureen E Hillenmeyer. The chemical genomic portrait of yeast. *World Health*, 362(2008):362–5, 2008.

[93] Maureen E Hillenmeyer. The chemical genomic portrait of yeast. *World Health*, 362(2008):362–5, 2008.

[94] Katherine A Hoadley, Christina Yau, Denise M Wolf, Andrew D Cherniack, David Tamborero, Sam Ng, Max DM Leiserson, Beifang Niu, Michael D McLellan, Vladislav Uzunangelov, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, 2014.

[95] Matan Hofree, John P Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature methods*, 10(11):1108–1115, 2013.

[96] Da Wei Huang, Brad T Sherman, Qina Tan, Joseph Kir, David Liu, David Bryant, Yongjian Guo, Robert Stephens, Michael W Baseler, H Clifford Lane, et al. David bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic acids research*, 35(suppl 2):W169–W175, 2007.

[97] S. S. Huang, D. C. Clarke, S. J. Gosline, A. Labadorf, C. R. Chouinard, W. Gordon, D. A. Lauffenburger, and E. Fraenkel. Linking proteomic and transcriptional data through the interactome and epigenome reveals a map of oncogene-induced signaling. *PLoS Comput Biol*, 9(2):e1002887, 2013.

[98] T R Hughes, M J Marton, A R Jones, C J Roberts, R Stoughton, C D Armour, H A Bennett, E Coffey, H Dai, Y D He, and et al. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2000.

[99] T R Hughes, M J Marton, A R Jones, C J Roberts, R Stoughton, C D Armour, H A Bennett, E Coffey, H Dai, Y D He, and et al. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2000.

[100] T. Hunter. Tyrosine phosphorylation: thirty years and counting. *Curr Opin Cell Biol*, 21(2):140–6, 2009.

[101] P T Iau, R D Macmillan, and R W Blamey. Germ line mutations associated with breast cancer susceptibility. *European Journal of Cancer*, 37(3):300–321, 2001.

[102] Trey Ideker, Owen Ozier, Benno Schwikowski, and Andrew F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18(suppl 1):S233–S240, 2002.

[103] Nathan T Ihle, Robert Lemos, Peter Wipf, Adly Yacoub, Clint Mitchell, Doris Siwak, Gordon B Mills, Paul Dent, D Lynn Kirkpatrick, and Garth Powis. Mutations in the phosphatidylinositol-3-kinase pathway predict for antitumor activity of the inhibitor px-866 whereas oncogenic ras is a dominant predictor for resistance. *Cancer Research*, 69(1):143–150, 2009.

[104] Bing-Hua Jiang and Ling-Zhi Liu. Pi3k/pten signaling in angiogenesis and tumorigenesis. *Advances in cancer research*, 102:19–65, 2009.

[105] G Joshi-Tope, Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal, GR Gopinath, GR Wu, Lisa Matthews, et al. Reactome: a knowledgebase of biological pathways. *Nucleic acids research*, 33(suppl 1):D428–D432, 2005.

[106] C. Kandoth, M. D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J. F. McMichael, M. A. Wyczalkowski, M. D. Leiserson, C. A. Miller, J. S. Welch, M. J. Walter, M. C. Wendl, T. J. Ley, R. K. Wilson, B. J. Raphael, and L. Ding.

Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–9, 2013.

[107] Thomas Kelder, Martijn P van Iersel, Kristina Hanspers, Martina Kutmon, Bruce R Conklin, Chris T Evelo, and Alexander R Pico. Wikipathways: building research communities on biological pathways. *Nucleic acids research*, 40(D1):D1301–D1307, 2012.

[108] C. D. Kelstrup, C. Young, R. Lavallee, M. L. Nielsen, and J. V. Olsen. Optimized fast and sensitive acquisition methods for shotgun proteomics on a quadrupole orbitrap mass spectrometer. *J Proteome Res*, 11(6):3487–97, 2012.

[109] E. Khurana, Y. Fu, J. Chen, and M. Gerstein. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol*, 9(3):e1002886, 2013.

[110] Ekta Khurana, Yao Fu, Jieming Chen, and Mark Gerstein. Interpretation of genomic variants using a unified biological network approach. *PLoS Comput Biol*, 9(3):e1002886, 2013.

[111] Edna T Kimura, Marina N Nikiforova, Zhaowen Zhu, Jeffrey A Knauf, Yuri E Nikiforov, and James A Fagin. High prevalence of braf mutations in thyroid cancer genetic evidence for constitutive activation of the ret/ptc-ras-braf signaling pathway in papillary thyroid carcinoma. *Cancer Research*, 63(7):1454–1457, 2003.

[112] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*, volume 2009. MIT Press.

[113] Risi Imre Kondor and John Lafferty. Diffusion kernels on graphs and other discrete input spaces. *Machine-Learning International*, 2002.

[114] Robert L Kortum and Robert E Lewis. The molecular scaffold ksr1 regulates the proliferative and oncogenic potential of cells. *Molecular and Cellular Biology*, 24(10):4407–4416, 2004.

[115] A. Lachmann and A. Ma'ayan. Kea: kinase enrichment analysis. *Bioinformatics*, 25(5):684–6, 2009.

[116] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-philippe Brunet, Aravind Subramanian, Kenneth N Ross, and et al. The connectivity map : Using. *Science*, 1929(2006):1929–35, 2006.

[117] Alex Lan, Ilan Y Smoly, Guy Rapaport, Susan Lindquist, Ernest Fraenkel, and Esti Yeger-Lotem. Responsenet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic Acids Research*, 39(Web Server issue):W424–W429, 2011.

[118] J. Larkin, P. A. Ascierto, B. Dreno, V. Atkinson, G. Liszkay, M. Maio, M. Mandala, L. Demidov, D. Stroyakovskiy, L. Thomas, L. de la Cruz-Merino, C. Dutriaux, C. Garbe, M. A. Sovak, I. Chang, N. Choong, S. P. Hack, G. A. McArthur,

and A. Ribas. Combined vemurafenib and cobimetinib in braf-mutated melanoma. *N Engl J Med*, 371(20):1867–76, 2014.

[119] Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013.

[120] Celine Lefebvre, Presha Rajbhandari, Mariano J Alvarez, Pradeep Bandaru, Wei Keat Lim, Mai Sato, Kai Wang, Pavel Sumazin, Manjunath Kustagi, Brygida C Bisikirska, et al. A human b-cell interactome identifies myb and foxm1 as master regulators of proliferation in germinal centers. *Molecular systems biology*, 6(1), 2010.

[121] H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, and R Durbin. Sam format and samtools. *Processing*, 25(October):2078–2079, 2009.

[122] W. Li, N. Ai, S. Wang, N. Bhattacharya, V. Vrbanac, M. Collins, S. Signoretti, Y. Hu, F. M. Boyce, K. Gravdal, O. J. Halvorsen, H. Nalwoga, L. A. Akslen, E. Harlow, and R. S. Watnick. Grk3 is essential for metastatic cells and promotes prostate tumor progression. *Proc Natl Acad Sci U S A*, 111(4):1521–6, 2014.

[123] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[124] Wei Keat Lim, Eugenia Lyashenko, and Andrea Califano. Master regulators used as breast cancer metastasis classifier. *Pacific Symposium on Biocomputing*, 14:504–515, 2009.

[125] A. M. Lin, B. I. Rini, V. Weinberg, K. Fong, C. J. Ryan, J. E. Rosenberg, L. Fong, and E. J. Small. A phase ii trial of imatinib mesylate in patients with biochemical relapse of prostate cancer after definitive local therapy. *BJU Int*, 98(4):763–9, 2006.

[126] L. Liu, P. A. Mayes, S. Eastman, H. Shi, S. Yadavilli, T. Zhang, J. Yang, L. Seestaller-Wehr, S. Y. Zhang, C. Hopson, L. Tsvetkov, J. Jing, S. Zhang, J. Smothers, and A. Hoos. The braf and mek inhibitors dabrafenib and trametinib: Effects on immune function and in combination with immunomodulatory antibodies targeting pd-1, pd-l1, and ctla-4. *Clin Cancer Res*, 2015.

[127] Ivana Ljubi, Ren Weiskircher, Ulrich Pferschy, Gunnar W. Klau, Petra Mutzel, and Matteo Fischetti. *An Algorithmic Framework for the Exact Solution of the Prize-Collecting Steiner Tree Problem (Mathematical Programming Series)*, volume 105.

[128] DB Longley and PG Johnston. Molecular mechanisms of drug resistance. *The Journal of Pathology*, 205(2):275–292, 2005.

[129] José Lozano, Rosie Xing, Zhenzi Cai, Heather L Jensen, Carol Trempus, Willie Mark, Ron Cannon, and Richard Kolesnick. Deficiency of kinase suppressor of

ras1 prevents oncogenic ras signaling in mice. *Cancer research*, 63(14):4232–4238, 2003.

[130] S. Lu, S. Y. Tsai, and M. J. Tsai. Regulation of androgen-dependent prostatic cancer cell growth: androgen regulation of cdk2, cdk4, and cki p16 genes. *Cancer Res*, 57(20):4511–6, 1997.

[131] Ji Luo, Michael J Emanuele, Danan Li, Chad J Creighton, Michael R Schlabach, Thomas F Westbrook, Kwok-Kin Wong, and Stephen J Elledge. A genome-wide rnai screen identifies multiple synthetic lethal interactions with the ras oncogene. *Cell*, 137(5):835–848, 2009.

[132] Ji Luo, Nicole L Solimini, and Stephen J Elledge. Principles of cancer therapy: oncogene and non-oncogene addiction. *Cell*, 136(5):823–837, 2009.

[133] Aviv Madar, Alex Greenfield, Eric Vanden-Eijnden, and Richard Bonneau. Dream3: network inference using dynamic context likelihood of relatedness and the inferelator. *PloS one*, 5(3):e9803, 2010.

[134] Barbara Mair, Stefan Kubicek, and Sebastian Nijman. Exploiting epigenetic vulnerabilities for cancer therapeutics. *Trends in pharmacological sciences*, 35(3):136–145, 2014.

[135] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–34, 2002.

[136] Heiko A Mannsperger, Stefan Uhlmann, Christian Schmidt, Stefan Wiemann, zgr Sahin, and Ulrike Korf. Rnai-based validation of antibodies for reverse phase protein arrays. *Proteome Science*, 8(1):69, 2010.

[137] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo D Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*, 7(Suppl 1):S7, 2006.

[138] David L Masica and Rachel Karchin. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Research*, 71(13):4550–4561, 2011.

[139] Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, and Gad Getz. Gistic2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12(4):R41, 2011.

[140] Craig H Mermel, Steven E Schumacher, Barbara Hill, Matthew L Meyerson, Rameen Beroukhim, Gad Getz, et al. Gistic2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*, 12(4):R41, 2011.

[141] A. Michalski, E. Damoc, J. P. Hauschild, O. Lange, A. Wieghaus, A. Makarov, N. Nagaraj, J. Cox, M. Mann, and S. Horning. Mass spectrometry-based pro-

teomics using q exactive, a high-performance benchtop quadrupole orbitrap mass spectrometer. *Mol Cell Proteomics*, 10(9):M111 011015, 2011.

[142] Bertrand Mjm. Hanahan d and weinberg ra. the hallmarks of cancer. cell 100: 57-70, 2000. *Annals of Oncology*, page 2008, 2009.

[143] Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *Society for Industrial and Applied Mathematics (Review)*, 45:3–, 2003.

[144] Gemma Molyneux, Felipe C Geyer, Fiona-Ann Magnay, Afshan McCarthy, Howard Kendrick, Rachael Natrajan, Alan MacKay, Anita Grigoriadis, Andrew Tutt, Alan Ashworth, et al. ¡ i¿ brca1¡/i¿ basal-like breast cancers originate from luminal epithelial progenitors and not from basal stem cells. *Cell stem cell*, 7(3):403–417, 2010.

[145] Joris Mooij. libdai: A free and open source c++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, 11:2169–2173, 2010.

[146] P. J. Morin, A. B. Sparks, V. Korinek, N. Barker, H. Clevers, B. Vogelstein, and K. W. Kinzler. Activation of beta-catenin-tcf signaling in colon cancer by mutations in beta-catenin or apc. *Science*, 275(5307):1787–90, 1997.

[147] Chris J Needham, James R Bradford, Andrew J Bulpitt, and David R West-

head. A primer on learning in bayesian networks for computational biology. *PLoS Computational Biology*, 3(8):8, 2007.

[148] Cancer Genome Atlas Network et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, 2012.

[149] Cancer Genome Atlas Network et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[150] Cancer Genome Atlas Research Network et al. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456):43–49, 2013.

[151] Cancer Genome Atlas Research Network et al. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315–322, 2014.

[152] Cancer Genome Atlas Research Network et al. Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, 159(3):676–690, 2014.

[153] The Cancer Genome Atlas Research Network*. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455, October 2008.

[154] Joseph R Nevins and Anil Potti. Mining gene expression profiles: expression signatures as cancer phenotypes. *Nature Reviews Genetics*, 8(8):601–609, 2007.

[155] R. H. Newman, J. Hu, H. S. Rho, Z. Xie, C. Woodard, J. Neiswinger, C. Cooper, M. Shirley, H. M. Clark, S. Hu, W. Hwang, J. S. Jeong, G. Wu, J. Lin, X. Gao,

Q. Ni, R. Goel, S. Xia, H. Ji, K. N. Dalby, M. J. Birnbaum, P. A. Cole, S. Knapp, A. G. Ryazanov, D. J. Zack, S. Blackshaw, T. Pawson, A. C. Gingras, S. Desiderio, A. Pandey, B. E. Turk, J. Zhang, H. Zhu, and J. Qian. Construction of human activity-based phosphorylation networks. *Mol Syst Biol*, 9:655, 2013.

[156] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003.

[157] Pauline C Ng and Ewen F Kirkness. Whole genome sequencing. *Methods In Molecular Biology Clifton Nj*, 628:215–226, 2010.

[158] Sam Ng, Eric A Collisson, Artem Sokolov, Theodore Goldstein, Abel Gonzalez-Perez, Nuria Lopez-Bigas, Christopher Benz, David Haussler, and Joshua M Stuart. Paradigm-shift predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics*, 28(18):i640–i646, 2012.

[159] X. Ning, Z. Shi, X. Liu, A. Zhang, L. Han, K. Jiang, C. Kang, and Q. Zhang. Dnmt1 and ezh2 mediated methylation silences the microrna-200b/a/429 gene and promotes tumor progression. *Cancer Lett*, 359(2):198–205, 2015.

[160] J. V. Olsen, B. Blagoev, F. Gnad, B. Macek, C. Kumar, P. Mortensen, and M. Mann. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, 127(3):635–48, 2006.

[161] Larsson Omberg, Kyle Ellrott, Yuan Yuan, Cyriac Kandoth, Chris Wong, Michael R Kellen, Stephen H Friend, Josh Stuart, Han Liang, and Adam A Mar-

golin. Enabling transparent and collaborative computational analysis of 12 tumor types within the cancer genome atlas. *Nature genetics*, 45(10):1121–1126, 2013.

[162] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[163] M. Pandey, S. Sahay, P. Tiwari, D. S. Upadhyay, S. Sultana, and K. P. Gupta. Involvement of ezh2, suv39h1, g9a and associated molecules in pathogenesis of urethane induced mouse lung tumors: potential targets for cancer control. *Toxicol Appl Pharmacol*, 280(2):296–304, 2014.

[164] H. U. Park, S. Suy, M. Danner, V. Dailey, Y. Zhang, H. Li, D. R. Hyduke, B. T. Collins, G. Gagnon, B. Kallakury, D. Kumar, M. L. Brown, A. Fornace, A. Dritschilo, and S. P. Collins. Amp-activated protein kinase promotes human prostate cancer cell growth and survival. *Mol Cancer Ther*, 8(4):733–41, 2009.

[165] Evan O Paull, Daniel E Carlin, Mario Niepel, Peter K Sorger, David Haussler, and Joshua M Stuart. Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (tiedie). *Bioinformatics*, 29(21):2757–2764, 2013.

[166] T. Pawson. Specificity in signal transduction: from phosphotyrosine-sh2 domain interactions to complex cellular systems. *Cell*, 116(2):191–203, 2004.

[167] Tony Pawson and John D Scott. Signaling through scaffold, anchoring, and adaptor proteins. *Science*, 278(5346):2075–2080, 1997.

[168] C M Perou, T Srlie, M B Eisen, M Van De Rijn, S S Jeffrey, C A Rees, J R Pollack, D T Ross, H Johnsen, L A Akslen, and et al. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, 2000.

[169] J Pea. An improved bayesian structural em algorithm for learning bayesian networks for clustering. *Pattern Recognition Letters*, 21(8):779–786, 2000.

[170] MA Pierotti, I Bongarzone, MG Borrello, C Mariani, C Miranda, G Sozzi, and A Greco. Rearrangements of trk proto-oncogene in papillary thyroid carcinomas. *Journal of endocrinological investigation*, 18(2):130–133, 1995.

[171] Christine A Pratilas, Barry S Taylor, Qing Ye, Agnes Viale, Chris Sander, David B Solit, and Neal Rosen. V600ebraf is associated with disabled feedback inhibition of raf–mek signaling and elevated transcriptional output of the pathway. *Proceedings of the National Academy of Sciences*, 106(11):4519–4524, 2009.

[172] Yan Qi, Yasir Suhail, Yu-yi Lin, Jef D. Boeke, and Joel S. Bader. Finding friends and enemies in an enemies-only network: A graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Research*, 18(12):1991–2004, 2008.

[173] Suresh K Rayala, Poonam R Molli, and Rakesh Kumar. Nuclear p21-activated

kinase 1 in breast cancer packs off tamoxifen sensitivity. *Cancer research*, 66(12):5985–5988, 2006.

[174] Boris Reva, Yevgeniy Antipin, and Chris Sander. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research*, 39(17):e118–e118, 2011.

[175] C. Robert, B. Karaszewska, J. Schachter, P. Rutkowski, A. Mackiewicz, D. Stroiakovski, M. Lichinitser, R. Dummer, F. Grange, L. Mortier, V. Chiarion-Sileni, K. Drucis, I. Krajsova, A. Hauschild, P. Lorigan, P. Wolter, G. V. Long, K. Flaherty, P. Nathan, A. Ribas, A. M. Martin, P. Sun, W. Crist, J. Legos, S. D. Rubin, S. M. Little, and D. Schadendorf. Improved overall survival in melanoma with combined dabrafenib and trametinib. *N Engl J Med*, 372(1):30–9, 2015.

[176] Charles WM Roberts and Stuart H Orkin. The swi/snf complexchromatin and cancer. *Nature Reviews Cancer*, 4(2):133–142, 2004.

[177] PJ Roberts and CJ Der. Targeting the raf-mek-erk mitogen-activated protein kinase cascade for the treatment of cancer. *Oncogene*, 26(22):3291–3310, 2007.

[178] D. Robinson, E. M. Van Allen, Y. M. Wu, N. Schultz, R. J. Lonigro, J. M. Mosquera, B. Montgomery, M. E. Taplin, C. C. Pritchard, G. Attard, H. Beltran, W. Abida, R. K. Bradley, J. Vinson, X. Cao, P. Vats, L. P. Kunju, M. Hussain, F. Y. Feng, S. A. Tomlins, K. A. Cooney, D. C. Smith, C. Brennan, J. Siddiqui, R. Mehra, Y. Chen, D. E. Rathkopf, M. J. Morris, S. B. Solomon, J. C. Durack,

V. E. Reuter, A. Gopalan, J. Gao, M. Loda, R. T. Lis, M. Bowden, S. P. Balk, G. Gaviola, C. Sougnez, M. Gupta, E. Y. Yu, E. A. Mostaghel, H. H. Cheng, H. Mulcahy, L. D. True, S. R. Plymate, H. Dvinge, R. Ferraldeschi, P. Flohr, S. Miranda, Z. Zafeiriou, N. Tunariu, J. Mateo, R. Perez-Lopez, F. Demichelis, B. D. Robinson, M. Schiffman, D. M. Nanus, S. T. Tagawa, A. Sigaras, K. W. Eng, O. Elemento, A. Sboner, E. I. Heath, H. I. Scher, K. J. Pienta, P. Kantoff, J. S. de Bono, M. A. Rubin, P. S. Nelson, L. A. Garraway, C. L. Sawyers, and A. M. Chinnaiyan. Integrative clinical genomics of advanced prostate cancer. *Cell*, 161(5):1215–28, 2015.

[179] Dan Robinson, Eliezer M Van Allen, Yi-Mi Wu, Nikolaus Schultz, Robert J Lonigro, Juan-Miguel Mosquera, Bruce Montgomery, Mary-Ellen Taplin, Colin C Pritchard, Gerhardt Attard, et al. Integrative clinical genomics of advanced prostate cancer. *Cell*, 161(5):1215–1228, 2015.

[180] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[181] C F Rochlitz, G K Scott, J M Dodson, E Liu, C Dollbaum, H S Smith, and C C Benz. Incidence of activating ras oncogene mutations associated with primary and metastatic human breast cancer. *Cancer Research*, 49(2):357–360, 1989.

[182] M. A. Rubin, M. Putzi, N. Mucci, D. C. Smith, K. Wojno, S. Korenchuk, and K. J.

Pienta. Rapid ("warm") autopsy study for procurement of metastatic prostate cancer. *Clin Cancer Res*, 6(3):1038–45, 2000.

[183] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[184] A. J. Saldanha. Java treeview–extensible visualization of microarray data. *Bioinformatics*, 20(17):3246–8, 2004.

[185] Carl F Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, and Kenneth H Buetow. Pid: the pathway interaction database. *Nucleic acids research*, 37(suppl 1):D674–D679, 2009.

[186] M Schena, D Shalon, R W Davis, and P O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[187] H. I. Scher, K. Fizazi, F. Saad, M. E. Taplin, C. N. Sternberg, K. Miller, R. de Wit, P. Mulders, K. N. Chi, N. D. Shore, A. J. Armstrong, T. W. Flaig, A. Flechon, P. Mainwaring, M. Fleming, J. D. Hainsworth, M. Hirmand, B. Selby, L. Seely, and J. S. de Bono. Increased survival with enzalutamide in prostate cancer after chemotherapy. *N Engl J Med*, 367(13):1187–97, 2012.

[188] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin,

B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–504, 2003.

[189] Carol Huang Shao-shan and Ernest Fraenkel. Integration of proteomic, transcriptional, and interactome data reveals hidden signaling components. *Sci Signal*, June 2010.

[190] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nature Biotechnology*, 26(10):1135–1145, 2008.

[191] B. Shi, J. Liang, X. Yang, Y. Wang, Y. Zhao, H. Wu, L. Sun, Y. Zhang, Y. Chen, R. Li, Y. Zhang, M. Hong, and Y. Shang. Integration of estrogen and wnt signaling circuits by the polycomb group protein ezh2 in breast cancer cells. *Mol Cell Biol*, 27(14):5105–19, 2007.

[192] Yigong Shi and Joan Massagu. Mechanisms of tgf-beta signaling from cell membrane to the nucleus. *Cell*, 113(6):685–700, 2003.

[193] Jose M Silva, Mamie Z Li, Ken Chang, Wei Ge, Michael C Golding, Richard J Rickles, Despina Siolas, Guang Hu, Patrick J Paddison, Michael R Schlabach, and et al. Second-generation shrna libraries covering the mouse and human genomes. *Nature Genetics*, 37(11):1281–1288, 2005.

[194] Ngak-Leng Sim, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C Ng. Sift web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*, pages 1–6, 2012.

[195] Paula Soares, Vítor Trovisco, Ana Sofia Rocha, Jorge Lima, Patrícia Castro, Ana Preto, Valdemar Maximo, Tiago Botelho, Raquel Seruca, and Manuel Sobrinho-Simoes. Braf mutations and ret/ptc rearrangements are alternative events in the etiopathogenesis of ptc. *Oncogene*, 22(29):4578–4580, 2003.

[196] Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf. Large scale multiple kernel learning. *The Journal of Machine Learning Research*, 7:1531–1565, 2006.

[197] Daniel E. Carlin Evan O. Paull Kiley Graim Chris Wong Adrian Bivol Peter Ryabinin Kyle Ellrott Joshua M. Stuart* Artem Sokolov HPN-DREAM Consortium (other authors TBA) Steven M. Hill, Laura M. Heiser. Empirical assessment of causal network learning through a community-based efforts. *Nature Methods (accepted)*, 2015.

[198] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and et al. 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*, 102(43):15545–15550, 2005.

[199] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. 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*, 102(43):15545–15550, 2005.

[200] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. 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*, 102(43):15545–15550, 2005.

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

[202] Barry S Taylor, Nikolaus Schultz, Haley Hieronymus, Anuradha Gopalan, Yonghong Xiao, Brett S Carver, Vivek K Arora, Poorvi Kaushik, Ethan Cerami, Boris Reva, et al. Integrative genomic profiling of human prostate cancer. *Cancer cell*, 18(1):11–22, 2010.

[203] John J Tentler, Aik Choon Tan, Colin D Weekes, Antonio Jimeno, Stephen Leong, Todd M Pitts, John J Arcaroli, Wells A Messersmith, and S Gail Eckhardt. Patient-derived tumour xenografts as models for oncology drug development. *Nature reviews Clinical oncology*, 9(6):338–350, 2012.

[204] John J Tentler, Aik Choon Tan, Colin D Weekes, Antonio Jimeno, Stephen Leong, Todd M Pitts, John J Arcaroli, Wells A Messersmith, and S Gail Eckhardt. Patient-derived tumour xenografts as models for oncology drug development. *Nature reviews Clinical oncology*, 9(6):338–350, 2012.

[205] Raoul Tibes, YiHua Qiu, Yiling Lu, Bryan Hennessy, Michael Andreeff, Gordon B. Mills, and Steven M. Kornblau. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Molecular Cancer Therapeutics*, 5(10):2512–2521, 2006.

[206] Cristian Tomasetti and Bert Vogelstein. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, 347(6217):78–81, 2015.

[207] Ryota Tomioka and Taiji Suzuki. Sparsity-accuracy trade-off in mkl. *arXiv preprint arXiv:1001.2615*, 2010.

[208] Nurcan Tuncbag, Alfredo Braunstein, Andrea Pagnani, Shao-Shan Carol Huang, Jennifer Chayes, Christian Borgs, Riccardo Zecchina, and Ernest Fraenkel. Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. *Journal of Computational Biology*, 20(2):124–136, 2013.

[209] Nicholas C Turner, Christopher J Lord, Elizabeth Iorns, Rachel Brough, Sally Swift, Richard Elliott, Syndonia Rayter, and Alan Tutt, Andrew N Ashworth.

A synthetic lethal sirna screen identifying genes mediating sensitivity to a parp inhibitor. *EMBO J*, 27:1368–1377, 2008.

[210] Tusher, Virginia Goss, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[211] P. W. Twardowski, J. H. Beumer, C. S. Chen, A. S. Kraft, G. S. Chatta, M. Mitsuhashi, W. Ye, S. M. Christner, and M. B. Lilly. A phase ii trial of dasatinib in patients with metastatic castration-resistant prostate cancer treated previously with chemotherapy. *Anticancer Drugs*, 24(7):743–53, 2013.

[212] Robert J. Vanderbei. Loqo: An interior point code for quadratic programming. Technical report, Optimization Methods and Software.

[213] Fabio Vandin, Patrick Clay, Eli Upfal, and Benjamin J. Raphael. Discovery of mutated subnetworks associated with clinical data in cancer. *Pacific Biosciences Symposium*, 2012.

[214] Fabio Vandin, Eli Upfal, and Benjamin J. Raphael. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, 18:507–522, 2011.

[215] Fabio Vandin, Eli Upfal, and Benjamin J Raphael. De novo discovery of mutated driver pathways in cancer. *Genome Research*, 22(2):375–85, 2011.

[216] Ignacio Varela, Patrick Tarpey, Keiran Raine, Dachuan Huang, Choon Kiat Ong,

Philip Stephens, Helen Davies, David Jones, Meng-Lay Lin, Jon Teague, et al. Exome sequencing identifies frequent mutation of the swi/snf complex gene pbrm1 in renal carcinoma. *Nature*, 469(7331):539–542, 2011.

[217] Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, David Haussler, and Joshua M Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. *Bioinformatics*, 26(12):i237–i245, 2010.

[218] J. A. Vizcaino, E. W. Deutsch, R. Wang, A. Csordas, F. Reisinger, D. Rios, J. A. Dianes, Z. Sun, T. Farrah, N. Bandeira, P. A. Binz, I. Xenarios, M. Eisenacher, G. Mayer, L. Gatto, A. Campos, R. J. Chalkley, H. J. Kraus, J. P. Albar, S. Martinez-Bartolome, R. Apweiler, G. S. Omenn, L. Martens, A. R. Jones, and H. Hermjakob. Proteomexchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol*, 32(3):223–6, 2014.

[219] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler. Cancer genome landscapes. *science*, 339(6127):1546–1558, 2013.

[220] G. Wang, K. A. Ahmad, G. Unger, J. W. Slaton, and K. Ahmed. Ck2 signaling in androgen-dependent and -independent prostate cancer. *J Cell Biochem*, 99(2):382–91, 2006.

[221] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[222] Laura D Wood, D Williams Parsons, Sin Jones, Jimmy Lin, Tobias Sjblom, Rebecca J Leary, Dong Shen, Simina M Boca, Thomas Barber, Janine Ptak, and et al. The genomic landscapes of human breast and colorectal cancers. *Science*, 318(5853):1108–1113, 2007.

[223] Zongbing You, Daniel Saims, Shaoqiong Chen, Zhaocheng Zhang, Denis C. Guttridge, Kun-liang Guan, Ormond A. MacDougald, Anthony M.C. Brown, Gerard Evan, Jan Kitajewski, and Cun-Yu Wang. Wnt signaling promotes oncogenic transformation by inhibiting c-myc-induced apoptosis. *J. Cell Biol.*, 157(3):429–440, 2002.

[224] G. Yu, Y. C. Lee, C. J. Cheng, C. F. Wu, J. H. Song, G. E. Gallick, L. Y. Yu-Lee, J. Kuang, and S. H. Lin. Rsk promotes prostate cancer progression in bone through ing3, ckap2, and ptk6-mediated cell survival. *Mol Cancer Res*, 13(2):348–57, 2015.

[225] Travis I Zack, Steven E Schumacher, Scott L Carter, Andrew D Cherniack, Gordon Saksena, Barbara Tabak, Michael S Lawrence, Cheng-Zhong Zhang, Jeremiah Wala, Craig H Mermel, et al. Pan-cancer patterns of somatic copy number alteration. *Nature genetics*, 45(10):1134–1140, 2013.

[226] Laura (Laura A.) Zager. Graph similarity and matching. 2005.

[227] B. Zhang, J. Wang, X. Wang, J. Zhu, Q. Liu, Z. Shi, M. C. Chambers, L. J. Zimmerman, K. F. Shaddox, S. Kim, S. R. Davies, S. Wang, P. Wang, C. R. Kinsinger, R. C. Rivers, H. Rodriguez, R. R. Townsend, M. J. Ellis, S. A. Carr, D. L. Tabb, R. J. Coffey, R. J. Slebos, D. C. Liebler, and Cptac Nci. Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513(7518):382–7, 2014.

[228] A. Zimman, S. S. Chen, E. Komisopoulou, B. Titz, R. Martinez-Pinna, A. Kafi, J. A. Berliner, and T. G. Graeber. Activation of aortic endothelial cells by oxidized phospholipids: a phosphoproteomic analysis. *J Proteome Res*, 9(6):2812–24, 2010.

Appendix A

Appendix

This chapter describes the technologies that I have used or are directly relevant to my thesis work and summarizes the prior methods that have inspired my research. This chapter can be used as either an introductory background to the thesis, or as an appendix that may be referred to while reading the later chapters.

A.1 Measuring Biology: Biological Networks and Data Types

Modeling of biological networks is a decades-old field that relies on the combination of multiple forms of evidence to construct interaction models of real biology. These models can present multiple types of interactions: signal propagation through phosphorylation cascades, activating transcriptional responses, transcriptional repression and complex associations may all be present. To perform inference with these

networks, whose utility lies in their ability to combine such disparate data sources, we need to know the individual characteristics of each interaction type and mode of curation.

A.1.1 Signaling Interaction Types

A.1.1.1 Ligand/Receptor Mediated Signaling:

In this model, an external ligand binds to a cell-surface exposed receptor that begins a process of signal transduction and mediation, resulting in a transcriptional or functional response from the cell. A good example of this process is the binding of TGF-β, which begins a signaling cascade by binding together type 1 and type 2 serine/threonine kinases on the cell surface [192]. This action allows the phosphorylation of the receptor 1 kinase by the receptor 2 kinase, which then propagates the signal by phosphorylating SMAD proteins, changing their conformation and converting them to receptor-activated (R-SMAD) proteins (see figure A.1). [192].

A.1.1.2 Complex Formation and Transcriptional Activation:

The phosphorylated R-SMAD-2 and SMAD-3 proteins then form a trimeric complex with SMAD4 proteins, and this activated complex then accumulates in the nucleus of the cell [67]. Once in the nucleus, this complex transcriptionally activates a number of genes [67, 192], including pro-apoptotic genes in the BCL2 family, and the cyclin depending kinase inhibitor CDKN1A.

In this example, we can observe several interaction types: ligand-binding in-

teractions, activating links in signal transduction pathways, complex formation, and transcriptional activation. In addition, recent research has suggested that mutant versions of the tumor-suppressor p53 may disrupt the complex formation of SMAD2/3 and SMAD4 – an example of a repressive interaction [67].

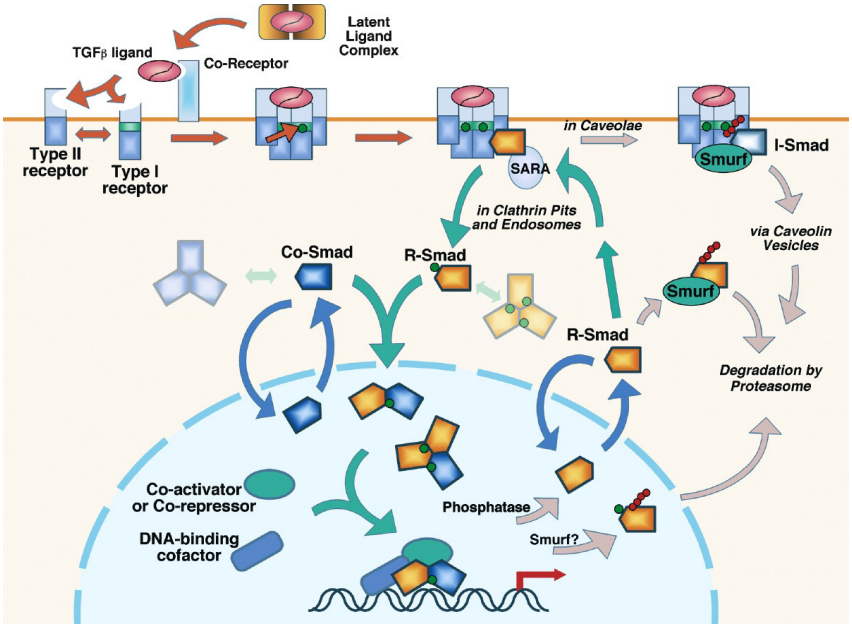

Figure A.1: Figure of TGF-β / SMAD signaling transduction from the Shi-Massague [192] paper.

A.1.2 Genome Sequencing and Variant Analysis

High-throughput genome sequencing is a set of technologies that allow us to read the exact DNA sequences of both normal and cancer cells, these techniques allow us to find the genomic differences that have caused the observed cancer phenotype. Several hardware platforms use short-read technology to read small bits of DNA (from a few dozen to a few hundred nucleotides long): these second generation sequencing

technologies include 454, Illumina and SOLiD [190]. These current technologies are accurate and cost-effective enough to use for sequencing of hundreds of tumor and normal samples, as used in The Cancer Genome Atlas projects [153], and potentially many more cancer patients in the near future. In addition, promising next-generation technologies like nanopore sequencing may soon provide individual sequencing for as little as $1,000 [29].

Because of the short-read nature of the current technologies, algorithmic approaches are needed to find the genomic locations that have been mutated or had regions duplicated or deleted in cancer cells. The University of California, Santa Cruz, has developed tools to produce variant "calls" from genomic sequencing data [121], as has the BROAD Institute, with MutSig (that operates on cohort-wide data) and the MuTect algorithm [23]. In addition, algorithms are available to find genes that are significantly targeted by focal amplification or deletion event in cancer cells [139].

A.1.3 Gene Expression Microarrays and Analysis

Gene expression microarrays are an older technology than genome sequencing but still play a key role in the analysis of cancer cells. Expression microarrays use small oligonucleotide probes that hybridize with specific regions of known genes, and that binding is usually detected with fluorescence; by comparing the binding in two difference conditions, the relative abundance of a known target can be assessed [186].

Microarray analysis has vastly enhanced our understanding of cancer biology since it's inception, having been used to predict the molecular subtypes of various can-

cers [168], drug efficacy [116] and many other clinically relevant features.

A.1.4 RNA-Seq

As with gene expression microarrays, RNA-Seq is a (newer) technology to quantify mRNA transcripts, and is a widely used tool in the analysis of cancer biology.

With this technology, a population of RNA is converted to a library of cDNA fragments with adaptors attached to one or both ends, after which each molecule (with or without amplification) is then sequenced in a high-throughput manner to obtain short sequences from one end (single-end sequencing) or both ends (pair-end sequencing) [221]. Short reads are generated (typically 30-400 bp [221]), and then, in the case of tumor DNA, aligned to a human reference genome. Compared with micro-array data, RNA-Seq has very low background signal and very high dynamic range, due to the ability to uniquely map reads [221], making this technology much better at detecting lowly expressed RNAs as well as quantifying expression change between experimental conditions.

A.1.5 Reverse Phase Protein Arrays (RPPA)

Proteomics deals with large-scale determination of gene and cellular function at the protein level, combining a collection of technical disciplines [15]. Reverse phase protein arrays (RPPA) are an extremely important new technology for detecting the status of proteins in tumor cells. Unlike transcriptional or genome sequence analysis, RPPA allows for a direct assay of the functional status (phosphorylation, cleavage,

mutational) and abundance of proteins [205].

With this technique, samples are spotted in parallel on solid-phase carriers (plates) [136] to allow for high-throughput detection of protein activity, at a very high level of accuracy and repeatability for known proteins and states [205]. Primary antibodies are used for protein capture, with secondary antibodies and fluorescence used to quantify measurements. One downside of the antibody approach, however, is that only known proteins and functional mutations can be captured [136], which means that novel functional mutations may be missed from the assay.

A.1.6 Mass Spectrometry for Protein Phosphorylation Quantification

Like RPPA technology, mass spectrometry techniques allow for large scale quantification of protein abundance, allowing biologists to directly measure the functional status of the proteome. A mass spectrometer consists of an ion source, a mass analyser that measures the mass-to-charge ratio *(m/z)* of the ionized analytes, and a detector that registers the number of ions at each m/z value [15]. This technique allows for specific identification and quantification of protein molecules.

Through the development of specific antibodies that recognize only phosphorylated protein residues, it is possible to use mass-spectrometry to ascertain the quantity of phosphorylated protein-peptides [81]. Because reversible phosphorylation of proteins is a major conduit of information flow from transmembrane receptors to the nucleus [81, 167], this technique is useful in determination of cellular state, in a way that is highly complementary to measurements of gene-expression and genomic state.

A.1.7 RNAi techniques for drug target identification

Tumor growth is achieved by an evolutionary process that allows tumors to rapidly proliferate, evade immune detection, attract new nutrient-carrying blood vessels and eventually metastasize [132, 142], via a number of gain of function and loss of function genomic (and epi-genomic) alterations. Because of this characteristic oncogene and non-oncogene addiction [132], cancer therapies must try to hit key functional nodes that will not only cripple tumor growth, but also be selective enough to limit damage to normal cells. Making this task more difficult, is the surprising diversity among tumors regarding the presence or absence of a number of low frequency mutations. It seems that these low frequency driver mutations may all contribute to the extensive rewiring and phenotypic changes seen in tumor cells [132, 222] and makes the problem of drug-target selection for a given tumor more difficult.

To solve this problem, we need a strategy to test the effects of inhibiting each of many different targets, and in combination with a number of different drugs or other targeted therapies. One can view this as a large multi-dimensional landscape of drug/target(s) combinations, to which we may apply a sampling strategy to explore. This is a similar concept to synthetic-lethality, which has been used to find the combinations of gene knock-outs that cause growth defects in yeast [49], or chemical-genomic lethality used to find the same with drug-gene combinations [92].

One experimental technology that has been used to successfully sample the landscape of possible target therapies is RNAi inhibition. This technique mimics natural

small interfering (siRNA) and micro (miRNA) RNA pathways, by introducing short hairpin RNAs (shRNA) that bind to transcribed mRNAs, forming a double stranded molecule that is then cleaved by Dicer enzymes or otherwise silenced [193]. These shRNAs are designed to target specific transcripts, allowing a "library" of silencing RNAs to be created; and, with high throughput welling techniques a number of potential targets can be screened this way [193, 209].

A.2 Sub network Computation

A.2.1 Algorithms for sub network extraction

Many of the current approaches to biological pathway search use some form of linear optimization to connect sets of genes over a curated network pathway. Flow based methods like ResponseNet [117] attempt to connect "source" genes to "target" genes over a network $G = (V, E)$, (which represents the vertices and directed edges, respectively), by providing each edge with a capacity [48] to allow a "flow" between source and target nodes. Because the capacities limit the amount of flow between each edge, the problem of finding the maximum flow between source and targets can be solved efficiently and exactly with an algorithm that iteratively tries to reroute data along edges with available capacities (the Ford-Fulkerson algorithm [48]). This algorithm is formulated as a linear optimization problem, where the goal is to minimize an objective function by applying either a depth first or breadth-first search (the Edmonds-Karp algorithm). To maximize flow between multiple sources and targets, this problem can

be modified by creating a "dummy" source node that feeds into only the true source node set, and a dummy target node that takes flow from the target set.

To adapt this algorithm to gene networks, one could imagine the source set of nodes as a set of proteins that have an effect on cellular signaling (cell surface receptors, kinases) that conduct an external signal through a set of intermediate nodes to a target set of transcription factors, which then modify the expression and phenotype of the cell. This signal is modeled as a flow, and the solution would be the most direct or parsimonious route that the signal might take to reach the target genes, given the known network topology. Lan *et al.* [117] used this approach to find signaling pathways connecting genetic hits to expression responses in yeast, defining an objective function and solving it using the linear optimization library LOQO [212].

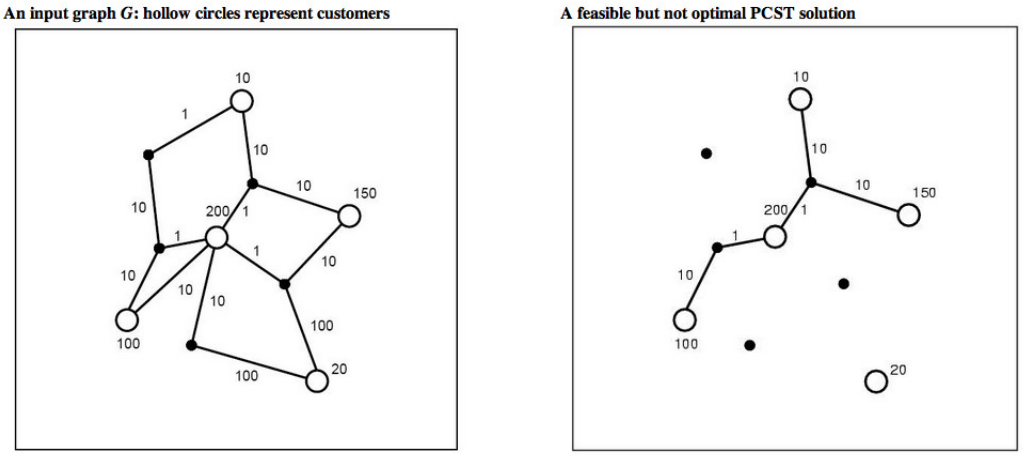

Figure A.2: An example of the Prize Collecting Steiner Tree problem [127].

The Prize Collecting Steiner Tree (PCST) problem defines a similar approach to a slightly different problem; under this formulation the problem is to connect a set

of weighted "prize" nodes over a interaction network where each edge has a weighted penalty. This method was used by Dittrich *et al.* [61] to identify functional modules in PPI networks, and the algorithm now part of the BioNet R-Package [27], a publicly downloadable library for the R language. The greater complexity of the PCST formulation makes it a costly computational problem to solve for larger gene sets over well connected networks. Code for an exact solution is available to those with an academic CPLEX library as part of Ivana Ljubi's dhea package [127], but, as the PCST problem is NP-hard the worst-case running time is often sufficiently long to make an exact solution impractical for sets of a few hundred or more nodes (genes) over a densely connected network. Fortunately, a fast-heuristic is built-in to the BioNet package, and an approximate solution is usually sufficient for the purposes of connecting sets of genes over a biological interaction network.

Huang *et al.* used the PCST framework to connect perturbed "genetic hits to the transcriptional responses accompanying those perturbations though a curated signaling network in yeast [189]. This framework can also be extended to connect genomic perturbations such as mutations, and copy number variations in key cancer genes to differentially transcribed genes in human networks.

A.3 Master Regulator Analysis / VIPER Analysis

The "Master Regulator Analysis" algorithm [124] uses a modification of the Gene Set Enrichment (GSEA) ALGORITHM [199] to infer the activity of a set of

transcription factors (TF), from the expression of the TFs downstream targets (see the diagram in figure A.3 below). These activity scores (or the significance of the scores tested against random target sets of identical size) predict the protein activity of TFs and have been shown to provide a much more robust measure than the gene expression alone.

The VIPER algorithm is available as a Bioconductor/R (bioconductor.org) package, and implements master regulator analysis and a sample-specific version of the algorithm.

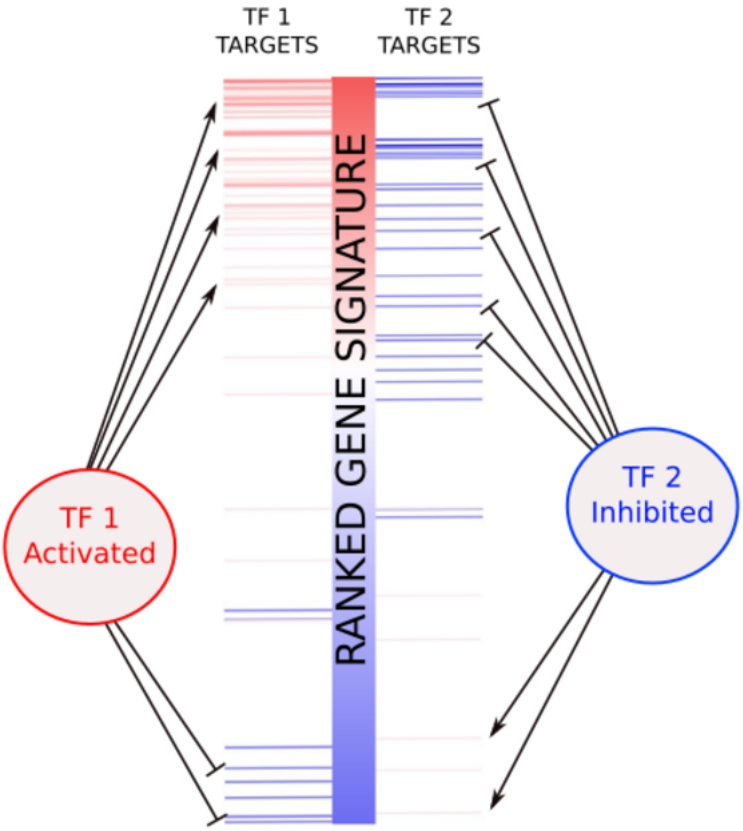

Figure A.3: An illustration of the Master Regulator Analysis (MRA) / MARINa algorithm, taken from [17].

A.4 Probabilistic Graphical Models

A Bayes belief network is a directed, a-cyclic graphical structure that represents the conditional dependencies between a set of random variables. This structure allows us to represent the complete joint distribution of a relatively large set of variables in a simplified manner, because each variable is dependent on only it's direct neighbors. For a biological network, a Bayes net model can be used to represent the regulatory links between biological entities, and to discover complex interactions between nodes. This type of probabilistic graphical model (PGM) can perform two important tasks: inference and learning. The former allows us to use a PGM to answer questions related to it's model of the world: for example, if gene A activates gene B in our model, and we know that gene A has high activity in a given data set, we can infer that gene B is also up-regulated from our model. Learning in a PGM framework allows us to construct a model that approximates some feature of a set of experimental observations [112] – a particularly useful and data-driven approach.

One immediate issue that arises when applying a Bayes netork model to biological networks is the requirement that these networks be a-cyclic; this problem is solved with approximate or "loopy" belief propagation, which can produce good inference approximations on cyclic graphs in many cases [112]. This type of message passing inference algorithm has been used to do inference on biological networks, including with the PARADIGM tool, developed by Vaske *et al.* [217]. PARADIGM uses the C++ library LibDAI to perform inference [145] and learning, which has been proven

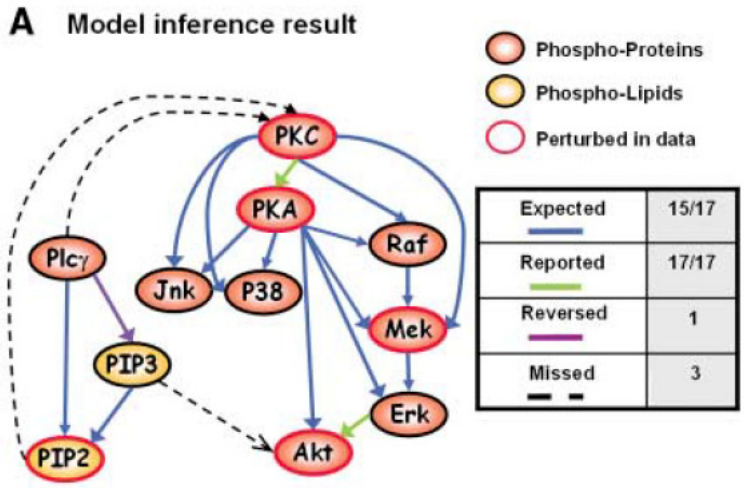

Figure A.4: Figure from Sachs *et al.* [183] shows the results of applying a structure learning algorithm to a small biological network and input data set.

to be computationally efficient over even large biological networks (thousands of nodes). Under this model, learning the parameter space of a PGM network representation is accomplished with a factor graph model [217] and an interactive expectation maximization algorithm to find optimum parameter values of the PGM [112].

Structure learning is another important use of probabilistic graphical models: this task involves learning not only the parameter space of a given PGM, but also searching though the space of possible edge additions and deletions, finding an optimal structure to explain a given data set. The structural EM algorithm developed by Nir Friedman are able to search both structure and parameter spaces concurrently [72], and has since been further developed for efficiency [169]. Structure learning has been used to infer biological networks before (see figure A.4 above), but because of the complexity of searching for all possible links, it is often infeasible to apply to networks larger than

a few dozen nodes [112].